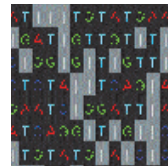




InfoQuest™ FP Software
Instruction Manual | Version 5



NOTES

SUPPORT BY BIO-RAD LABORATORIES, INC.

While the best efforts have been made in preparing this manuscript, no liability is assumed by the authors with respect to the use of the information provided.

Bio-Rad will provide support to research laboratories in developing new and highly specialized applications, as well as to diagnostic laboratories where speed, efficiency and continuity are of primary importance. Our software thanks its current status for a part to the response of many customers worldwide. Please contact us if you have any problems or questions concerning the use of InfoQuest™ FP, or suggestions for improvement, refinement or extension of the software to your specific applications:

Bio-Rad Laboratories, Inc.
Life Science Group
2000 Alfred Nobel Drive
Hercules, CA 94547

Technical Support: (800) 424-6723 and (510) 741-6910

FAX: (510) 741-5802

E-MAIL: LSG.TechServ.US@Bio-Rad.com (US)

LSG.TechServ.Intl@Bio-Rad.com (International)

The InfoQuest FP software and this accompanying guide are subject to the terms and conditions outlined in the License Agreement. The support, entitlement to upgrades and the right to use the software automatically terminate if the user fails to comply with any of the statements of the License Agreement.

No part of this guide may be reproduced by any means without prior written permission of the authors.

URL: <http://www.consult.bio-rad.com>

Copyright (C) 1998, 2008, Bio-Rad Laboratories, Inc. All rights reserved.

LIMITATIONS ON USE

InfoQuest FP is a registered trademark of Bio-Rad Laboratories, Inc.

All other product names or trademarks are the property of their respective owners.

InfoQuest FP includes a library for XML input and output from Apache Software Foundation (<http://www.apache.org>).

Table of contents

NOTES

TABLE OF CONTENTS

1. INTRODUCTION 7

1.1 The concepts of InfoQuest FP.....9

- 1.1.1 The programs9
- 1.1.2 The database and the experiments.....9
- 1.1.3 Multi-database setup9
- 1.1.4 Home directory and databases.....11
- 1.1.5 Modules and features.....11
- 1.1.6 The InfoQuest FP script language12

1.2 About this guide13

- 1.2.1 Conventions.....13
- 1.2.2 Toolbars13
- 1.2.3 Floating menus13

1.3 Installing the software as a standalone

license15

- 1.3.1 The InfoQuest FP Setup program15
- 1.3.2 Example database.....16

1.4 Installing the InfoQuest FP network

software.....19

- 1.4.1 Introduction.....19
- 1.4.2 Setup.....19
- 1.4.3 Advanced features of the Netkey server
program.....21
- 1.4.4 Features of the client program
(InfoQuest FP).....23

1.5 Starting and setting up InfoQuest FP ..25

- 1.5.1 The InfoQuest FP Startup program25
- 1.5.2 Creating a database25

- 1.5.3 Installing plugin tools in a new database...27

1.6 The InfoQuest FP user interface.....29

- 1.6.1 Introduction to the InfoQuest FP
user interface29
- 1.6.2 The InfoQuest FP main window29
- 1.6.3 General appearance of InfoQuest FP
windows30
- 1.6.4 Display of panels31
- 1.6.5 Configuring toolbars32
- 1.6.6 Grid panels33
- 1.6.7 Zoom sliders35
- 1.6.8 Behaviour of InfoQuest FP windows36
- 1.6.9 Navigator pane.....36

2. DATABASE 37

- 2.1 Introduction39
- 2.1.1 Local and connected databases39
- 2.1.2 Elementary structure of a database39
- 2.1.3 Location of a database39
- 2.1.4 Setting up a new database.....40
- 2.1.5 Protecting a database.....41
- 2.1.6 Log files41

2.2 Database functions43

- 2.2.1 Adding entries to the database.....43
- 2.2.2 Creating information fields.....43
- 2.2.3 Entering information fields.....44
- 2.2.4 Attaching files to database entries.....45
- 2.2.5 Information field properties46
- 2.2.6 Configuring the database layout47
- 2.2.7 Selections of database entries48
- 2.2.8 Manual selection functions48
- 2.2.9 Automatic search and select functions49
- 2.2.10 The advanced query tool.....49
- 2.2.11 Subsets53
- 2.2.12 Opening an additional database55

2.3 Connected databases	57
2.3.1 Advantages of a connected database	57
2.3.2 Setting up a new connected database	58
2.3.3 Configuring the connected database link in InfoQuest FP	59
2.3.4 Working in a connected database	61
2.3.5 Linking to an existing database with standard InfoQuest FP table structure	61
2.3.6 Linking to an existing database with table structure not in InfoQuest FP format	62
2.3.7 Converting a local database to a connected database	63
2.3.8 Opening and closing database connections	65
2.3.9 Restricting queries	66
2.3.10 Protecting connected databases with a password	69
2.4 Levels and relations in a database	71
2.4.1 Introduction	71
2.4.2 Creating levels	72
2.4.3 Creating new relation types	72
2.4.4 Defining relations between entries	73
2.4.5 Relations and scripts	75
2.4.6 Different relation types	75
2.5 Importing data in a InfoQuest FP database	77
2.5.1 Importing data using the Import plugin ...	77
2.5.2 Importing data via an ODBC link	77
2.6 Database exchange tools	81
2.6.1 Solutions for data exchange: bundles and XML files	81
2.6.2 Using bundles in InfoQuest FP	81
2.6.3 Export and import using XML files	83
2.7 Taking backups from a InfoQuest FP database	85
2.7.1 Backing up a local database	85
2.7.2 Backing up a connected database	85

3. EXPERIMENTS	87
3.1 Experiment types available in InfoQuest FP	89
3.2 Setting up fingerprint type experiments	91
3.2.1 Defining a new fingerprint type	91
3.2.2 Processing gels	93
3.2.3 Defining pattern strips on the gel	94
3.2.4 Defining densitometric curves	99
3.2.5 Normalizing a gel	101
3.2.6 Defining bands and quantification	105
3.2.7 Advanced band search using size-dependent threshold	107
3.2.8 Quantification of bands	108
3.2.9 Editing the fingerprint type settings	110
3.2.10 Adding gel lanes to the database	112
3.2.11 Adding information to fingerprint files and fingerprint lanes	113
3.2.12 Superimposed normalization based on internal reference patterns	114
3.2.13 Import of molecular size tables as fingerprint type	117
3.2.14 Conversion of gel patterns from GelCompar versions 4.1 and 4.2	119
3.2.15 Dealing with multiple reference systems within the same fingerprint type	119
3.3 Setting up character type experiments	121
3.3.1 Defining a new character type	121
3.3.2 Editing a character type	121
3.3.3 Input of character data	123
3.3.4 Character type settings restricted to connected databases	125
3.3.5 Import of character data by quantification of images scanned as TIFF files	126
3.4 Setting up sequence type experiments	135
3.4.1 Defining a new sequence type	135

3.4.2 Importing sequences	135	4.1.11 Dendrogram display functions	175
3.4.3 Input of sequences using the InfoQuest FP Assembler program	136	4.1.12 Working with Groups	176
3.5 Setting up trend data type experiments	151	4.1.13 Cluster significance tools.	179
3.5.1 Introduction.	151	4.1.14 Matrix display functions	180
3.5.2 Defining a new trend data type	151	4.1.15 Group statistics.	181
3.5.3 Entering trend data in the database.	156	4.1.16 Printing a cluster analysis.	183
3.5.4 Displaying trend data	157	4.1.17 Exporting rendered trees.	185
3.5.5 Additional comparison parameters.	158	4.1.18 Analysis of the congruence between techniques	186
3.5.6 Comparison settings.	158	4.2 Cluster analysis of fingerprints	191
3.6 Setting up matrix type experiments . .	159	4.2.1 Fingerprint comparison settings	191
3.6.1 Defining a new matrix type.	159	4.2.2 Fingerprint display functions.	192
3.7 Setting up composite data sets	161	4.2.3 Defining 'active zones' on fingerprints.	193
3.7.1 Introduction.	161	4.2.4 Calculation of optimal position tolerance optimization and settings	194
3.7.2 Defining a new composite data set	161	4.3 Band matching and polymorphism analysis	197
3.8 Experiment display and edit functions	163	4.3.1 Introduction	197
3.8.1 The experiment card.	163	4.3.2 Creating a band matching.	197
3.8.2 Gelstrips.	163	4.3.3 Manual editing of a band matching	199
3.8.3 Character experiment cards	164	4.3.4 Adding entries to a band matching.	201
3.8.4 Sequence experiment cards.	165	4.3.5 Saving band classes to a fingerprint type	201
4. COMPARISONS	167	4.3.6 Band and band class filters	202
4.1 General comparison functions	169	4.3.7 Exporting band matching information	203
4.1.1 Definition.	169	4.3.8 Tools to display selective band classes	203
4.1.2 The Pairwise comparison window.	169	4.3.9 Creating a band matching table for polymorphism analysis.	204
4.1.3 The Comparison window	170	4.3.10 Finding discriminative bands between entries	205
4.1.4 Adding and removing entries.	172	4.4 Cluster analysis of characters	207
4.1.5 Rearranging entries in a comparison.	172	4.4.1 Character comparison settings.	207
4.1.6 Saving and loading comparisons	173	4.4.2 Character display functions	209
4.1.7 Interaction between subsets and comparisons	173	4.4.3 Advanced analysis of massive character sets using GeneMaths XT.	209
4.1.8 Cluster analysis: introduction.	174	4.5 Multiple alignment and cluster analysis of sequences	211
4.1.9 Calculating a dendrogram.	174	4.5.1 An introduction to sequence analysis	211
4.1.10 Calculation priority settings	174		

4.5.2 Calculating a cluster analysis based on pairwise alignment	212	4.6.15 Matrix display functions.	237
4.5.3 Calculating a multiple alignment	213	4.6.16 Printing and exporting a sequence alignment	237
4.5.4 Multiple alignment display options	214	4.6.17 Finding sequence positions in an alignment	239
4.5.5 Editing a multiple alignment	215	4.6.18 Sequence translation	239
4.5.6 Drag-and-drop manual alignment	215	4.6.19 Subsequence search.	239
4.5.7 Inserting and deleting gaps	215	4.6.20 Mutation search	241
4.5.8 Removing common gaps in a multiple alignment.	217	4.6.21 Defining bookmarks in a sequence alignment	243
4.5.9 Changing sequences in a multiple alignment	217	4.7 Cluster analysis of trend data	245
4.5.10 Finding a subsequence	217	4.7.1 Trend data comparison settings	245
4.5.11 Calculating a clustering based on the multiple alignment (steps 5 and 6)	218	4.7.2 Display options for trend data.	245
4.5.12 Adding entries to and deleting entries from an existing multiple alignment	218	4.8 Cluster analysis of composite data sets	247
4.5.13 Automatically realigning selected sequences	219	4.8.1 Principles.	247
4.5.14 Sequence display and analysis settings	219	4.8.2 Calculating a dendrogram from a composite data set	247
4.5.15 Exporting a multiple alignment.	220	4.8.3 Finding discriminative characters between entries	250
4.5.16 Converting sequence data to categorical character sets	220	4.8.4 Transversal clustering.	250
4.5.17 Excluding regions from the sequence comparisons.	222	4.9 Phylogenetic clustering methods	253
4.5.18 Writing comments in the alignment	223	4.9.1 Introduction	253
4.6 Sequence alignment and mutation analysis	225	4.9.2 Maximum parsimony of fingerprint and character type data	253
4.6.1 Introduction	225	4.9.3 Maximum parsimony clustering of sequence data	254
4.6.4 Creating a new alignment project	226	4.9.4 Maximum likelihood clustering	256
4.6.5 The Alignment window	227	4.10 Advanced clustering and consensus trees	257
4.6.6 General functions.	228	4.10.1 Introduction	257
4.6.7 Adding and removing entries	229	4.10.2 Degeneracy of dendrograms	257
4.6.8 Aligning sequences	230	4.10.3 Consensus trees	258
4.6.9 Calculating a consensus sequence.	232	4.10.4 Advanced clustering tools	259
4.6.10 Display options for sequences and curves	233	4.10.5 Displaying the degeneracy of a tree	259
4.6.11 Editing an alignment.	235	4.10.6 Creating consensus trees	261
4.6.12 Calculating a global cluster analysis.	235		
4.6.13 Dendrogram display functions	236		
4.6.14 Cluster significance tools	237		

4.11 Minimum spanning trees for population**modelling263**

- 4.11.1 Introduction.....263
- 4.11.2 Minimum spanning trees in InfoQuest FP263
- 4.11.3 Calculating a minimum spanning tree from character tables.....264
- 4.11.4 Interpreting and editing a minimum spanning tree266
- 4.11.5 Calculating a minimum spanning tree from a similarity matrix.....269

4.12 Dimensioning techniques271

- 4.12.1 Introduction.....271
- 4.12.2 Calculating an MDS.....271
- 4.12.3 Editing an MDS.....271
- 4.12.4 Calculating a PCA272
- 4.12.5 Calculating a discriminant analysis.....276
- 4.12.6 Self-organizing maps.....276
- 4.12.7 Multivariate analysis of variance (MANOVA) and discriminant analysis.....278

4.13 Chart and statistics tools281

- 4.13.1 Introduction.....281
- 4.13.2 Basic terminology281
- 4.13.3 Charts and statistics.....283
- 4.13.4 Using the plot tool291
- 4.13.5 Bar graph.....292
- 4.13.6 Contingency table.....292
- 4.13.7 2-D scatterplot.....294
- 4.13.8 3-D scatterplot.....295
- 4.13.9 ANOVA plot.....296
- 4.13.10 1-D numerical distribution297
- 4.13.11 3-D Bar graph297
- 4.13.12 Colored bar graph297

5. IDENTIFICATION 299**5.1 Identification with database entries ..301**

- 5.1.1 Creating lists for identification.....301
- 5.1.2 Identifying unknown entries301

- 5.1.3 Fast band-based database screening of fingerprints.....302
- 5.1.4 Fast character-based identification303
- 5.1.5 Fast sequence-based identification304
- 5.1.6 Probabilistic identification304
- 5.1.7 BLAST sequence matching.....308

5.2 Identification using libraries313

- 5.2.1 Creating a library313
- 5.2.2 Identifying entries against a library.....314
- 5.2.3 Creating a neural network315

5.3 Decision networks319

- 5.3.1 Introduction319
- 5.3.2 Creating a new decision network319
- 5.3.3 Operators.....320
- 5.3.4 Building a decision network.....321
- 5.3.5 Display and output options for decision networks325
- 5.3.6 Working with layers in a decision network325
- 5.3.7 Using confidence values326
- 5.3.8 Building decisions relying on multiple states326
- 5.3.9 Creating charts from a decision network ..327
- 5.3.10 Executing a decision network from the InfoQuest FP main window.....328
- 5.3.11 Decision trees329

6. INFOQUEST FP 2D..... 333**6.1 Analyzing 2D gels335**

- 6.1.1 Proteomics in a broader context: the InfoQuest FP Platform.....335
- 6.1.2 Data sources for InfoQuest FP 2D335
- 6.1.3 Applications for InfoQuest FP 2D336
- 6.1.4 Automated workflow for experiments with repeats336
- 6.1.5 Automated workflow for multiplex experiments (DIGE).....336
- 6.1.6 Getting started with InfoQuest FP 2D.....336
- 6.1.7 Creating a new database.....337
- 6.1.8 Defining a new 2D gel type337

6.1.9 Importing 2D gel image files	338	7.1.3 Table ATTACHMENTS	379
6.1.10 Processing 2D gel images	338	7.1.4 Character Values table	379
6.1.11 Step 1: Spot detection	343	7.1.5 Character Fields table	379
6.1.12 Step 2: Calibration	346	7.1.6 Table COMPARISONS	380
6.1.13 Step 3: Normalization	348	7.1.7 Table DBSCHEMAS	380
6.1.14 Step 4: Defining metrics	351	7.1.8 Table DBSETTINGS	380
6.1.15 Step 5: Describing the 2D gel in the database	354	7.1.9 Table DECISNTW	380
6.1.16 Step 6: Normalization of other 2D gels ...	354	7.1.10 Table ENLEVELS	380
6.2 Comparing 2D gels	359	7.1.11 Table ENRELATIONS	381
6.2.1 Introduction	359	7.1.12 Table ENRELATIONTYPES	381
6.2.2 Matching spots on different gels	359	7.1.13 Table ENTRYTABLE	381
6.2.3 Creating 2D spot queries	363	7.1.14 Table EVENTLOG	381
6.2.4 Listing spots in tables	367	7.1.15 Table EXPERATTACH	381
6.2.5 Comparing spots in scatter plots	369	7.1.16 Table EXPERIMENTS	382
6.2.6 Clustering and statistical analysis of 2D gels in InfoQuest FP	371	7.1.17 Table FPRBNDCLS	382
6.2.7 Analyzing 2D gel spot tables with GeneMaths XT	372	7.1.18 Table FPRINT	382
6.2.8 Editing reference systems	373	7.1.19 Table FPRINTFILES	383
6.2.9 Creating synthetic gels	374	7.1.20 Table MATRIXVALS	383
7. APPENDIX	377	7.1.21 Table SEQTRACEFILES	383
7.1 Connected database table structure ..	379	7.1.22 Table SEQUENCES	384
7.1.1 Introduction	379	7.1.23 Table SUBSETMEMBERS	384
7.1.2 Table ALIGNPROJ	379	7.1.24 Table TRENDATA	384
		7.1.25 Indices in the database	384
		7.2 Regular expressions	387
		INDEX	

1. INTRODUCTION

1.1 The concepts of InfoQuest FP

1.1.1 The programs

The InfoQuest FP software is composed of two executable units: a **Startup program** that creates and manages the *databases* and associated directories and that starts the **Analyze program** with a selected database. All import and analysis functions are done in the Analyze program. This includes processing of gel files starting from TIFF images, including lane finding and normalization. Independent plugins allow the import of data in different file formats.

1.1.2 The database and the experiments

The logical flow of processing raw experiment files is represented in Figure 1-1. The basis of InfoQuest FP is a relational *database* consisting of *entries*. The entries correspond to the individual organisms or samples under study: animals, plants, fungi, bacterial or viral strains, organic samples, tissue samples,.... Each database entry is characterized by a unique *key*, assigned either automatically by the software or manually, and by a number of user-defined information fields. The organization and functions of InfoQuest FP databases are discussed in Chapter 2. Each entry in a database may be characterized by one or more *experiments* that can be linked easily to the entry. What we call experiments in InfoQuest FP are in fact the experimental data that are the numerical results of the biological experiments or assays performed to estimate the relationship between the samples. Chapter 3 of this manual deals with setting up experiments in InfoQuest FP. All biological experiments are functionally classified into six different classes, called *Experiment Types*:

- **Fingerprint types** (Section 3.2): Any densitometric record seen as a one-dimensional profile of peaks or bands can be considered as a fingerprint type. Examples are of course gel and capillary electrophoresis patterns, but also gas chromatography or HPLC profiles, spectrophotometric curves, etc. fingerprint types can be derived from TIFF or bitmap files as well, which are two-dimensional bitmaps. The condition is that one must be able to translate the patterns into densitometric curves.
- **2D gel types**: Any two-dimensional bitmap image seen as a profile spots or defined labelled structures. Examples are e.g. 2D protein gel electrophoresis patterns, 2D DNA electrophoresis profiles, 2D thin layer chromatograms, or even images from radioactively labelled cryosections or short half-life radiotracers. The analysis of 2D gels is discussed in Chapter 6.

- **Character types** (Section 3.3): Any array of named characters, binary or continuous, with fixed or undefined length can be classified within the character types. The main difference between character types and electrophoresis types is that in the character types, each character has a well-determined name, whereas in the electrophoresis types, the bands, peaks or densitometric values are unnamed (a molecular size is NOT a well-determined name!).
- **Sequence types** (Section 3.4): Within the sequence types, the user can enter nucleic acid (DNA and RNA) sequences and amino acid (protein) sequences.
- **Trend data types** (): Reactions to certain substrates or conditions are sometimes recorded as multiple readings in function of a changing factor, defining a trend. Examples are the kinetic analysis of metabolic and enzymatic activity, real-time PCR, or time-course experiments using microarrays. Although multiple readings per experiment are mostly done in function of time, they can also depend on another factor, for example readings in function of different concentrations.
- A sixth type, **matrix types** (Section 3.6), is not a native experiment type, but the result of a comparison between database entries, expressed as similarity values between certain database entries. An example of a matrix type is a matrix of DNA homology values. DNA homology between organisms can only be expressed as pairwise similarity, not as native character data.

Each experiment type is available as a module of the InfoQuest FP software (see 1.1.5).

Essentially, adding a single organism (entry) with its associated experiments to the database constitutes several steps (see Figure 1-1).

1.1.3 Multi-database setup

InfoQuest FP is a multi-database software, which supports the setup of different users in Windows NT. It is very important to understand the hierarchical structure of the user, database, and experiment setup in order to make optimal use of these features.

Windows NT (Windows 2000, XP and Vista): The InfoQuest FP users are associated with the Windows NT login users. Each Windows NT user can specify his/her InfoQuest FP databases directory, and InfoQuest FP saves this information in the user's system registry. For

example, suppose that a user X logs in on a Windows NT machine with InfoQuest FP installed. This user can create a directory, and specify this directory as the **home directory** in the Startup program. InfoQuest FP will save this information in this user's system registry, so that each time the user logs in, InfoQuest FP will automatically consider the same directory as the home directory. In this way, each Windows NT user can define his/her own InfoQuest FP home directory, without interfering with other users. Within this home directory, the user can specify as many *Databases* as desired. InfoQuest FP

allows two types of databases to be created: a local file-based database using a dedicated database mechanism developed by Bio-Rad, and a DBMS based type which relies on an external ODBC compatible relational database management engine (see Section 2.3 for more information). In the first type, protection of the InfoQuest FP databases depends on the protection of the specified directory by the Windows NT user. If a user protects the directory containing the InfoQuest FP databases, other users will not be able to change or to read the databases, depending on whether the directory is write or read

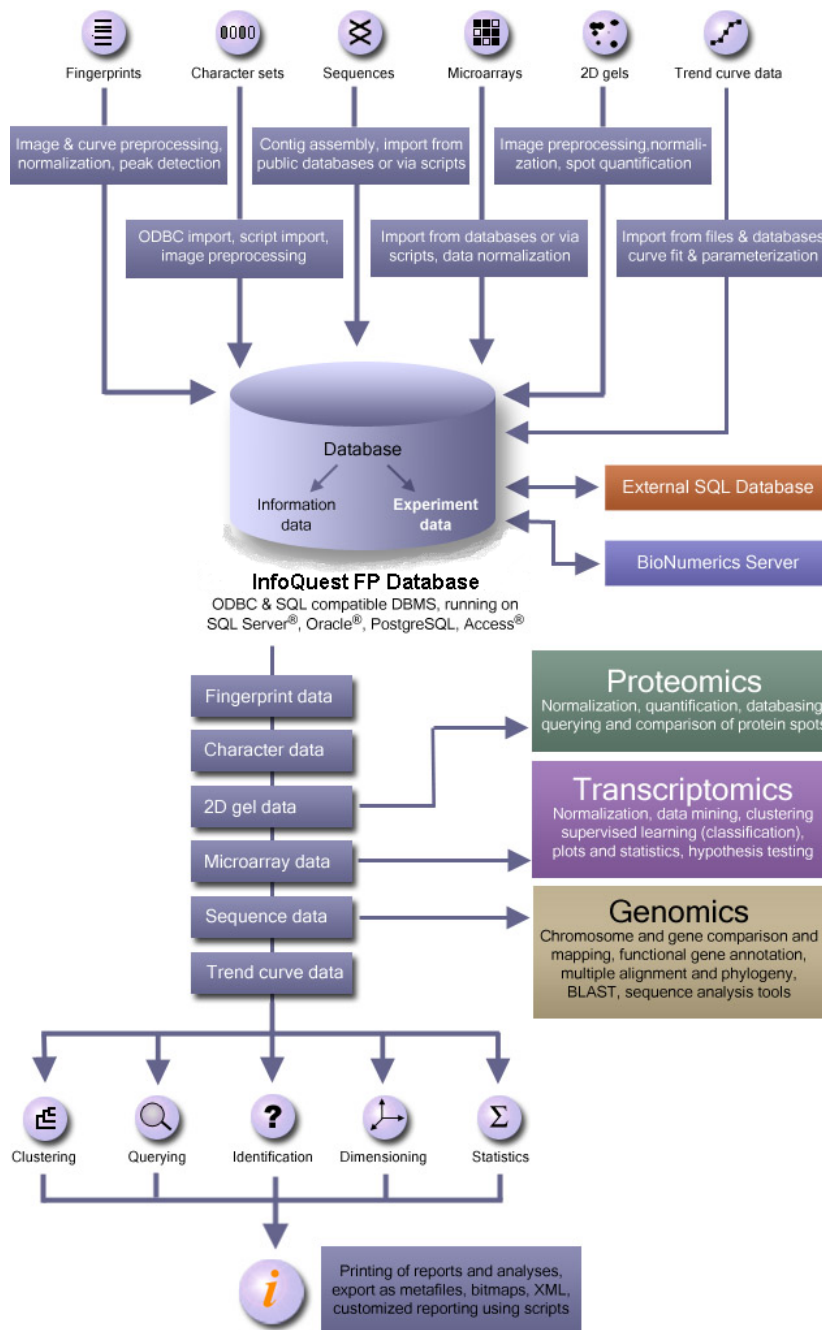


Figure 1-1. Flow chart of main steps in the acquisition and analysis of data in InfoQuest FP.

protected. In the second type, protection also relies on the protection and security measures provided by the DBMS (see Section 2.3.10).

1.1.4 Home directory and databases

As explained in the previous paragraph, InfoQuest FP recognizes its databases by means of a **home directory**. This home directory can be different per Windows login, and can even be on a different computer in the network.

By default, InfoQuest FP will install its databases under the home directory. However, a database can also be located in a different directory, and even on a different computer in the network. What is important is that in the home directory, a **Database descriptor file** is present for each database. These files have the name of the databases with the extension “.dbs”. The ***Database*.dbs** file is a pure text file which can be edited in Notepad or any other text editor.

The line after [BACKCOL] contains the RGB values for the window background color, and the line after [SAVELOGFILE] indicates whether log files are saved or not.

For databases created in a InfoQuest FP version prior to version 5.0, the line after the tag [DIR] indicates the full path where the database is located.

```
[DIR]
C:\Program Files\InfoQuestFP\data\DemoBase

[BACKCOL]
150 171 172

[SAVELOGFILE]
0
```

In databases created in version 5.0 or higher, the full path is replaced by a [HOMEDIR] tag. This tag points to the home directory as defined in the Startup screen. Because the database paths are stored relatively with respect to the home directory, databases can easily be copied to other locations or computers. After copying the database(s) (and their .dbs files) to another location, you only need to change the home directory (see 1.5.2.1).


```
[DIR]
[HOMEDIR]\DemoBase

[BACKCOL]
150 171 172

[SAVELOGFILE]
0
```

NOTES:







(1) If a database, created in a version prior to version 5.0, is moved from one computer to another, you need to edit the .dbs file and enter the correct path. The correct path for a database can also be entered from the Startup

screen by pressing  and selecting **Database settings** and **<Change directory>**.




(2) If you are working in version 5.0 or higher with databases created in a prior version, it may be useful to replace the paths in the .dbs files by a [HOMEDIR] tag. A script is available to store the paths relatively with respect to the home directory. Contact Bio-Rad to obtain this script.


(3) In case a database has been physically removed (or moved) from a computer, the ***Database*.dbs** file may still be present in the home directory, which causes the InfoQuest FP Startup program to list the database. When attempts are made to open or edit such a removed database, InfoQuest FP will produce an error. The only remedy is to delete the ***Database*.dbs** file.

1.1.5 Modules and features

The InfoQuest FP software consists of six application modules and four analysis modules. The **application modules** **Fingerprint types** , **2D gel types** , **Character types** , **Sequence types** , **Trend data types**  and **Matrix types**  correspond to each of the six experiment types described in 1.1.2.

The four **analysis modules** are not linked to a specific experiment type, but rather offer additional functionality for all experiment types:

- The **Comparison and cluster analysis** module  allows the user to create comparisons (see Chapter 4) and groups all functionality regarding cluster analysis.
- The **Identification** module  allows the user to identify unknown entries using the database, identification libraries, neural networks or decision networks (see Chapter 5).
- The **Dimensioning and Statistics** module  offers several non-hierarchical grouping methods, such as Principal Component Analysis, Multidimensional Scaling and Self-Organizing Maps (see Section 4.11). In addition, it comprises powerful statistics tools (see Section 4.12).

- The *Database sharing tools* module  allows researchers to exchange data with other institutions using bundles and XML files (see Section 2.3).

For each section in this manual where specific features are described, the required modules will be indicated in the section title.

The specific InfoQuest FP package you are working with might not include all modules. To check which modules are present, proceed as follows:

1.1.5.1 In the *InfoQuest FP main* window (see Figure 1-15), select **File > About**.

A window pops up, showing the version of the software, the package serial number, and a list with modules (see Figure 1-2). A module is present in the installed InfoQuest FP package when the module name is preceded by a hyphen.

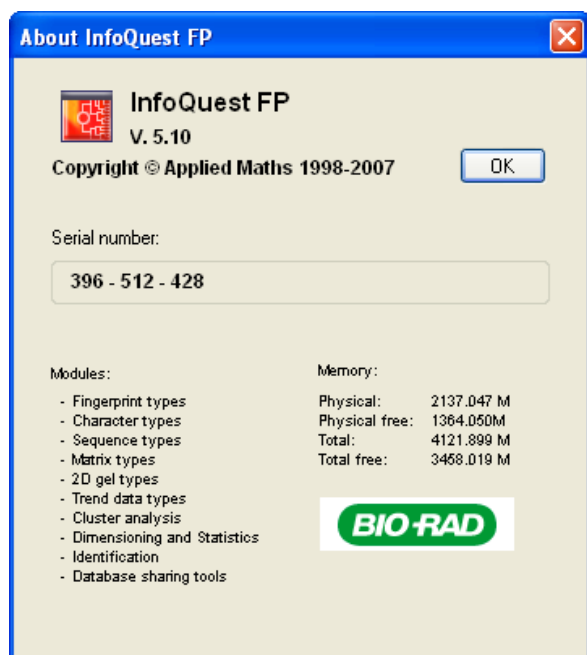


Figure 1-2. Window containing information about the installed InfoQuest FP package.

1.1.6 The InfoQuest FP script language




InfoQuest FP is a very comprehensive software package which has many data import, export and analysis functions already included in the software. Additional functionality - often related to specific applications - is bundled into convenient plugins (see 1.5.3). However, ultimate flexibility is offered by InfoQuest FP's own script language.


The InfoQuest FP script language is a programming language, containing a large array of specialized functions, that allows the user to create scripts (i.e. small programs) for automation of specific tasks e.g. the import of own data formats.

A script editor can be opened from the *InfoQuest FP main* window by selecting **Scripts > Edit script**. For a description of the *Script Editor* window and an explanation of the numerous script functions available, we refer to the separate script manual.

NOTE: The functionality to edit scripts is available in any InfoQuest FP configuration, but depending on the software configuration (modules present or not, see 1.1.5), some script functions might be disabled.

A number of general scripts are available on the website of Bio-Rad. These scripts can be launched from the *InfoQuest FP main* window, using **Scripts > Browse Internet**

or by pressing the  button. In the browser window that appears, click on a category to display the relevant scripts. When leaving the  checkbox in the browser toolbar unchecked, a script can be executed directly over the internet by clicking on its name (recognized by the preceding  icon). When the checkbox

is checked (), you will be prompted for a destination folder. Scripts can be saved in any folder, but two locations are predestined: Saving the script in **C:\Program files\InfoQuest FP\Scripts** makes the script available as a menu item in the *Scripts* menu of any InfoQuest FP database. To make a script only appear as a menu item in a specific database, save the script in the corresponding **[HOMEDIR]*dbname*\Scripts** folder.

*NOTE: Using **Scripts > Run script from file**, a script can be executed from any location.*

1.2 About this guide

1.2.1 Conventions

In the sections that follow, all menu commands and button text is typed in ***bold-italic***. Submenus are separated from parent menus by a “greater than” sign (>). Button text is always given between < and > signs.

For example, the following menu command (Figure 1-3) will be indicated as ***Edit > Arrange entries by field***.

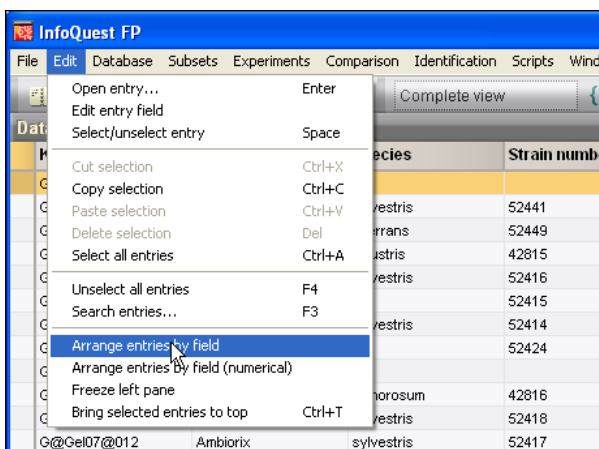


Figure 1-3. Illustration of the menu command *Edit > Arrange entries by field*.

In Figure 1-4, the following buttons are indicated as ***Consider absent values as zero*** (check box), ***<OK>***, ***<Cancel>***, ***<Apply>*** (disabled).

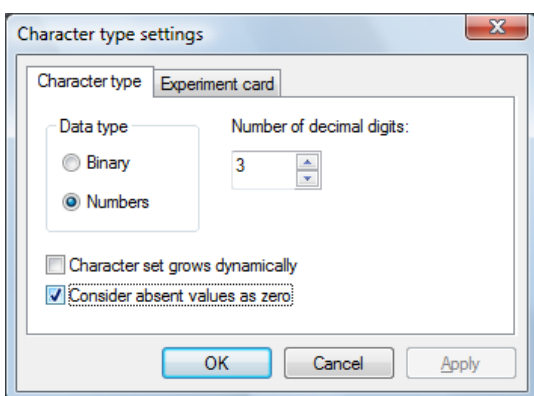


Figure 1-4. Illustration of buttons and check boxes in the *Character type settings* dialog box.

Each window and dialog box described in the guide will be given a name. This name is shown in *italic*, and usually corresponds to the name in the caption of the window or dialog box. For example, the dialog box in Figure 1-4 will be called the *Character type settings* dialog box.

Descriptive text, such as explaining the layout of windows, describing the function of available menu items and buttons, or providing background information on the use of different algorithms, etc., is always displayed in normal text layout (such as the present paragraph), without preceding paragraph number. Tutorial text, which guides the user through the program by applying the available analysis functions on example data, are always preceded by a paragraph number for easy reference. This is illustrated in the following example:

1.2.1.1 In the Startup screen, double-click on **DemoBase** to open the database for analysis. Alternatively, you can select **DemoBase** from the list and then press the



button.

Names of databases (e.g. **DemoBase**) and experiment types (e.g. **PhenoTest**) in InfoQuest FP are typed in **bold** face.

1.2.2 Toolbars

In almost every window in the InfoQuest FP Analyze program, there is a toolbar containing buttons for the most common functions available in the window. Placing the mouse pointer on a button for one second invokes a tool tip to appear, explaining the meaning of the button.

1.2.3 Floating menus

In almost every window in the InfoQuest FP Analyze program, the use of place-specific “floating menus” is supported. For example, if you *right-click* (clicking the right mouse button) on a database entry, a floating menu is popped up, showing you all the possible menu commands that apply to the selected entry (see Figure 1-5).



Figure 1-5. Floating menu appearing after clicking right on a database entry.

The floating menus make the use of InfoQuest FP easier and more intuitive for beginners, and much faster for experienced users. In describing menu commands in this guide, we will not usually mention the corresponding floating menu command. It is up to the user to try right-clicking in all window panels in order to find out which is more convenient in every specific case: calling the command from the window's menu or toolbar button or from the place-specific floating menu.

1.3 Installing the software as a standalone license

1.3.1 The InfoQuest FP Setup program

The InfoQuest FP software is delivered on CD-ROM or can be downloaded from the Bio-Rad website (www.bio-rad.com/softwaredownloads).

1.3.1.1 Insert the protection key (dongle) in the parallel or USB port of the computer.

1.3.1.2 If you insert the CD-ROM in the drive, the Setup program will automatically load if the *Auto insert notification* of the CD drive is enabled. If not, or if you have downloaded the setup files from the website, run **Setup.exe**.

1.3.1.3 On the Setup intro screen (Figure 1-6), click **Install**.

The *Installation* dialog box (Figure 1-7) allows you to change the *Installation directory*, to specify the *Database home directory*, *Install the Sample database*, and to *Install the Netkey Server program*.

- **Installation directory:** By default, the software installs itself in a subdirectory InfoQuest FP of the Program files directory. To change the installation path, click the **<Browse>** button and navigate to another path.

- **Database home directory:** The program offers two default options for the location of the home directory: *In Common Documents* and *In My Documents*. The



Figure 1-6. The InfoQuest FP Setup screen.

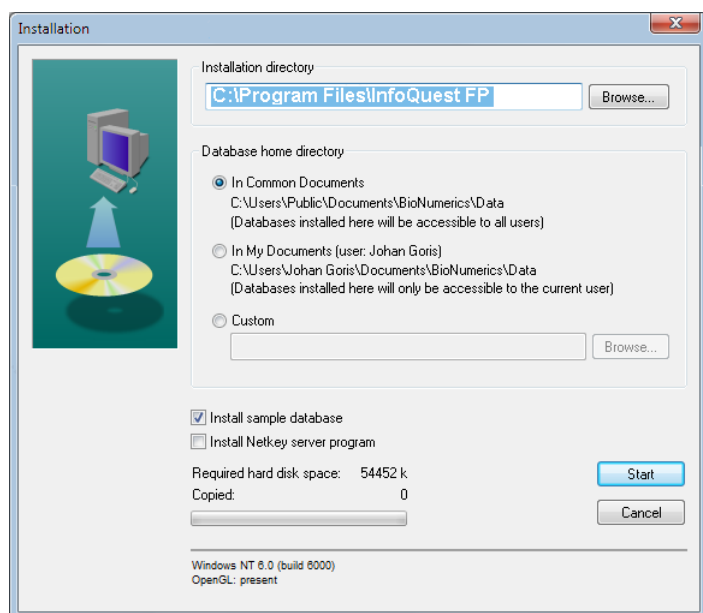


Figure 1-7. The *Installation* dialog box.

option *In Common Documents* makes the InfoQuest FP databases available for all users on that computer. The option *In My Documents* makes the databases only accessible to the current user. The third option, *Custom folder*, allows the user to specify any desired directory. You can even specify a directory on a network drive, on condition that this drive is permanently available.

- **Install sample database:** This option installs the sample (demo) databases **DemoBase** and **Demobase 2D** with the software.

- **Install Netkey server program:** see 1.4.2.

1.3.1.4 If this is the first installation of InfoQuest FP, you should allow the program to install the sample databases. Do not check the *Install Netkey server program* checkbox if you are installing a standalone version of the software.

When file copying is completed, the Setup program prompts to create a shortcut for InfoQuest FP on the desktop (recommended).

Upon completion, the installation program prompts you to confirm the installation of the drivers for the protection key.

1.3.1.5 If you are not installing a network license (see Section 1.4 for the installation and features of InfoQuest FP network licenses), press **<OK>** to allow the installation of the dongle drivers.

1.3.1.6 If you allowed the setup program to create a shortcut on the desktop, double-click on the InfoQuest FP icon to start the software. If not, open the Start menu and select *InfoQuest FP* under *All programs*.

When InfoQuest FP is started for the first time, it will prompt you to enter the *License String*. The License String is required to activate your license and has to be entered only one time after a new installation and again after the installation of an upgrade. It is stored in the Windows registry.

The License String is provided on the jewel case of the CD-ROM, or in case of an upgrade or an Internet license, you may have obtained it electronically.

Enter the 6 x 4 characters License String in the input fields and press **<OK>**. The *InfoQuest FP Startup* program now appears (see 1.5.1).

1.3.2 Example database

One database, **DemoBase**, is installed with the software, and this database will serve as a tutorial and as an example in this guide. This database contains experimental data on some fictitious (bacterial) genera. The database contains the following experiment types:

- **Fingerprint types:**

RFLP: Two different RFLP techniques, called **RFLP1** and **RFLP2**, resulting in two patterns for each bacterial strain.

AFLP: Amplified Fragment Length Polymorphism profiles (AFLP), run on an ABI PRISM 310 Genetic Analyser (Applied BioSystems).

- **Character types:**

FAME: Fatty Acid Methyl Esters (FAME) profiles obtained on a Hewlett Packard 5890A gas-liquid chromatography instrument. This is a typical example of an *open* data set: the number of fatty acids found depends on the group of entries analyzed. If more entries are added, more fatty acids will probably be found. Furthermore, FAME profiles are an example of a *continuous* character type: the percentage occurrence of a fatty acid in a bacterium can have any real value between zero and 100%.

PhenoTest: This is a fictitious phenotypic test assay that reveals the metabolic activity or enzyme activities of bacteria on 19 different compounds. The first cup of the test is a blank control. This is an example of a closed data set: the 20 characters are well-defined, and regardless of the number of entries examined, the number of characters in the experiment will always remain 20. Real examples of such types of assays exist as commercial test panels available on microplates or galleries. They can be interpreted in two ways. One can read the reactions by eye and score them as positive or negative; in this case the character type is *binary*. If the microplates are read automatically using a microplate reader, the reactions in the cups may have any real value between an OD of zero and 2.5 to 3.0, which is a *continuous* character type. In the example database, the reactions are scored as *continuous* characters.

In addition to the binary and continuous character types, one can also distinguish the *semi-quantitative* character types. These are tests that can have a number of discrete values, e.g. 0, 1, 2, 3, 4, or 5. In practice, a number of continuous character types are interpreted as multistate characters for convenience.

- **Sequence types:**

16S rDNA: For all of the strains, and a number of additional strains, the nearly complete 16S ribosomal RNA gene has been sequenced. The sequences are approximately 1500 bases long, but not all of them are sequenced completely.


- **Matrix types:**

A partial homology matrix based upon hybridization of total genomic DNA has been generated for the genera.

The **DemoBase** database which is installed with the software, is a *local* database (see Section 2.1.1 for more information on local and connected databases). A number of more advanced features in InfoQuest FP are only available for connected databases. Therefore, the data

contained in **DemoBase** is also provided as a connected database. Since the ODBC connection is dependent on the computer configuration, the data are provided as a **DemoBase_SQL.mdb** file on the installation CD-ROM. Proceed as follows to install this database:

1.3.2.1 In the InfoQuest FP Startup program (see 1.5.1),

press the  button to create a new database.

1.3.2.2 The *New database* wizard (see 1.5.2 for a detailed description) prompts for a database name, enter **DemoBase_SQL**.

1.3.2.3 Leave all settings at their defaults (press the **<Next>** button twice) and press **<Finish>** to complete the setup of the new database.

1.3.2.4 Leave the default option *New connected database (automatically created)* enabled and press **<Proceed>**.

1.3.2.5 Press **<Proceed>** in the *Plugin installation* toolbox, without installing any plugin (see 1.5.3 for more information on plugins).

The InfoQuest FP home directory (as defined in the Startup program, see 1.5.2) should now contain a database folder called **DemoBase_SQL**.

1.3.2.6 In Windows explorer, simply replace the automatically generated **DemoBase_SQL.mdb** file from the **DemoBase_SQL** folder with the **DemoBase_SQL.mdb** file from the installation CD-ROM.

When the connected database **DemoBase_SQL** is now opened from the InfoQuest FP Startup screen, it will contain the same example data as available in the local database **DemoBase**. Any tutorial paragraph in this manual that uses **DemoBase** is also applicable for **DemoBase_SQL**.

A separate demonstration database, **Demobase 2D** is available for exploring the *2D gel types* module of InfoQuest FP (see Chapter 6).

Additional example data are available on the installation CD-ROM, in the directory **Sample and Tutorial data**. Alternatively, the same data can be downloaded from the Bio-Rad website (www.bio-rad.com/softwaredownloads). Navigate to the download page and click on "Sample data" in the left menu to display a list with available data. Click on a list item to download the .zip file and unzip the sample data to a destination folder of your choice.

1.4 Installing the InfoQuest FP network software

1.4.1 Introduction

The InfoQuest FP network software is compatible with any TCP/IP supporting network in combination with Windows 2000, Windows NT 4.0, Windows XP and Windows Vista. The communication is based on TCP/IP sockets provided by Windows.

What we call a *server* is the computer that manages the network licenses. The *server* may be any Windows 2000, XP, NT 4.0, or Vista computer in the network. On the server computer, a *server* program called **Netkey** manages the network licenses. The *clients*, running InfoQuest FP, are all computers with the same *Domain Name*, including the server computer. The network software even allows licenses to be granted to physically distant locations via Dial-up connections, provided that the domain name for such distant clients is the same.

The system consists of three components: the *security driver program*, the *security key* (dongle) and the *client software*.

The **security key** is a hardware device (dongle). It attaches to the parallel port or USB port of a computer that is part of the network. This computer will be the *License server*.

The **security driver** is a program, **NETKEY.EXE**, that is available on the server computer. This program manages multiple licensing over the network. It is permanently running as a *Windows Service* on the server computer in the network, i.e. where the security key is attached.

The **client software** is a InfoQuest FP software version that contains the routines needed to register with the security server. This can be installed on *any* computer connected to the network, but only a restricted number of computers, the *license limit*, can run the software at the same time.

NOTE: In a TCP/IP network with Internet access, each computer has its own name in addition to its IP address. These computer names must be valid and registered names for all client computers, since the InfoQuest FP network software uses these names to recognize the client computers. If a Name Server is used, the names of the client computers must be validly registered in the Name Server of the

network, otherwise, license granting will not be possible!

1.4.2 Setup

The License server computer has the *security key* inserted in the LPT1 or USB port and runs the **Netkey server program**, which manages the network licenses. All computers connected to the network can have the InfoQuest FP *application software* installed, but only the number allowed by the *license limit* is able to run the software simultaneously. If the license limit is reached, a new license becomes free whenever the application is closed on one computer.

First, identify a suitable License server computer. The server should be a stable computer in terms of hardware and software configuration, that is permanently working and available over the network to other computers. A computer running Windows 2000/NT, XP Professional or Vista is to be preferred, only for reasons of stability of the operating system.

Once the License server is located, install InfoQuest FP on the License server. When installing InfoQuest FP on the server, you should check the option **Install Netkey server program** in the *Startup* wizard (see Figure 1-7)

After installation of the *server program* on the *server* computer, install InfoQuest FP as a standalone license on one or more client computers in the Network (see Section 1.3). Start perhaps with installing InfoQuest FP on just one client computer and take note of all steps needed to configure the network software. On the server computer, as well as on each of the client computers, the License String should be entered. The License String can be found on the jewel case of the CD-ROM or, in case of upgrades, are obtained electronically.

Do not forget to plug the security key (dongle) into the parallel or USB port of the License server computer!

After installation on the server, the following programs are installed under *Program files > InfoQuest FP*:

- InfoQuest FP
- Netkey

The **Netkey Server** program manages the network licenses. This program runs as a Service that is automatically started on the License server, and should never be halted as long as licenses are in use.

*NOTE: With Windows Vista as operating system, the Netkey program should be ran as administrator in order to run a service. To do so, select **NetKey.exe** in Windows explorer, right-click on it and select "Run as administrator" from the drop-down menu.*


Before the network software can be put into use successfully, there are some settings that will need to be made or changed. **If changes to the network settings of the computers are needed, we recommend to have these changes made by the system administrator or computer expert of your department or institution!**

•TCP/IP

Each computer that will be used in the InfoQuest FP network configuration needs the *TCP/IP protocol* installed on the network. The TCP/IP protocol is provided with the installation package of Windows.


•IP address and DNS host name

Furthermore, each computer in the InfoQuest FP network configuration needs a valid and unique *IP address*, to be specified in the TCP/IP properties. The IP address may be a permanent address assigned to the computer, or an IP address assigned by the DHCP server (Dynamic Host Configuration Protocol). It also must have a valid and unique *DNS host name*, which should not include spaces or periods. The DNS host name can be found by opening *Network* in the Control Panel, selecting *TCP/IP* and clicking *Properties*, under *IP address* and *DNS Configuration*. If permanent, the IP address can be found in the same window. They can also be seen in the InfoQuest FP Startup program by

pressing  and selecting *License settings*. In the *License settings* box that appears, click *Info* to show the computer name, domain name and the IP address. Note down the DNS host names of the client computers and the server computer, and if permanent, also note down the IP addresses.

•Initial settings

On each computer, including the server computer, InfoQuest FP has created a settings file NETKEY.INI, which needs to be completed for the network. Run the InfoQuest FP Startup program on the server, click on

 and select *License settings*. Under *Server computer name*, fill in the DNS host name without the domain name. For example, if a computer is known as **computer.dept.univ.ext**, you should fill in **computer** without the domain name **dept.univ.ext**. You should not change the *Port number* unless there is a conflict with other software that uses the same port number.

You can also edit NETKEY.INI in Notepad by double-clicking on the file name in the Windows Explorer. When opened in Notepad, the contents of the file look as follows:

SERVERNAME=

SERVERPORT=2350

After **SERVERNAME=**, enter the DNS host name of the server computer and save the file. This change must be made on each client computer and on the server computer, in order to allow InfoQuest FP to find the server in the network.

The **SERVERPORT** is the TCP port that is used by the Netkey server and the clients to communicate with each other, and thus should be the same on all computers. In normal circumstances, you can leave **SERVERPORT** unchanged. However, in case there is a firewall between the Netkey server and the clients, or in case the client and/or server computer has a software firewall installed, you will have to open the specified TCP port. Any other port number can be specified, as long as the port number is correctly indicated in the **SERVERPORT** line, both on the Netkey server computer and on the clients.

Start the **Netkey** configuration program on the server computer. The following window appears (Figure 1-8).

Initially, the panel 'Registered computers' is empty and the panel 'Current connected users' does not appear, but a message "Unable to connect to the NetKey service" appears. This is because the service has not been started up yet.

Start the Netkey service by clicking the button **<Start service>**. Since the service is not installed yet, you will be asked to confirm to install it. When this is finished, a message "The NetKey service has been successfully installed" appears.

After clicking **<OK>** to this message, another message tells that the "Service has been started". The bottom panel now lists the currently connected users (empty; see Figure 1-8). The Netkey service is now ready to distribute licenses.

The upper panel lists the computers that are granted access to the InfoQuest FP network. The lower panel lists the computers where the software is currently in use. **Every computer that can get access to the InfoQuest FP network must be specified in the server program, by means of its IP address and DNS host name.** In large institutions, this feature allows control over which computers/users that are allowed to use the InfoQuest FP software.

•Configuring a client

On each client computer, configure the file NETKEY.INI in the same way as described above.

•Defining a client

On the server computer, add the client computer to the list of InfoQuest FP clients as follows: Click **<Add>**. Enter the DNS host name of the client computer in the

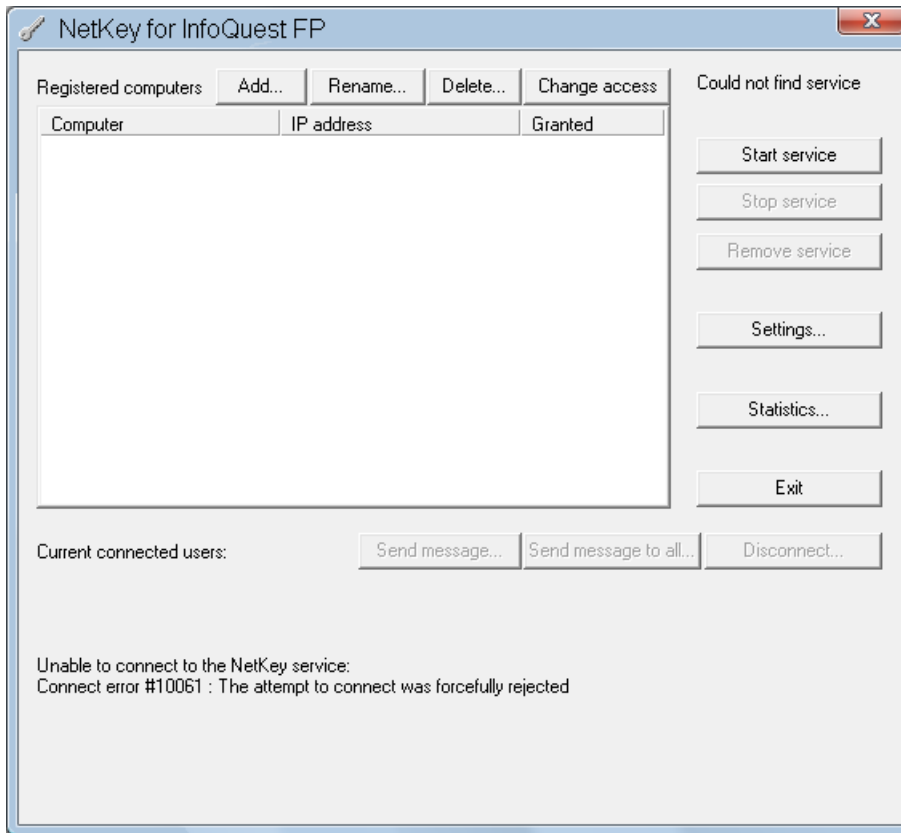


Figure 1-8. The Netkey configuration program, initial view.

dialog box. In non-DHCP configurations (i.e. in case of permanent IP addresses), also enter the IP address. Press **<OK>**. The client is now shown in the upper panel, with its name only (DHCP) or with its name and IP address (permanent). From this point on, the client has access to the InfoQuest FP network software.

NOTE: If you do not wish to define specific computers to have permission to obtain a license for InfoQuest FP, you can enter an asterisk () instead of the computer name, without specifying an IP address. When doing so, every computer in the LAN will be able to obtain a InfoQuest FP license.*

•Running InfoQuest FP

On the client computer, open a database in the Startup program. InfoQuest FP should load if the network is configured correctly and if the server name, the IP addresses and domain host names are filled in correctly.

On the server computer, the client that uses InfoQuest FP is now listed in the lower panel, showing its IP address, DNS host name, total usage time (elapsed) and idle time (1.4.3) (Figure 1-9).

More client computers can be added to the network by simply adding the IP address and the computer name as described in the previous paragraph.

1.4.3 Advanced features of the Netkey server program

The Netkey server program is a Windows Service. As such, it can be seen in the list of installed *Services*. The startup settings, i.e. Manual, Automatic or Disabled, can be specified from the Windows **Services** administration tool (**Control Panel > Administrative tools > Services**) If you close the Netkey configuration program, the service will not be halted. Even when the current user logs off, the service remains running in the background. To effectively shut down the service, click the **<Stop service>** button in the *Netkey configuration* window. If licenses are still in use, the program will produce a warning message, asking you to continue or not.

•License granting

Each computer in the network can be granted or refused access to the application software by the server program. To refuse access to a particular computer, select it in the upper panel, and refuse its access with **<Change access>**. The blue screen icon changes into a red screen. To grant the access again, click **<Change access>** a second time. To permanently remove a computer from the users list, select the computer in the upper panel, and click **<Delete>**.

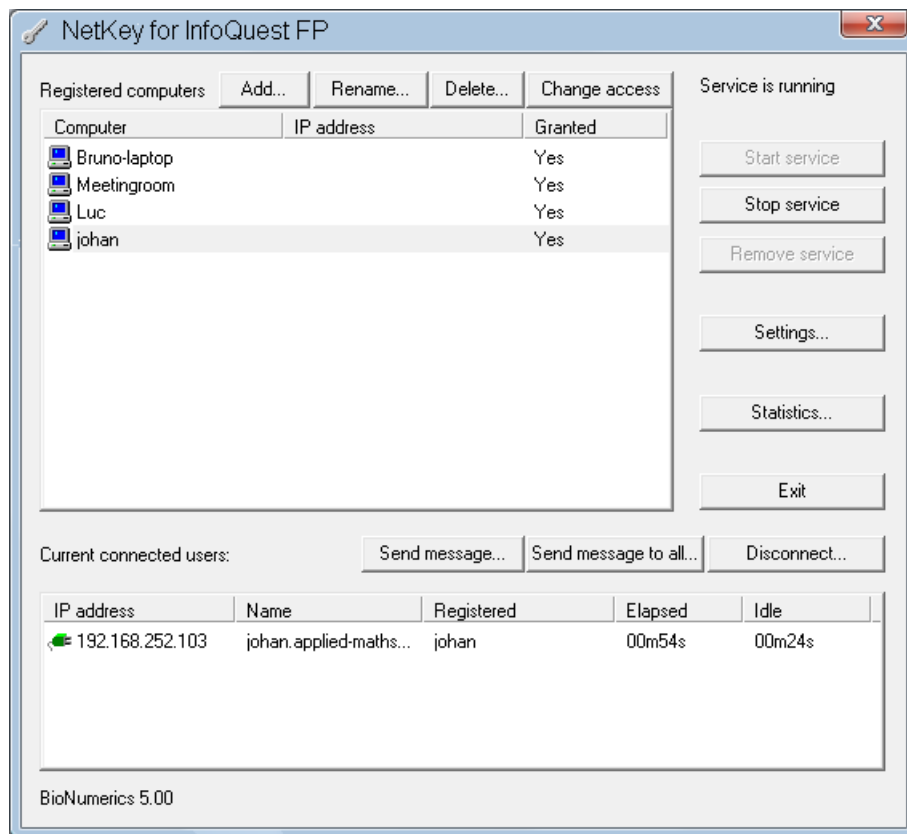


Figure 1-9. The Netkey configuration program, listing all computers that are granted access (top) and licenses in use (bottom).

• Disconnect users

The server can disconnect a client if needed. Select a user in the lower panel, and disconnect it (withdraw its license) with **<Disconnect>**.

• Time-out

The *idle time* of each user is recorded by the Netkey server program. A time-out for inactive licenses can be specified: in case there is a waiting list, a client for whom the idle time exceeds the time-out value will lose his license in favor of the first in the waiting list. Specify a maximum idle time with **<Settings>**, and enter the minutes of idle time. **Note that a user who has exceeded the idle time limit will not be disconnected by the server as long as there is no waiting list.**

• Maximum usage limit

The *usage time* by each client is recorded by the Netkey server program; it is the total connection time of the current session. A maximum usage time can be specified: in case there is a waiting list, a client for which the usage time exceeds the maximum usage time will lose his license in favor of the first in the waiting list. Specify a maximum usage limit with **<Settings>**, and enter the minutes of usage time. **Note that a user who has exceeded the maximum usage time limit will not be disconnected by the License server as long as there is no waiting list.**

• Messaging

The License server can send messages to any or all connected clients, for example in case the server computer will be shut down or if a client will be disconnected. Send a message to one user by selecting the user in the lower panel and press **<Send message>**. Enter a message string and press **<OK>**. The user will receive the message in a dialog box. Send a message to all users with **<Send message to all users>**. Enter a message string and press **<OK>**. All active users will receive the message in a dialog box.

• Usage statistics

The Netkey server program records every usage of each client. Graphical statistics can be displayed about the history of the usage over longer periods, and the relative usage of each client computer can be shown for any time interval. To view the usage history of the InfoQuest FP network version, click **<Statistics>**. The panel shows a detailed view of the number of computers that have used the software on a time scale divided in hours. You can scroll in this panel to view back in the past. The license limit is shown as a red line; computers in a waiting list are shown in red. The relative usage of each client computer can be shown by clicking the **<Relative usage>** tab. Enter the time period (from-to) in Days / Months / Years. The result is a circle diagram with the percentage usage time for each computer shown.

1.4.4 Features of the client program (InfoQuest FP)

•Waiting lists

In case the maximum license number is exceeded, the server program manages a waiting list. The client receives a message with its number in the waiting queue, and the InfoQuest FP software pops up as soon as the client's license becomes available.

The user can request an overview of the computers currently using a InfoQuest FP license by pressing



in the InfoQuest FP Startup program, selecting *License settings* and then clicking *<Status>*. It shows for each connected computer the IP address, the computer name, the total usage time and the idle time.

•Disconnection by server or license loss

If the client is disconnected by the server or loses its license, e.g. due to idle time or maximum usage limit, a warning box flashes that you should save any unsaved data and quit the program immediately. InfoQuest FP tries four times again to negotiate its license with intervals of 15 seconds. After the fourth time (1' 15" in total), the program halts automatically.


1.5 Starting and setting up InfoQuest FP


1.5.1 The InfoQuest FP Startup program

1.5.1.1 Double-click the “InfoQuest FP” icon on the desktop to run the **Startup program**. This program shows the *Startup screen* (see Figure 1-10). It allows you

to run the InfoQuest FP main application with ,

to create new databases () , and customize

various settings () such as the home directory (with *Change home directory*), the directory of a selected database (with *Database settings*) and the license and network settings (with *License settings*).

1.5.1.2 Use the  button when you are finished running the InfoQuest FP applications.

1.5.2 Creating a database

In order to facilitate the use of InfoQuest FP in different research projects, it is possible to set up *Databases*. The principles of a database are explained in . The InfoQuest FP Startup program will look for all databases in one *home directory*, specified by the user. Note that, in Windows 2000, Windows XP, and Windows Vista, each Windows user may specify a different home directory. The InfoQuest FP home directory is saved with the system registry of the user.

If you want to change the current home directory follow the steps below:

1.5.2.1 In the Startup screen, press the settings button


() and select *Change home directory*. This pops up the *Home directory* dialog box (Figure 1-11).



Figure 1-10. The InfoQuest FP Startup screen.

The program offers two default options for the location of the home directory: *In Common Documents* and *In My Documents*. The option *In Common Documents* makes the InfoQuest FP databases available for all Windows users on that computer. The option *In My Documents* makes the databases only accessible to the user currently logged on. The third option, *Custom folder*, allows the user to specify any desired directory. You can even specify a directory on a network drive, on condition that this drive is permanently available with write-access.

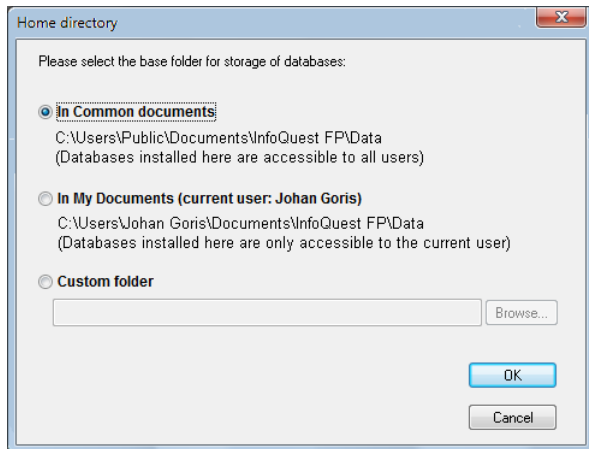



Figure 1-11. The *Home directory* dialog box.

1.5.2.2 Press **<OK>** to select the new home directory. The program updates the list of available databases in the new directory.

Create a new database as follows:

1.5.2.3 Press the  button to enter the *New database wizard*.

1.5.2.4 Enter a name for the database, e.g. **Example**, and press **<Next>**.

NOTES:

(1) The program automatically creates a folder within the current home directory for storage of the database-specific files and folders (in this example: **[HOMEDIR]\Example**). If you want to change this (not recommended), press **<Browse>**. This option allows you to select any directory from any permanent drive. It is recommended to create a new empty directory before you choose it as database directory.

(2) If you do not want the program to automatically create subdirectories, click **No** to this question (not recommended, unless the subdirectories already exist). In that case, you will have to create the subdirectories manually (see Figure 2-3).

1.5.2.5 Press **<Next>** again.

1.5.2.6 You are now asked whether or not you want to create log files. If you enable InfoQuest FP to create log files, every change made to a database component (entry, experiment etc.) is recorded to the log file with indication of the kind, the date, and the time of change. For more information on log files, see 2.1.6.

1.5.2.7 Press **<Finish>** to complete the setup of the new database.

1.5.2.8 A new dialog box pops up, prompting for the type of database (see Figure 1-12).

InfoQuest FP offers two alternative database solutions to store its databases: the program's own built-in database (= **local database**) or an external SQL and ODBC compatible database engine. The latter solution is called a **connected database**. Because of the multi-user access and the extended range of features only available in connected databases, the creation of a connected database is the default option in InfoQuest FP. The use of connected databases and the different options (see Figure 1-12) are discussed in .

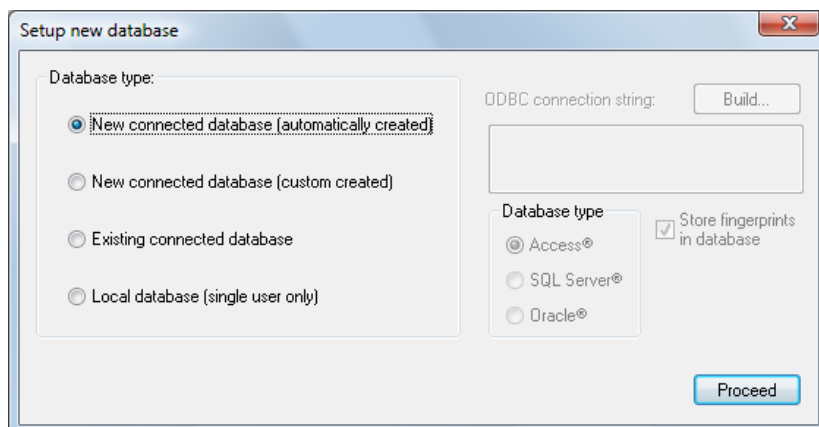


Figure 1-12. *Database selection* dialog box.

1.5.2.9 For this example, leave the default option *New connected database (automatically created)* enabled and press **<Proceed>**.

1.5.2.10 The *Plugin installation* toolbox appears. For more information on the installation of plugins, see 1.5.3.

1.5.2.11 Press **<Proceed>** to start working with the newly created database without installing any plugins.

1.5.3 Installing plugin tools in a new database


When a new database is opened for the first time (1.5.2.10), InfoQuest FP will provide the opportunity to install *Plugin tools*. Plugins are tools written in the InfoQuest FP script language, available as binary encoded packages. The plugins offer extra functionality, often to import or export various types of data, but also to deal with specific applications, such as multi-locus sequence typing (MLST), variable number of tandem repeats (VNTR) analysis, etc. Plugins can also provide extra functions related to dendrogram analysis, statistics, database management, etc.

1.5.3.1 Open the database **DemoBase** by selecting **DemoBase** in the Startup program, and click on the



button. Simply double-clicking on the database name does the same.

If this is the first time the database **DemoBase** is opened, the *Plugin installation* toolbox will appear, as illustrated in Figure 1-13. If not, the database will open without showing this toolbox. In that event, you can still open the *Plugin installation* toolbox from the *InfoQuest FP main* window (see Figure 1-15) with **File > Install/Remove**

plugins or by pressing the  button.

The listbox (left) shows the plugin tools that were delivered with the installation CD.

1.5.3.2 With **<Check for updates>** the latest versions and/or new plugins are downloaded from the Bio-Rad website and shown in the listbox. Installing new versions of the plugins might require administrator rights on your computer.

1.5.3.3 When a particular plugin is selected, a short description appears in the right panel along with a version number.

1.5.3.4 A selected plugin can be installed with the **<Install>** button.

1.5.3.5 Once a plugin is installed, it is marked with a green V-sign. It can be removed again with the **<Remove>** button.

1.5.3.6 If a manual exists for the plugin, a PDF manual is opened in Adobe Acrobat Reader with **<Manual>**.

1.5.3.7 If you have finished installing plugins, you can proceed to the *InfoQuest FP main* window with **<Proceed>**.

When installed, a plugin installs itself in a menu of the software, and is characterized by a plug icon left from the menu item. For example, if **"Fingerprint processing reports"** is installed, a new item **Print TIFF image** becomes available in the **File** menu of the *Fingerprint data* window (see Figure 1-14).

At the time of writing of this manual, the following plugins were available:


- **2D gels plugin** (

Figure 1-13. The *Plugin installation* toolbox.

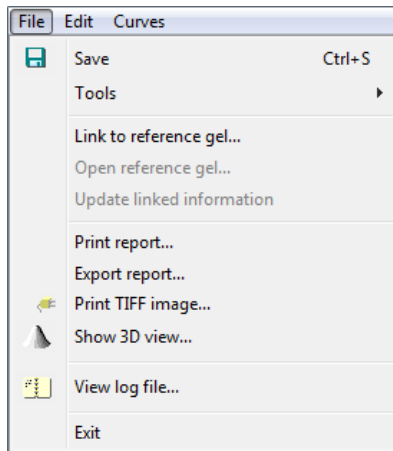


Figure 1-14. Plugin menu commands are characterized by a yellow plug icon left from the menu item.

Intermediate or Resistant) based on user-defined cut-off values for disk diffusion (zone diameter) tests and Minimum Inhibitory Concentration (MIC) values.

- **Batch sequence assembly plugin (SQ)**: A tool for the automatic assembly of trace files into (multiple) contigs. Allows parsing of entry keys and experiment names from the trace file names or from template files and displays an interactive error report.
- **InfoQuest FP Help plugin**: Contains a compiled help function for working with InfoQuest FP, opens a PDF version of this manual and offers an easy link to the online FAQ database and the support question webform.
- **Database tools**: Offers additional search functions (fuzzy search, find and replace) and database layout tools.
- **Dendrogram tools (CL)**: Contains tools for working with dendrograms in the *Comparison* window.
- **Fingerprint processing reports (FP)**: Contains tools for exporting data from the *Fingerprint processing* window.

- **Geographical plugin (DS)**: Plots database entries on a geographical map, e.g. for epidemiological investigations.
- **HDA plugin (FP, CH)**: Contains all necessary tools for automated Hetero Duplex Analysis (HDA) or Conformational Specific Capillary Electrophoresis (CSCE) analysis.
- **Import plugin**: Offers various convenient import routines for import of characters, sequences, trend data, sequencer fingerprint files, etc.
- **MLST plugin (SQ, CH)**: Offers the tools to automatically set up Multi-Locus Sequence Typing (MLST) experiments. Offers the option to link interactively to the PUBmlst.org website.
- **Sequence translation tools (SQ)**: Translates nucleic acid sequences into amino acid sequences.
- **Spa Typing plugin (SQ, CH)**: Provides the functionality to perform spa-typing on *Staphylococcus aureus*.
- **VNTR plugin (FP, CH)**: Offers the tools to import and analyze Variable Number of Tandem Repeat (VNTR) and Multi-Locus VNTR Analysis (MLVA) data.
- **XML Tools plugin (DS)**: Allows the import and export of database information as XML files.

IMPORTANT: Always check the *Plugin installation* toolbox after pressing **<Check for updates>** to find the latest versions and the most complete list of available plugins. Pressing the **<Check for updates>** button will update ANY plugin for which a newer version is available.

Some plugins depend on the functionality offered by specific modules (see 1.1.5), e.g. the Spa typing plugin requires the *Sequence types* and *Character types* modules to be present. If one of the required modules is missing, the plugin cannot be installed and an error message is generated.

1.6 The InfoQuest FP user interface

1.6.1 Introduction to the InfoQuest FP user interface

InfoQuest FP is a very comprehensive software package. For every specific task in InfoQuest FP, a **window** is available that groups the relevant functionality for that specific task. From an active window, dialog boxes and sub windows can be launched. Each window at its turn consists of several **panels** containing specific information. The InfoQuest FP user interface is very flexible and can be customized by the user at three different levels:

1. The behavior and general appearance of windows can be set.
2. For each separate window, the toolbars and panels that are displayed can be chosen, as well as the size and the location of panels.
3. In grid panels, columns can be displayed or hidden and the relative position of rows and columns can be chosen.

The InfoQuest FP user interface will be explained using the example database provided with the software, called **DemoBase** (see 1.3.2 for a description of the data available in this database).

NOTE: In comparison with previous versions, the user interface of InfoQuest FP 5.x is completely renewed. The major changes include: direct editable information fields, control over how the windows stack appears, preset color schemes, font types, dockable panels, toolbars that can be displayed or hidden, and zoom sliders. Even for the experienced InfoQuest FP user, it might be useful to read this chapter to take full advantage of these new features.

1.6.2 The InfoQuest FP main window

1.6.2.1 Open the database **DemoBase** by selecting **DemoBase** in the Startup program, and click on the



button. Simply double-clicking on the database name does the same.

If the *Plugin installation* toolbox appears instead of the database (Figure 1-13), it means that you are opening this database for the first time. See 1.5.3 for further explanation on the installation of plugin tools.

The *InfoQuest FP main window* in default configuration (Figure 1-15) consists of a menu, a toolbar for quick

access to the most important functions, a status bar, and the following eight panels:

- The *Database entries* panel, listing all the available entries in the database, with their information fields and their unique keys (see 1.1.2). A local InfoQuest FP database can contain up to 200,000 entries. A connected database can contain many more entries, but only 200,000 can be displayed in one view.
- The *Experiments* panel, showing the different experiment types, and the experiments that are defined under each type.
- The *Experiment presence* panel, which for each database entry shows whether an experiment is available (colored dot) or not. Clicking on a colored dot causes the *Experiment card* for that experiment to be popped up (see Figure 3-84).
- The *Files* panel, showing the available data files for the experiment type selected in the *Experiments* panel, with their date of creation, the date when the files were last modified and their location (Local or Shared for local database or connected database, respectively).
- The *Comparisons* panel, listing all comparisons that are saved, with their date of creation, the date when the comparisons were last modified and their location (Local or Shared for local database or connected database, respectively).
- The *Libraries* panel, which shows the available identification libraries and their location (Local or Shared for local database or connected database, respectively).
- The *Decision networks* panel, which shows the available decision networks, with their date of creation and the date when the networks were last modified. Decision networks can also be executed on selections of database entries from within this panel.
- The *Entry relations* panel, listing the available entry relations with their forward and reverse name and the database levels they relate to.
- The *Alignments* panel, listing all alignment projects that are saved, with their date of creation and the date when the alignment projects were last modified.

NOTES:

- (1) In default configuration, the *Entry relations panel* appears as tabbed view together with the *Experiments*

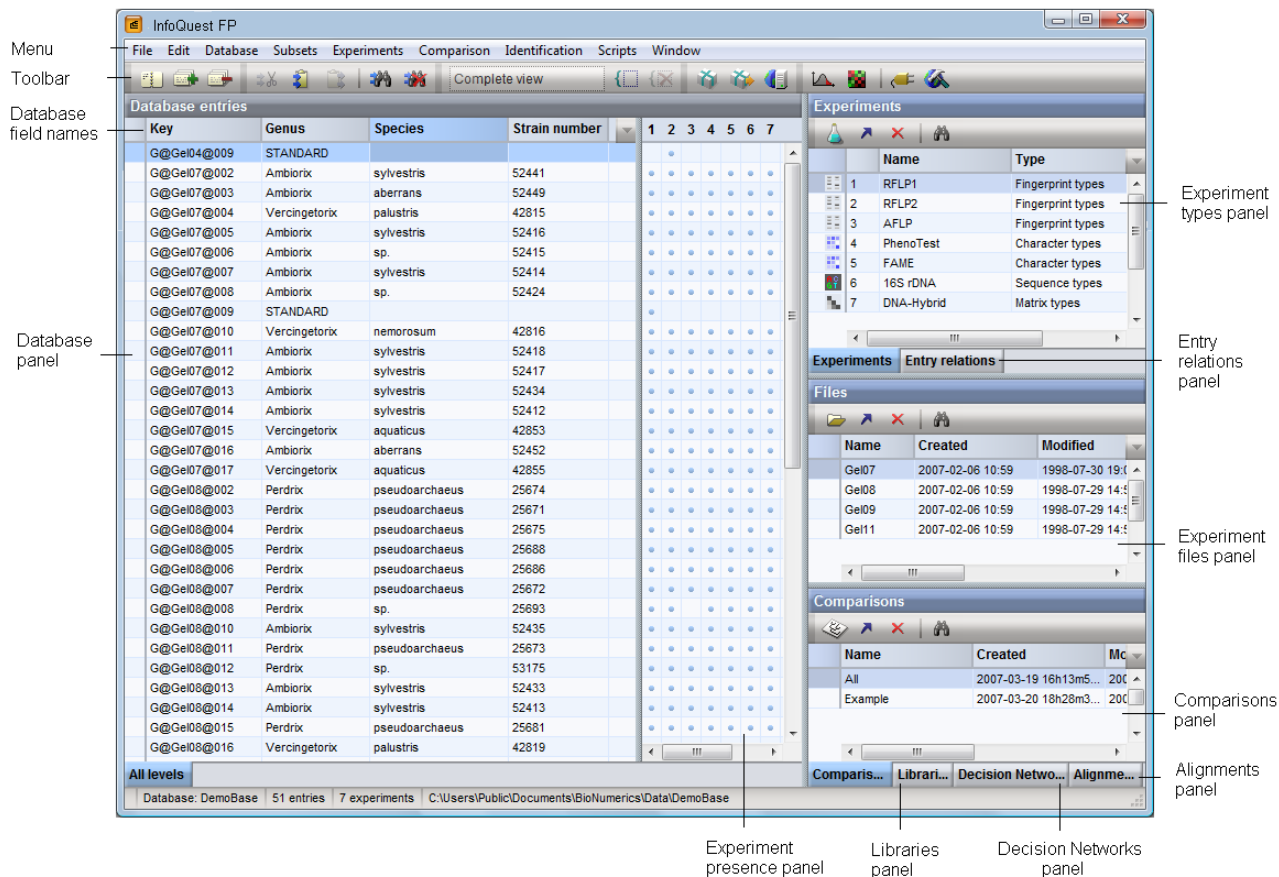


Figure 1-15. The *InfoQuest FP* main window.

panel, with the Experiments panel displayed. Likewise, the Comparisons, Libraries, Decision networks and Alignments panels are grouped as a tabbed view, with the Comparisons panel displayed by default.

(2) Unless otherwise stated, all screenshots in this manual are taken using default settings, but with the blue color scheme applied. If the *InfoQuest FP* main window is displayed differently on your screen than in the screen shot in Figure 1-15, then your current settings might be different from the default settings. How to return to the default settings is described in the next paragraphs.

1.6.3 General appearance of InfoQuest FP windows

InfoQuest FP offers the choice between eight preset color schemes, allows the adjustment of brightness and visual effects and lets the user select the font type. These general appearance settings have an effect on **all** InfoQuest FP windows.

1.6.3.1 In the *InfoQuest FP* main window, select **File > Preferences**. The *Preferences* dialog box appears (Figure 1-16). From the list on the left side of the dialog box, select **Windows appearance**.

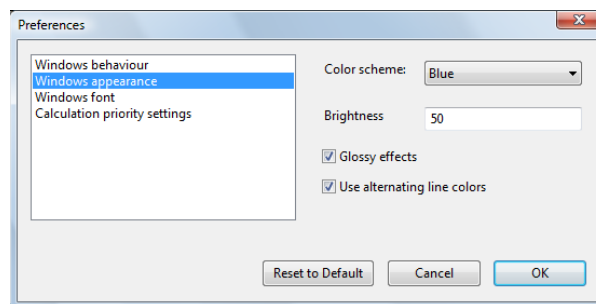


Figure 1-16. The *Preferences* dialog box, **Windows appearance** settings.

From the pull-down list next to **Color scheme**, a selection can be made from eight preset color schemes. The **Brightness** can be adjusted by entering a percentage. According to your own preferences, **Glossy effects** and **Use alternating line colors** can be either checked or unchecked.

1.6.3.2 Select any color scheme of your choice, try entering a different brightness percentage and/or uncheck any of the visual effects to notice their effect. Press **<OK>** to display the modified appearance of the *InfoQuest FP* main window and other InfoQuest FP windows. It might be necessary to restart InfoQuest FP to enable the applied changes.

1.6.3.3 Call the *Preferences* dialog box again with *File > Preferences* and select *Windows appearance* from the list. Pressing *<Reset to Default>* will restore the default appearance settings: Orange and Gray color scheme, brightness 50% and both *Glossy effects* and *Use alternating line colors* enabled. Press *<OK>* to have the InfoQuest FP windows displayed with default settings again. It might be necessary to restart InfoQuest FP to enable the applied changes.

1.6.3.4 In the same *Preferences* dialog box, clicking on *Windows font* enables you to set the type and size of the font used in all panels. Pressing *<Reset to Default>* will restore the default font settings: font Arial, size 11.

NOTE: All preferences are saved at a database level, allowing the user to specify different preference settings for different databases.

1.6.4 Display of panels

InfoQuest FP windows can be customized up to a high degree by the user. All panels can be resized to make optimal use of the display by dragging the horizontal separators between the panels up or down or by dragging the vertical separators left or right.

Two types of panels are available in InfoQuest FP windows: fixed panels and dockable panels. The displayed information in **fixed** panels is indispensable in any type of analysis. Therefore, these panels are always displayed in their corresponding window. Depending on the nature of the experiment, the type of analysis performed and user preferences, **dockable** panels may not always be essential. Therefore, InfoQuest FP offers the possibility to either display or hide dockable panels. This feature allows the user to hide infrequently used panels that would otherwise clutter the workspace. For example, if you do not have the *Identification* module, the *Libraries* panel in the *InfoQuest FP main window* can be hidden for the sake of clarity. Several InfoQuest FP windows contain dockable panels, which all behave identically. The principles are illustrated here for the *InfoQuest FP main window*.

1.6.4.1 In the *InfoQuest FP main window*, click on the *Libraries* tab to display the *Libraries* panel.

1.6.4.2 Select *Window > Show / Hide panels* in the *InfoQuest FP main window*. This displays a submenu, listing all available panels. Panel names for which a check mark is present left of the menu item are shown in the *InfoQuest FP main window*.

1.6.4.3 Click on *Libraries* in the submenu. The *Libraries* panel is now hidden from the *InfoQuest FP main window*.

Furthermore, dockable panels can be placed on the screen in one of two modes: floating or docked. **Floating** allows the window to be placed anywhere on the screen, similar to a normal window of a base size (not maxi-

mized). The **docked** mode automatically places the panel in one of five locations: top, bottom, left, right, or stacked onto another panel (tabbed view). The position of a panel is controlled with a **docking guide**.

1.6.4.4 Click in the header of the *Files* panel and - while keeping the mouse button pressed - drag it upwards in the window. As soon as the panel leaves its original position, a docking guide appears in the center of the *Experiments* panel. Release the mouse button on any place next to the docking guide to leave the panel floating in the window.

A floating window can be repositioned to any place on your monitor.

1.6.4.5 Click in the header of the *Files* panel and drag it towards the *Experiments* panel again. Drop the floating panel on the top part of the docking guide that appears (see Figure 1-17). This action will make the *Files* panel appear above the *Experiments* panel in the *InfoQuest FP main window*.

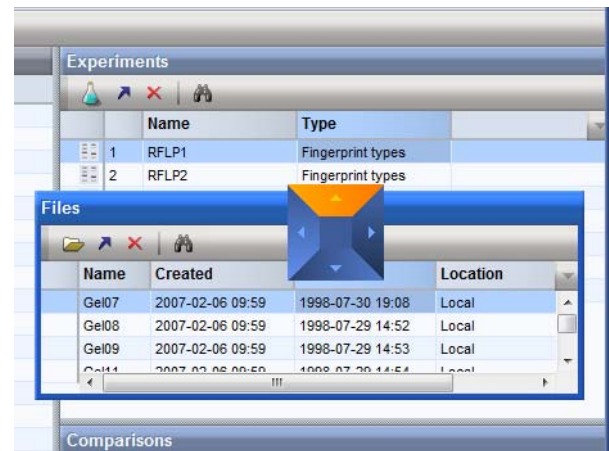


Figure 1-17. Docking the *Files* panel above the *Experiments* panel using the docking guide.

1.6.4.6 Click in the *Files* panel header and drag it towards the *Experiments* panel again. This time, drop the *Files* panel on the center of the docking guide (see Figure 1-18). As a result, the *Files* panel is now displayed as a tabbed view with the *Experiments* and *Entry relations* panel (see Figure 1-19).

NOTE: To re-locate a panel that is presently displayed as a tabbed view with other panels, click on the panel tab instead of the panel header to drag the panel to its new position.

After making some changes to the window configuration, it is always possible to return to the default configuration for the active window. This might be useful e.g. to make comparison with screenshots shown in this manual easier. If you intend to revert to the user-defined configuration afterwards, then you can save the user-defined configuration first and recall it afterwards.

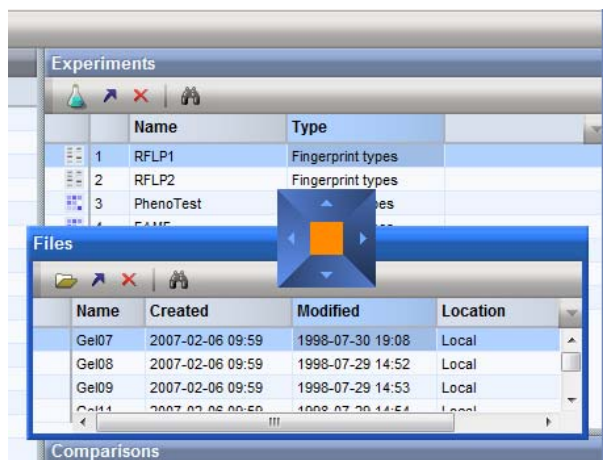


Figure 1-18. Docking the *Files* panel as a tabbed view with the *Experiments* and *Entry relations* panels.

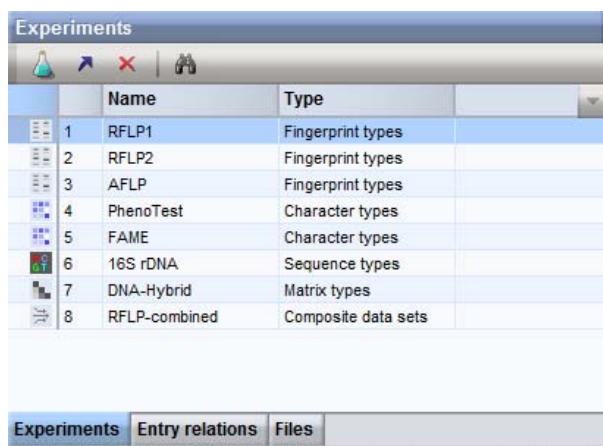


Figure 1-19. Result of the action depicted in Figure 1-18: tabbed view of the *Experiments*, *Entry relations* and *Files* panel.

1.6.4.7 Select *Window > Save current configuration* to store the configuration that you have just defined.

1.6.4.8 To restore the default configuration, select *Window > Restore default configuration*. The window now appears back in its original configuration.

1.6.4.9 Recall the user-defined configuration with *Window > Recall saved configuration*. Notice that the changes you made to the window configuration are introduced again.

In case you do not wish to save the introduced configuration changes, steps 1.6.4.7 to 1.6.4.9 can be skipped:

1.6.4.10 Select *Window > Restore default configuration* to restore the default configuration of the *InfoQuest FP* main window again.

Any window configuration can be protected from accidental changes via *Window > Lock configuration*. A check mark is present in the menu left of *Lock configu-*

ration if the configuration of the active window is locked. Configuration changes will be enabled if *Window > Lock configuration* is selected again.

1.6.5 Configuring toolbars

In addition to the pull-down menu's that are available for executing the commands (see Figure 1-20 for an example), InfoQuest FP also displays toolbars for

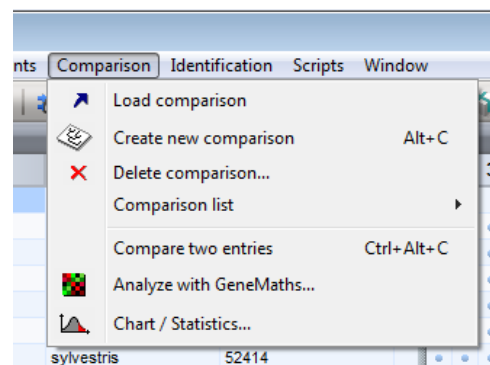



Figure 1-20. The pull-down menu *Comparison*.

frequently used commands. Toolbars consist of buttons that are arranged in groups, according to their function.

An example is the database toolbar,  which is located in the header of the *InfoQuest FP* main window. Many panels have their own toolbar, grouping panel-specific commands. Toolbars can either be displayed or hidden, a feature which allows the user to hide infrequently used toolbars.

NOTE: When using Microsoft Windows Vista as operating system, the corresponding toolbar icons appear left of the menu items as a visual aid (see Figure 1-20). Since earlier operating systems do not support this feature, toolbar icons in the pull-down menus may not be displayed on your computer screen.

1.6.5.1 In the *InfoQuest FP* main window, select *Window > Show / Hide toolbar* and, for example, click on *Database* to hide the Database toolbar.

When the toolbar is displayed, a check mark is present next to the corresponding menu item. Toolbars specific to certain panels are listed under the corresponding submenu's.

1.6.5.2 Select *Window > Show / Hide toolbar > Files panel* and click on *File tools* to hide the toolbar specific for the *Files* panel.

The position of a toolbar within a window or panel can also be altered.

1.6.5.3 In the header of the *InfoQuest FP* main window, click on the dark gray area in a toolbar, left of a set of buttons. The mouse pointer will take the shape of a hand

on top of two arrows. Drag the toolbar left or right to change the order in which the toolbars appear.

1.6.5.4 In the header of the *InfoQuest FP main* window, drag another toolbar slightly downwards to make the toolbars appear in two rows.

1.6.5.5 Click on a toolbar again and drag it to the left (or right or bottom) part of the window to dock it on the left (or right or bottom) part of the window.

The position of panel-specific toolbars can be customized much in the same way as general toolbars, with the restriction that they cannot be positioned outside their corresponding panel.

Individual buttons can be hidden from their toolbars.

1.6.5.6 Right-click on any toolbar. A floating menu appears, listing all buttons of the toolbar (see Figure 1-21). By default, all button names are checked in the menu and the corresponding buttons will appear in the toolbar.

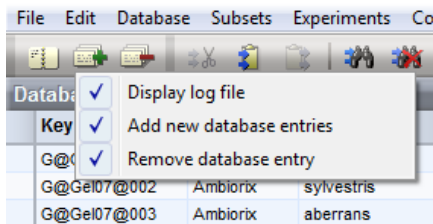


Figure 1-21. Floating menu for the Database toolbar, listing all available buttons.

1.6.5.7 Select the button that you want to hide from the floating menu.


The toolbar button can be displayed again by repeating the above actions.

As for the display of panels, the configuration of toolbars can be restored to default using *Windows > Restore default configuration*. In case you want to save the current configuration, follow steps 1.6.4.7 to 1.6.4.9.

1.6.5.8 Select *Window > Restore default configuration* to restore the default configuration of the *InfoQuest FP main* window again.

1.6.6 Grid panels

Grid panels contain data in tabular format, i.e. organized in columns and rows. In the *InfoQuest FP main* window, examples are the *Database entries*, *Experiments*, *Files*, and *Comparisons* panels. All grid panels can be customized up to a high degree by the user. The width of columns can be changed by moving the separator line in the column heading left or right. Other column properties can be accessed via the column properties button

, located on the right hand side in the information fields header. The column properties of the *Files* panel are illustrated in Figure 1-22. This pull-down menu contains a list of information fields that are available in the grid panel and which can either be displayed or hidden (check mark resp. present or absent).

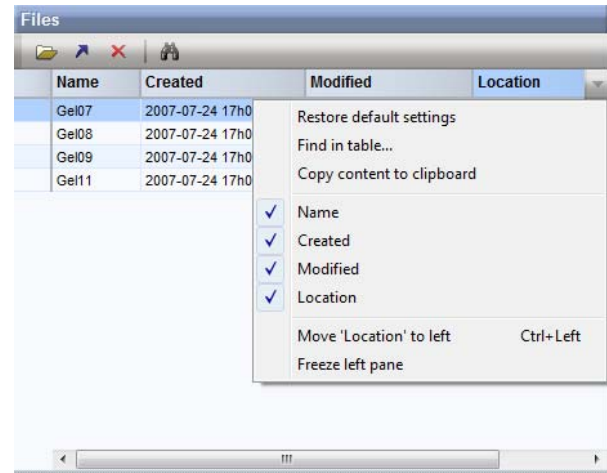




Figure 1-22. Column properties of the Files panel.


1.6.6.1 Click on the column properties button () of the *Files* panel (see Figure 1-22) and click on *Created* to hide the column displaying the date of creation.

The relative position of a selected column within the panel can be changed using the menu items *Move 'column_name' to left* and *Move 'column_name' to right*.


1.6.6.2 In the *Files* panel, select the column 'Location' and click on the column properties button (). Click *Move 'Location' to left* or *Move 'Location' to right* to shift the 'Location' column to the left or right, respectively. The shortcut keys CTRL+left arrow and CTRL+right arrow can be used for the same purposes.

The option *Freeze left pane* allows the user to freeze one or more information fields so that they always remain visible left from the scrollable area.

Similar as for the window configuration (see 1.6.4.10), it is possible to revert to the default column properties settings for the active panel.

1.6.6.3 Click on the column properties button () of the *Files* panel and select *Restore default settings* to disable all introduced changes to the column properties of the *Files* panel.

A grid panel can be searched for the occurrence of a search string in any displayed information field:

1.6.6.4 Click on the column properties button () of the *Database entries* panel and select *Find in table*. This pops up the *Find in table* dialog box (see Figure 1-23).

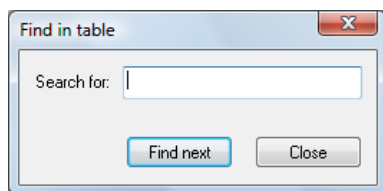



Figure 1-23. The *Find in table* dialog box.

1.6.6.5 Enter a search string, e.g. “verc” and press <Find next> repeatedly to find all occurrences of the entered search string in the displayed information fields (in this case, all Vercingetorix in the ‘Genus’ field).


The information contained in any grid panel can be exported to the clipboard for use in other programs:

1.6.6.6 Click on the column properties button () of the *Database entries* panel and select *Copy content to clipboard*.

1.6.6.7 Paste the content of the clipboard in e.g. Notepad to view the information that was copied from the *Database entries* panel.

In grid panels, rows can be sorted according to a certain field by right-clicking on the field (column) header and selecting *Arrange ... by field*.

A number of predefined information fields are automatically created when creating a new database.

1.6.6.8 Select the column properties button () of the *Database entries* panel.

The predefined information fields listed in the pull-down menu (see Figure 1-24) can either be displayed or hidden (check mark resp. present or absent).

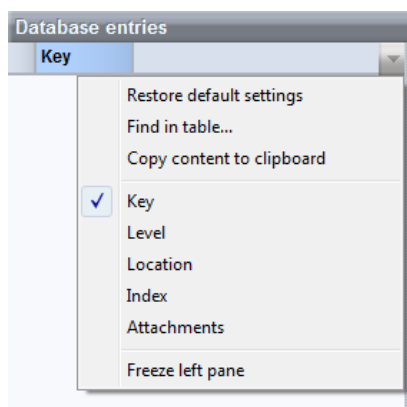


Figure 1-24. Automatically created information fields in the *Database entries* panel.

NOTE: For a local database, the default information fields Created and Modified follow standard Microsoft Windows behavior. Therefore, if a file is copied from a different location, the moment of copying is taken as Created date and will be more recent than the date displayed in Modified.

In addition to these default information fields, extra information fields can be added with *Add new information field* or removed with *Remove information field*. The menu commands can be accessed in each panel by right-clicking in their information field headers.

Information in non-default information fields can be edited by clicking twice (not double-click) on an item, or by pressing CTRL+Enter on the keyboard. The information then appears selected blue against a bright colored background and can be modified. This is illustrated in Figure 1-25, where information about gel staining is added to the *Files* panel.

Name	Created	Modified	Location	Staining
Gel07	2007-02-06 09:59	1998-07-30 19:08	Local	EtBr
Gel08	2007-02-06 09:59	1998-07-29 14:52	Local	EtBr
Gel09	2007-02-06 09:59	1998-07-29 14:53	Local	SybrGreen
Gel11	2007-02-06 09:59	1998-07-29 14:54	Local	

Figure 1-25. Editing information within non-default information fields, illustrated for the *Files* panel. Clicking twice on an information fields enables direct editing.

The *Database entries* panel behaves just as other grid panels, with a few peculiarities. As in earlier versions of the software, double-clicking on a database entry or pressing Enter opens its *Entry edit* window (see Figure 1-26). More information on this window is provided in .

If desired, the direct field editing behavior can be modified:

1.6.6.9 Select *File > Preferences* to call the *Preferences* dialog box and click on *Windows behaviour* in the list on the left hand side.

1.6.6.10 Check or uncheck *Single click field editing* to enable or disable information field editing after a single mouse click.


NOTE: When Single click field editing is enabled, the Entry edit window (see Figure 1-26) is opened by double-clicking in the margin or on the Key field of the database entry.

For information fields in the *Database entries* panel, properties can be set. One of these properties includes a different background color for each field state. The



extend of the background color can be set in the *Preferences* dialog box.

1.6.6.11 Select *File > Preferences* to call the *Preferences* dialog box and click on *Windows behaviour* in the list on the left hand side.

1.6.6.12 Uncheck *Use color background for complete field* (checked by default) to limit the background color to a small rectangle, preceding the information field content.

In the panels *Experiments, Files, Comparisons, Libraries, Decision networks* and *Alignments* from the *InfoQuest FP* main window, pressing the  icon calls the *Field query* dialog box (Figure 1-27). This allows the list of available experiments, files, comparisons or libraries to be searched for name and any user-defined information field (if present). Items that match the search criteria are marked with a small colored triangle in the left column. When the option *Bring selected to top* is checked, the selected items appear on top of the list. Further options are available to replace currently selected items with new items, add the newly found items to the list, to search within the selection and to use regular expressions (see Section 7.2 on how to use regular expressions) in the search string.

1.6.7 Zoom sliders

In many *InfoQuest FP* windows, those panels containing graphical information can be zoomed in or out to make optimal use of the display. Zooming in or out can be done via the  and  buttons in the toolbar or via the corresponding menu commands. Shortcut keys

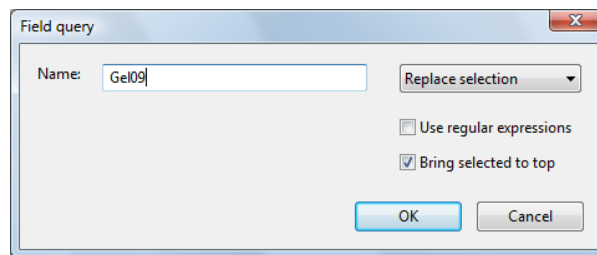





Figure 1-27. *Field query* dialog box for the *Files* panel.

for these actions are CTRL+PageUp and CTRL+Page-Down, respectively.

In addition, graphical panels or windows in *InfoQuest FP* are equipped with *zoom sliders* in the shape of a narrow vertical or horizontal pane, featuring a colored bar (see Figure 1-28 for an example). Increasing the bar size, by dragging it with the mouse, zooms in on the image. Decreasing the bar size with the mouse zooms out on the image. The zoom slider can also be operated by hovering over it and using the scroll wheel of the mouse. Alternatively, press the CTRL or SHIFT (in case more than one zoom slider is present) key on the keyboard and use the scroll wheel. Image proportions are maintained when (as in Figure 1-28) the  icon is displayed in the zoom slider. When the  and  icons are shown in the zoom sliders, horizontal and vertical zooming can be performed separately. The gray line in the zoom slider bar corresponds to the original image size (x 1.00). Similar to toolbars, the position of the zoom sliders (left, right, top or bottom) can be changed by clicking on the area above the zoom icons and dragging the zoom sliders in position with the mouse.

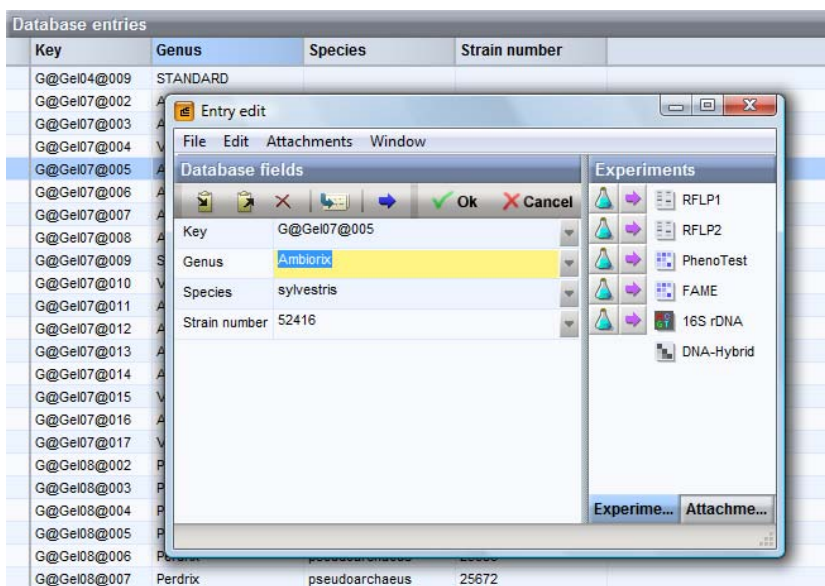


Figure 1-26. Using the *Entry edit* window to modify information fields.

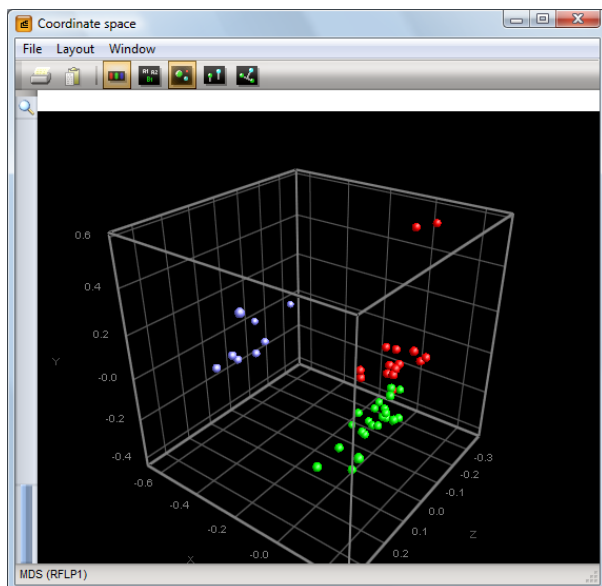


Figure 1-28. Zoom slider (left), illustrated for the *Coordinate space* window.

1.6.8 Behaviour of InfoQuest FP windows

1.6.8.1 Open the database **DemoBase** by selecting **DemoBase** in the Startup program, and click on the



button. Simply double-clicking on the database name does the same.

Via *File > Preferences > Windows behaviour*, the behaviour of the InfoQuest FP windows stack can be controlled. The windows stack can be set either *Fixed* or *Flexible*. In *fixed* mode, the various InfoQuest FP windows are stacked in a fixed order. For example, a *Comparison* window always appears on top of the *InfoQuest FP main* window and in order to view the complete *InfoQuest FP main* window, the *Comparison* window needs to be closed or minimized. Furthermore, in the Taskbar of your operating system, only one tab will appear for the InfoQuest FP software. In *flexible* mode, any type of InfoQuest FP window can be on top of another, regardless of its “rank”. For each InfoQuest FP window, a separate tab becomes available in the Taskbar of your operating system.

1.6.9 Navigator pane

For both fixed and flexible mode of the windows stack, a Navigator pane is available (Figure 1-29). The Navigator

pane displays all open InfoQuest FP windows in a tree-like hierarchical structure to facilitate navigation between windows. The active window is shown in orange type, inactive windows are shown in white type. Under default settings, the Navigator pane is enabled and appears when moving the mouse to the far right side on the screen.

The position of the Navigator pane on the computer display (top, bottom, left or right) can be modified:

1.6.9.1 Click on the structured part in the Navigator pane and drag it to the desired position with the mouse.

Other display properties of the Navigator pane can be set by right-clicking in the structured part and selecting them from the drop-down menu:

1.6.9.2 Right-click on the structured part of the Navigator pane and uncheck *Always on top* if you want the Navigator pane to appear *stacked* with other open windows instead of *on top* of all open windows.

1.6.9.3 Uncheck *Auto hide* if you want the Navigator pane to be permanently displayed.

1.6.9.4 Select *Disable* from the floating menu and press <OK> in the confirmation dialog box that appears to disable the Navigator pane.

The Navigator can be enabled again from the *InfoQuest FP main* window:

1.6.9.5 Select *File > Preferences > Windows behaviour* and check *Show navigator*.

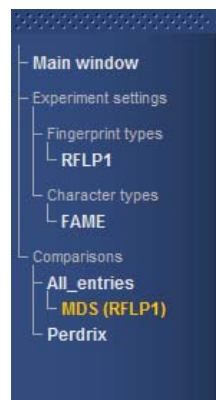


Figure 1-29. Navigator pane of a InfoQuest FP session in which two *Experiment type* windows and two *Comparison* windows are open. For one comparison, an MDS is also available. The MDS window is currently active.

2. DATABASE

2.1 Introduction

2.1.1 Local and connected databases

InfoQuest FP offers two possibilities to store its data: the program's own local database engine (the *local database*) or an external ODBC compatible database engine. The latter solution is called a *connected database*. Currently supported database engines are Microsoft Access, Microsoft SQL Server, Oracle, PostgreSQL, MySQL, and DB2. Others may work as well but are not guaranteed to be fully compatible in a standard setup.

The *local database* is a generic file-based databasing environment with limited possibilities but simple to handle. It was the first database solution available in InfoQuest FP, and is still maintained for compatibility reasons. However, we recommend to use the connected database option to create new databases. As connected databases rely on market standard database software, they offer a number of security and sharing/exchange options and are much more extensible than local databases. In addition, connected databases offer a richer database structure within InfoQuest FP, by providing character mapping, character and fingerprint lane fields, quality scores on sequences, and last but not least, the *Levels* and *Relations* (see Section 2.3). Connected databases are the default option in InfoQuest FP.

2.1.2 Elementary structure of a database

The core unit of a InfoQuest FP database is the *entry*. Entries represent the biological entities for which data is sampled, digitized and imported, to be further compared and analyzed. Each entry is identified by a unique *key*, through which various pieces of information are linked to the appropriate entries: information fields, attachments, fingerprints, contig projects and sequences, character data, etc. (see Figure 2-1).

2.1.3 Location of a database

A InfoQuest FP database can be located on the local computer or anywhere on the network, as long as InfoQuest FP has sufficient privileges to write to the database and its associated folders. InfoQuest FP recognizes and inventories the available databases by looking in the *home directory*, a folder that can be specified by the user, and which contains a database descriptor file for each database. The files have the extension ".dbs" and basically contain a tag [DIR] under which the full database path or network location is written. If the databases are subfolders of the home directory (the default setting), the full path is relativized as [HOMEDIR] *dbname* (*dbname* is the name of the database folder). This notation is only used from version 5.0 onwards, and is not compatible with earlier versions. The relative reference has the advantage that the home directory with all data-

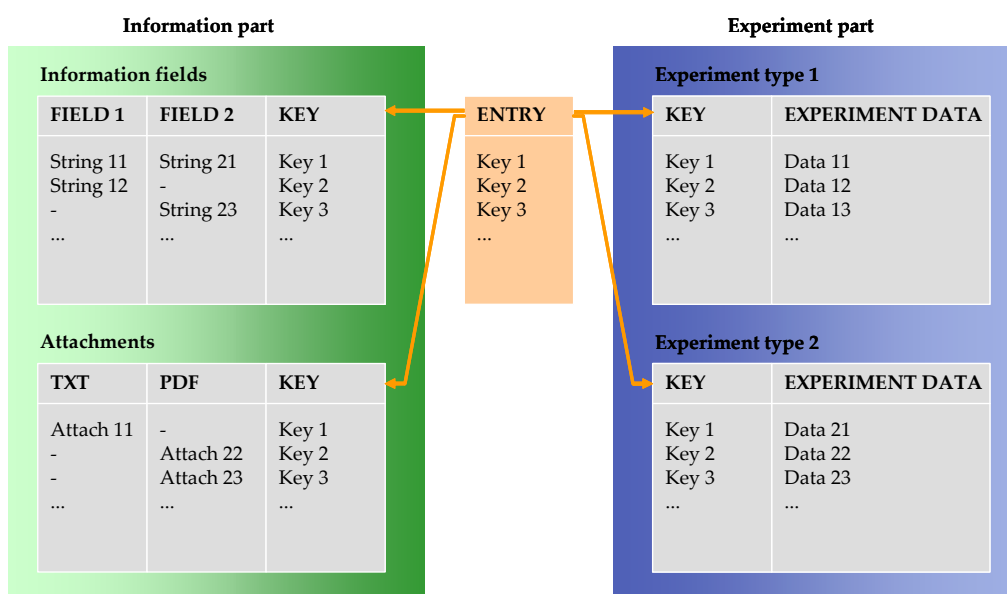


Figure 2-1. Linking various sorts of information to database entries through unique keys.

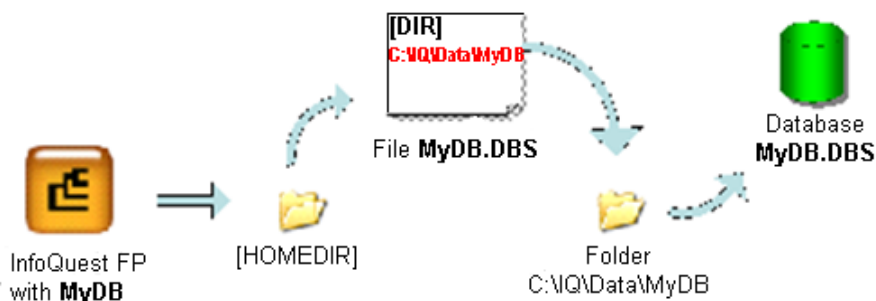


Figure 2-2. Steps in opening a database “MyDB”: (1) InfoQuest FP looks in the home directory for file MyDB.dbs (2) File MyDB.dbs is opened to obtain the database path; (3) the database is opened in the database path found.

bases as subfolders can physically be moved to another drive or computer without having to change the database path in the .dbs files. The different steps in opening a database are schematically represented in Figure 2-2.

A InfoQuest FP database requires a number of subfolders to be present in the database folder (see Figure 2-3). These folders are created automatically by

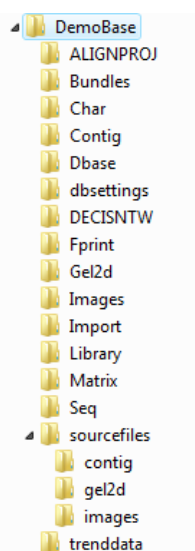


Figure 2-3. A InfoQuest FP database folder with its subfolders.

the software when the database is set up, or when the software is launched. In a local database setup, they may contain a number of files that store the experiment data, information fields, experiment and analysis settings, etc. For backup purposes, the entire database folder with all its subfolders should be backed up (see Section 2.4). In a connected database, most, but not all of these folders are empty. Window and viewing settings, for example, are stored locally. Optionally, imported files such as gel TIFF images or sequencer trace files, can also be stored locally in a connected database setup, but the default setting is to store all information, including import files, inside the connected database. Figure 2-4 illustrates a typical connected database setup.

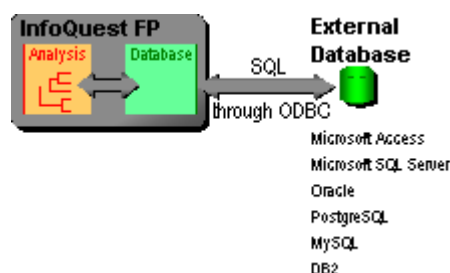



Figure 2-4. Connected database setup: all data is passed on to the SQL database through ODBC; InfoQuest FP' local database is empty.

2.1.4 Setting up a new database

The creation of a new database is described in . This paragraph also describes how to locate and change the home directory. A number of settings can be changed both for local and connected databases:

2.1.4.1 In the Startup program, select a database and press the settings button (). From the menu that is displayed, select *Database settings*. The *Database settings* dialog box appears (Figure 2-5).

Here you can change the database directory with *<Change directory>*. This option will overwrite the [DIR] tag in the *Database*.dbs file (Figure 2-2).

With *Enable log files*, it is possible to log all events that alter database information (see 2.1.6). In a connected database, logging events are written to a table, whereas in a local database, log files are created.

With *<ID code>*, you can install an ID code to protect all important settings in a local database (see 2.1.5). This function has no meaning in a connected database, where other, more advanced protection mechanisms are available (see 2.3.10).

2.1.4.2 Press *<OK>* or *<Cancel>* to exit the *Database settings* dialog box.

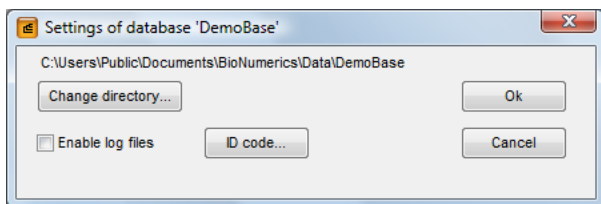



Figure 2-5. The *Database settings* dialog box in the Startup program.

2.1.4.3 To remove a database, select it from the list, press

the settings button () again and select *Delete database*. The program will ask for confirmation before deleting the database. When the database contains non-empty subfolders, which is usually the case, its folder structure will not physically be removed from the volume. However, it will be removed from the database list in the Startup program.

2.1.5 Protecting a database

The protection tools discussed in the present section are only meant to reduce the chances of incidental erroneous changes or damage to important database or experiment settings. **They are by no means secure enough to prevent others to making changes to the database!** Other, more advanced and secure protection mechanisms are available for connected databases (see 2.3.10).

In order to protect a database against incidental data loss, it is possible to lock the important settings and data files in the database. The following files can be locked:

- The settings files of the experiment types: as long as this file is locked, the settings for the experiment type cannot be changed.
- The data files for the experiment types: data of existing entries cannot be changed, however, new information can be added to the experiment data file. In a connected database, this option only applies to fingerprint data.
- Libraries for identification (locally stored libraries only): nothing can be changed in a locked library, but it still can be used for identification.

Each file can be locked and unlocked separately, so that it is possible to lock and protect "final" files and leave other files open for additional input.


2.1.5.1 The setting files, data files, and libraries in the database can be locked using the *File > Lock* command in the file's edit window. Once the settings are locked they cannot be changed anymore, until they are unlocked again by executing the command *File > Lock*.

A locked file is shown with a small key icon left from the filename in the *Files* panel of the *InfoQuest FP main*

window, and a key icon also appears in the *Fingerprint file information* panel of the file's edit window.

To protect a local database against modification by others or misuse, InfoQuest FP allows an *ID code* to be set for a database. Once an ID code is set, the database settings can only be changed after entering the ID code. In addition, locked files can only be unlocked or vice versa after entering the ID code. For connected databases, we recommend to use the protection mechanisms provided by the database management software (DBMS). For instructions on how to protect Access databases, see 2.3.10. For other databases, we refer to the specific DBMS documentation.

2.1.5.2 In order to set an ID code for the local database, run the Startup program and press the settings button

(). From the menu that is displayed, select *Database settings*.

2.1.5.3 In the *Database settings* dialog box (Figure 2-5), press *<ID code>* and enter the ID code. Any string of characters is allowed. The program will ask you to confirm this by entering the ID code a second time.

2.1.5.4 If you want to remove an ID code, press *<ID code>* in the *Database settings* dialog box and leave the input box empty.

NOTE: If you forget the ID code, you will have to contact Bio-Rad.

2.1.6 Log files

In certified environments and laboratories where conscientious recording of manipulations is important, the *log files* in InfoQuest FP are a useful tool. For every InfoQuest FP session, the log files show the Windows user who has last made the changes together with the kind of changes and the date and hour. There are a number of differences between the logging in a local database and a connected database. Logging is more complete and is centrally maintained in a connected database. In a local database, separate log files are maintained for different data types.

Log files are recorded for the following data types:

- The database: the log file lists any changes in names of database fields, any entries that are added or deleted, and keys of entries that are changed. It also reports if new experiment types are created, if experiment types have been renamed or removed.
- The settings of the experiment types: for every change made, the kind of change is indicated. All settings are recorded in the log file, so that the user may restore the previous settings based upon the log file, if enabled.
- The data for the experiment types: if data for entries are changed, the log file lists these entries. It also

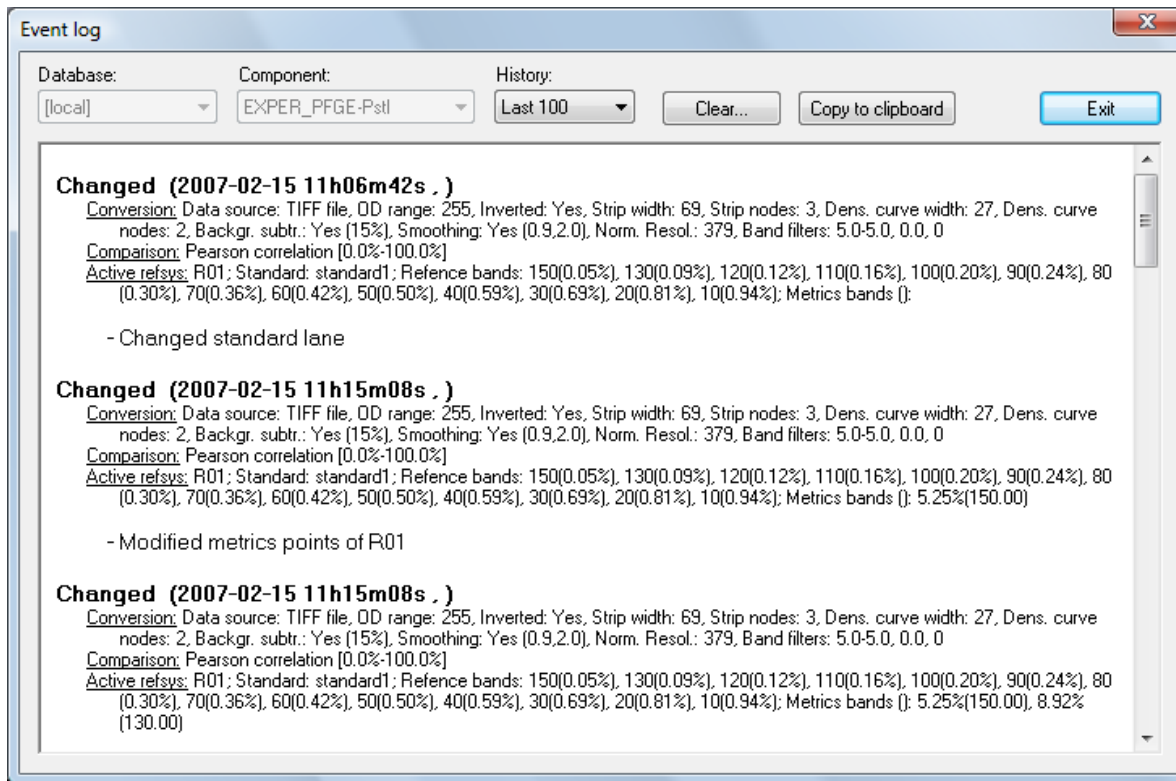




Figure 2-6. Event log viewer for a database component (local database).

mentions the creation of new experiments and the deletion of experiments.

- Libraries for identification: the log file keeps record of any changes in library units and records the addition and deletion of library units.

2.1.6.1 In order to enable the creation of log files for the database, press the settings button () in the Startup program and select *Database settings*.

2.1.6.2 In the *Database settings* dialog box (Figure 2-5), check the *Enable log files* checkbox.

2.1.6.3 In the data file windows, experiment file windows, the *InfoQuest FP main window*, or the *Library window*, select *File > View log file* or press  to display the log file.

In a local database (as opposed to a connected database; see Section 2.3), InfoQuest FP creates a temporary file, *DBASE.LOG* or *<EXPERIMENTNAME>.LOG*. For data files, it creates a log file *<DATAFILE>.LOG*.

2.1.6.4 The log files are loaded in InfoQuest FP' Event log viewer (Figure 2-6).

2.1.6.5 You can clear a log file with the command *<Clear>* or copy it to the clipboard with *<Copy to clipboard>*.

From the clipboard it can be pasted in other applications with a *Paste* command. The text is formatted as RTF (Rich Text Format) which enables the formatting to be retained in other software that supports RTF.

The items *Database* and *Component* are only applicable to connected databases (see Section 2.3).

2.2 Database functions

2.2.1 Adding entries to the database

In the database **DemoBase**, there are already entries defined. In most further exercises in this guide, we will work on our own database **Example**. Therefore we will start InfoQuest FP again with this new database:

2.2.1.1 Close the **DemoBase** database with *File > Exit*.


2.2.1.2 Back in the Startup program, select the database

Example and click on the  button. Simply double-clicking on the database name does the same.

Adding entries to the database can happen in two ways:

- You can add one or more entries directly to the database. Initially, these entries will be empty and no experiments will be linked to them. When you import experiment data later on, you can link the data to the entries.
- When you import a file of experiments, the program will ask you whether you want it to automatically create a corresponding database entry for each experiment.

We will now create a few database entries without importing experiments.

2.2.1.3 Select *Database > Add new entries* or press  in the toolbar.

A dialog box appears, asking for the number of new entries to create, and the database where they should be created. When there is a connected database associated with the database (see Section 2.3), there is a possibility to add the new entries either in the local database or in the connected database.

The input field in the bottom of the window allows a key to be entered by the user. This input field is only accessible when one single entry is added. As soon as the number of entries is specified to be more than one, the field is disabled.

2.2.1.4 Enter the number of entries you want to create, e.g. 3, and press **<OK>**.

The database now lists three entries with a unique key automatically assigned by the software. Usually, one will not want to change this entry key, but in special cases, it may be useful to change or correct the key manually. This can be done as follows:


2.2.1.5 Select the entry and *Database > Change entry key*.

2.2.1.6 Change the entry key in the input box, e.g. *Entry 1*, and press **<OK>**.

The key is a critical identifier of the database entries, and if you already have unique labels that identify your organisms under study, you can use these labels as keys in InfoQuest FP. In the latter case, they can be effectively used as a database field. As we will explain later, the key is also an important component in automatically linking experiments to existing database entries.

*NOTE: Remember the use of floating menus as described in : right-clicking in the database panel of the InfoQuest FP main window pops up the menu **Add new entries** and **Change entry key** (if you click on an entry).*

2.2.1.7 To remove an entry from the database, select one of the entries, e.g. the third one, and *Database > Remove*

entry or  in the toolbar. The program asks to confirm this action, and will warn you if there is any experiment information linked to the entry.

2.2.1.8 To remove all selected entries at once, choose *Database > Remove all selected entries*. See 2.2.7 to 2.2.9 for more information on the selection of entries.

WARNING: There is no undo function for this action and removed entries are irrevocably lost, together with any experiment information linked to them!

2.2.1.9 To remove all entries that have no experiment linked to them, you can select *Database > Remove unlinked entries*. In the case of our example database this would result in removal of all entries, since none has an experiment linked yet.

2.2.2 Creating information fields

A number of predefined information fields are automatically created when creating a new database. All predefined information fields are listed in the pull-down menu's in the information fields header of each panel and can be either displayed or hidden (see 1.6.6).

In addition to these default information fields, extra information fields can be added with *Add new information field* or removed with *Remove information field*. These menu commands can be accessed in each panel by right-clicking in their information toolbars. Information

fields can also be added to the *Database* panel with the corresponding menu commands.

2.2.2.1 Select *Database > Add new information field*.

2.2.2.2 Enter the name of the database information field, for example *Genus*, and press **<OK>**.

2.2.2.3 Select *Database > Add new information field* again to define the second field, *Species*.

2.2.2.4 Then, select *Database > Add new information field* again to define a third field, *Subspecies*.

2.2.2.5 Finally, select *Database > Add new information field* again to define a field *Strain no*.

The menu functions *Database > Rename information field* and *Database > Remove information field* can be used to rename and remove an information field, respectively.

NOTE: Renaming information fields in InfoQuest FP is not possible when using a connected Access database (.mdb and .accdb). In this case, you should open the database with Access (see 2.1.3 on how to locate the database) and rename the corresponding column in the ENTRYTABLE table. When the database is again loaded in InfoQuest FP, both the old and the renamed information field will appear. The old information field (now empty) can then be removed.

An information field in a local or connected database may contain up to 80 characters. In a local database, a maximum of 150 fields can be defined. In a connected database, many more fields can be defined, but only 150 can be displayed at the same time.

2.2.3 Entering information fields

2.2.3.1 By double-clicking, or pressing **Enter** on a database entry, the *Entry edit* window appears (Figure 2-7). Right-clicking on the entry, and selecting *Open entry* also works.

In default configuration, the left panel of the *Entry edit* window shows the information fields and the right panel shows the available experiments for the entry. The tab in the bottom right of the *Experiments* panel gives access to the *Attachments* panel. The latter allows attachments to be added and viewed for the entry (see 2.2.4). The *Entry edit* window can be rescaled to see more and/or longer information fields. The relative size of the panels can also be modified by dragging the separator line between the panels. All panels in the *Entry edit* window are dockable and their display can be customized as described in .

2.2.3.2 Enter some information in each of the fields (see Figure 2-7).

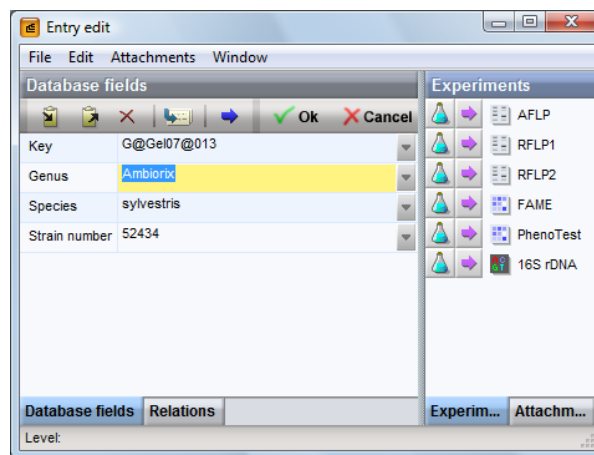






Figure 2-7. The *Entry edit* window.


2.2.3.3 If a number of entries have mostly the same fields, you can copy the complete entry information to the clipboard using the **F7** key or .



2.2.3.4 To clear the complete information of the entry, press .

2.2.3.5 To paste the information from the clipboard, press the **F8** key or .

If some of the information fields are the same as entered for previous entries (for example genus and species name), you can drop down a history list for each information field. The history lists can contain up to 10 previously entered strings for the information field. Using the history lists is recommended (i) to save time and work and (ii) to avoid typographical errors.

2.2.3.6 Drop down a history list by clicking the  button on the right hand from the information field. A floating menu appears from which you can select an information string.

The  button is related to ODBC communication with an external database (see Section 2.3).

2.2.3.7 Using , you can select or unselect the opened entry in the database (see Figure 2-7), for the construction of comparisons. When the entry is selected, this button shows as .

2.2.3.8 Press the **Enter** key or **<OK>** to close the *Entry edit* window and store the information, or press the **Escape** key or **<Cancel>** to close the window without changing any information.

In order to quickly enter the same information for many entries, the use of the keyboard is recommended: use the

Arrow Up and Down keys to move through the entries in the database, use the Enter key to edit an entry, use the F7 and F8 keys to copy and paste information, and use the Enter key again to close the *Entry edit* window.




Alternative to using the *Entry edit* window, information in non-default information fields in the database and in other grid panels can be edited directly by clicking twice on an information field in the database. The information will appear highlighted and can be edited. When field states are defined (see 2.2.5), they now become available as a drop-down list.


NOTE: Single click field editing can be enabled via File > Preferences in the InfoQuest FP main window (see 1.6.6.9).

2.2.4 Attaching files to database entries

Besides its information fields and the experiments linked to it, a database entry can also have files attached to it. Usually the attachment is a link to a file, except for text attachments, which are physically contained in the database. In addition to the attachment itself, InfoQuest FP also allows a description to be entered for the attachment. The following data types are supported as attachment:

- **Text:** Plain ASCII text attachments of unlimited length. InfoQuest FP contains its own editor (similar to Notepad) to paste or type text strings.
- **Bitmap image:** images of the following bitmap types are supported: TIFF, JPEG (JPG), GIF, BMP, PNG, and WMF. InfoQuest FP contains its own viewer for image attachments.
- **HTML documents:** HTML and XML documents can be attached as well as URLs. InfoQuest FP contains its own HTML viewer.
- **Word[®] document:** Documents in Microsoft[®] Word[®] format can be attached. The default editor or viewer registered by your Windows system will be opened if you want to edit or view the document.
- **Excel[®] document:** Documents in Microsoft[®] Excel[®] format can be attached. The default editor or viewer registered by your Windows system will be opened if you want to edit or view the document.
- **PDF[®] document:** Documents in Adobe[®] PDF[®] format can be attached. The default editor or viewer registered by your Windows system will be opened if you want to edit or view the document.

2.2.4.1 To create an attachment for an entry, open the *Entry edit* window as described in 2.2.3 and select the *Attachments* tab. The *Attachment* panel contains three buttons, respectively to create a new attachment , to open (view) an attachment  and to delete an attachment . The same commands are available from the menu as *Attachment > Add new*, *Attachment > Open*, and *Attachment > Delete*, respectively.

2.2.4.2 Press  to create a new attachment. The *Entry attachment* dialog box appears (Figure 2-8).

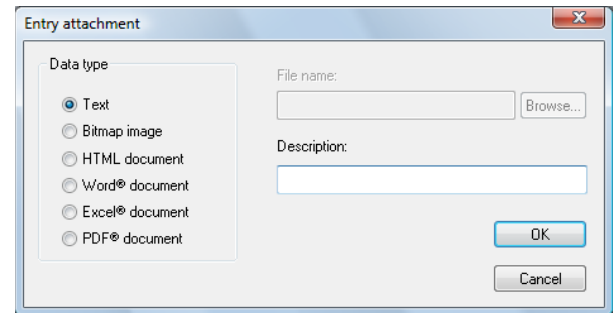


Figure 2-8. The *Entry attachment* dialog box.

2.2.4.3 Under *Data type*, specify one of the supported data types, as described earlier in this paragraph.




2.2.4.4 All data types except Text link to a file on the computer or the network. You can enter a path and a file name or use the **<Browse>** button to browse to a file of the specified type. Text attachments are stored inside the InfoQuest FP database.

2.2.4.5 A *Description* input field allows you to enter a description line for the attachment. The description will appear next to the attachment icon in the *Entry edit* window (Figure 2-7) and for text, bitmap, and HTML type attachments, it will also appear in the viewer or editor (text) window when the attachment is opened.

2.2.4.6 To open an attachment, double-click on the attachment icon in the *Entry edit* window.

2.2.4.7 In case of a text attachment, a *Text attachment* editor is opened (Figure 2-9) where one can type or paste a text document of unlimited length. The format should be pure text; any formatting will be lost while pasting texts from other editors. The editor contains a

Save button , an *Undo*  (shortcut: CTRL+Z) and *Redo*  (shortcut: CTRL+Y) button, as well as a

Cut  (shortcut: CTRL+X), **Copy**  (shortcut: CTRL+C) and **Paste**  (shortcut: CTRL+V) button.

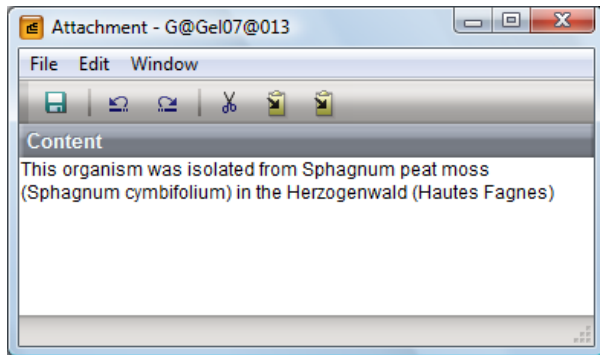





Figure 2-9. The *Text attachment* editor.

2.2.4.8 In case of a bitmap image (TIFF, JPEG [JPG], GIF, BMP, PNG, and WMF), InfoQuest FP' own viewer is opened with the bitmap displayed. The window contains a **Zoom in**  and a **Zoom out**  button.

2.2.4.9 In case of HTML and XML attachments, InfoQuest FP' own browser is opened with the HTML or XML file displayed. Note that an HTML document can be a link to a website, in which case the browser will display the website. The browser contains a **Back** button  to return to the previous page.

2.2.4.10 Word[®], Excel[®] and PDF[®] attachments are opened in the default programs registered by your Windows system for these file types.

2.2.4.11 To edit the link of an attachment or its description line, use **Attachments > Edit** in the menu of the

Entry editor. The *Entry attachment* dialog box appears as shown in Figure 2-8.

If the predefined information field **Attachments** is checked in the pull-down menu of the *Database* panel, the number of attachments is displayed for all entries (see Figure 2-10).

Database entries				
Key	Attachments	Genus	Species	Strain number
G@Gel07@002	1	Ambiorix	sylvestris	52441
G@Gel07@003	2	Ambiorix	aberrans	52449
G@Gel07@004		Vercingetorix	palustris	42815

Figure 2-10. Detail of database panel in *InfoQuest FP main* window, showing two entries with attachments.

2.2.5 Information field properties

In InfoQuest FP, properties can be assigned to a database information field. These properties include a list of *field states*, i.e. possible content that can be contained within the information field. Individual states can each be displayed against a differently colored background, for an improved display in grid panels. Field states also provide an additional display option in an unrooted tree or coordinate space window. Finally, when field states are defined, they become available as a drop-down list, facilitating and harmonizing the input of data via direct field editing.

2.2.5.1 In the *InfoQuest FP main* window with **DemoBase** loaded, click on the header of the information field for which you want to set the properties (e.g. the 'Genus' field).

2.2.5.2 Select **Database > Information field properties**. The *Information field properties* dialog box appears (see Figure 2-11). The *Field states* list is initially empty.

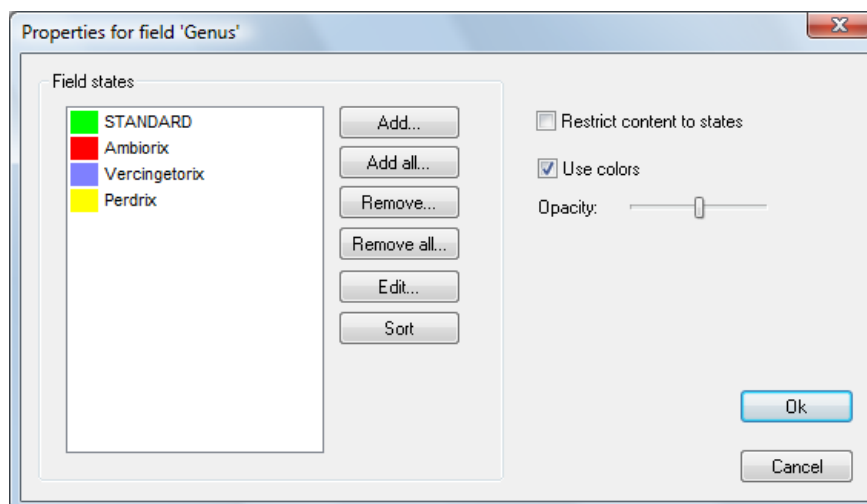


Figure 2-11. The *Field properties* dialog box for the 'Genus' field in DemoBase.

2.2.5.3 Press **<Add>** and enter a possible content (a *state*) for the 'Genus' information field, e.g. Ambiorix. If **Use colors** is checked (see 2.2.5.9), the same dialog box also allows you to set the **Background color**.

The state Ambiorix is now listed under **Field states**. If the information field already contains information (as it is the case with **DemoBase**), this information can be used to automatically create a list of **Field states**.

2.2.5.4 Press **<Add all>** to automatically create all existing states for the field 'Genus'.

2.2.5.5 To edit a field state, select it from the list and press **<Edit>**.

2.2.5.6 Likewise, if you want to remove a field state, select it from the list and press **<Remove>**.

2.2.5.7 All field states can be removed at once by pressing **<Remove all>**. The program will ask for confirmation.

2.2.5.8 Press **<Sort>** to have the field states sorted alphabetically.

2.2.5.9 Check **Use colors** to display a specific color code for each field state. The *Field properties* dialog box should now look the same as in Figure 2-11.


Color codes will be automatically generated for the first 30 field states, but can be modified if desired using the **<Edit>** button. The **Opacity** slider sets the applied color intensity.


2.2.5.10 Check **Restrict content to states** if the **Fields states** list contains all possible states for this information field.

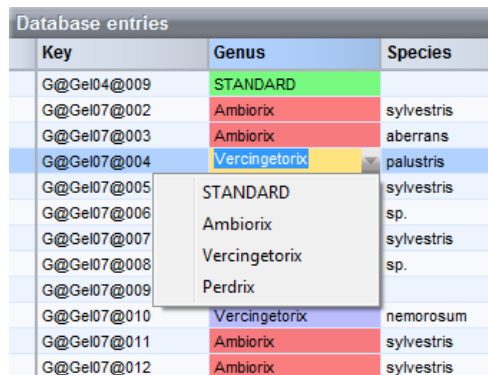
Turning on this feature forces database users to provide consistent information, as no other values than the ones specified in the **Fields states** list will be accepted for the information field.

2.2.5.11 Press **<OK>** in the *Field properties* dialog box.

In the *Database entries* panel of the *InfoQuest FP main* window (and also in the *Information fields* panel of the *Comparison* window), the different states of the 'Genus' field will each appear against their own background color. The extent of the background coloring can be modified using **File > Preferences**.

2.2.5.12 Click twice on an information field to enable direct editing and to display a  button on the right-hand side of the information field.

2.2.5.13 Press the  button (or press CTRL+down on the keyboard) to drop down a list from which you can select a field state (see Figure 2-12).




Key	Genus	Species
G@Gel04@009	STANDARD	
G@Gel07@002	Ambiorix	sylvestris
G@Gel07@003	Ambiorix	aberrans
G@Gel07@004	Vercingetorix	palustris
G@Gel07@005	STANDARD	sylvestris
G@Gel07@006	Ambiorix	sp.
G@Gel07@007	Vercingetorix	sylvestris
G@Gel07@008	Perdrix	sp.
G@Gel07@009	Perdrix	
G@Gel07@010	Vercingetorix	nemorosum
G@Gel07@011	Ambiorix	sylvestris
G@Gel07@012	Ambiorix	sylvestris

Figure 2-12. Detail of the *Database entries* panel, showing the drop-down list with field states for 'Genus'.

NOTES:

(1) If **Restrict content to states** (see 2.2.5.10) is checked in the *Field properties* dialog box for the *informations* field, its content cannot be edited by typing. Only the states available from the drop-down list can be selected as content.

(2) The field states drop-down list that appears after clicking the  button in the *Database entries* panel is different from the history drop-down list in the *Entry edit* dialog box (see 2.2.3.6). The latter automatically remembers the 10 last values entered for the field, while the former is not updated when a new state is entered by typing; the field state list needs to be updated via the *Field properties* dialog box.

2.2.6 Configuring the database layout

Since the *Database entries* panel is a grid panel, all display and customizing features discussed in are valid for this panel as well. Some features that are particularly useful in the context of database layout will be discussed here in detail.

Entries in the database can be ordered alphabetically by any of the information fields.



2.2.6.1 Click on one of the database field names in the information fields header of the *Database entries* panel.

2.2.6.2 Select **Edit > Arrange entries by field**.

When two or more entries have identical strings in a field used to rearrange the order, the existing order of the entries is preserved. As such it is possible to categorize entries according to fields that contain information of different hierarchical rank, for example *genus* and *species*. In this case, first arrange the entries based upon the field with the lowest hierarchical rank, i.e. *species*, and then upon the higher rank, i.e. *genus*.

When a field contains numerical values, which you want to sort according to increasing number, use *Edit > Arrange entries by field (numerical)*. In case numbers are combined numerical and alphabetical, for example entry numbers [213, 126c, 126a, 126c], you can first arrange the entries alphabetically (*Edit > Arrange entries by field*), and then numerically using *Edit > Arrange entries by field (numerical)*. The result will be [126a, 126b, 126c, 213].


The user can determine which information fields are displayed and the order in which they are shown.


2.2.6.3 Click on the column properties button  in the database information fields header. From the pull-down menu that appears, click on any field name to either display ( icon shown in the menu) or hide (no icon shown) the information field in the *Database entries* panel.

This feature can be used to hide fields that are non-informative for the user. For example, if keys are automatically generated, they might not contain useful information for you and can therefore be hidden.

2.2.6.4 The width of each database field can be adjusted by dragging the separator lines between the database field names to the left or to the right.

For example, if the genus name for the organisms is known or mostly the same, you can abbreviate it to one character and drag the separator between Genus and Species to the left to show just one character.

2.2.6.5 To change the position of an information field, click on the header of the field you want to move and then on the column properties button . Select *Move 'FieldName' to left* or *Move 'FieldName' to right*. Shortcut keys are CTRL+left arrow and CTRL+right arrow, respectively.

2.2.6.6 It is possible to freeze one or more information fields, so that they always remain visible left from the scrollable area. For example, if you want to freeze the Key field, select the field right from the Key in the field header, and select *Edit > Freeze left pane*. Alternatively, you can select *Freeze left pane* from the column properties button . This feature, combined with the possibility to change the order of information fields makes it possible to freeze any subset of fields.

2.2.6.7 The width of the *Database entries* panel as a whole can be changed by dragging the separator lines between the *Database entries* panel, the *Experiment presence* panel and the remaining panels to the left or to the right.

Settings 2.2.6.3 to 2.2.6.7 as well as all window sizes and positions are stored when you exit the software and are specific for each database.

When a new comparison is created or when an existing comparison is opened (see Chapter 4), the same layout as applied in the *Database entries* panel of the *InfoQuest FP main* window (which fields to display/hide, column width and order of information fields) is used for the *Information fields* panel of the *Comparison* window.

2.2.7 Selections of database entries

Selections in InfoQuest FP provide a tool to perform an action on a selected number of database entries, instead of on the database as a whole. As such, selections form the basis for the creation of comparisons (see Chapter 4), enabling a host of analysis tools to be applied on the selected entries. Selections can be cut, copied, pasted or deleted and furthermore allow the user to create subsets within a database (see 2.2.11), define library units for identification (see Section 5.2.1), select entries to be identified (see Section 5.2.2), run decision networks on specific entries, include entries in an alignment project and to share well-defined information with other users via the creation of bundles (see 2.6.2) or XML files (see 2.6.3).

Besides manual selection functions (see 2.2.8), automatic search and select functions are available in InfoQuest FP, with simple (see 2.2.9) and more advanced query functions (see 2.2.10).

2.2.8 Manual selection functions

The manual selection functions will be illustrated using the **DemoBase** database.

2.2.8.1 In the Startup screen, double-click on **DemoBase** to open the database for analysis. Alternatively, you can select **DemoBase** from the list and then press the




button.

A single entry can be selected by holding the CTRL key and left-clicking. Selected entries are marked by a colored arrow (Figure 2-13). Selected entries are unselected in the same way.

2.2.8.2 Select the first non-standard lane (CTRL + mouse click). The entry is now marked by a colored arrow.

2.2.8.3 In order to select a group of entries, hold the SHIFT key and click on another entry.

2.2.8.4 If you wish to select entries using the keyboard, you can scroll through the database using the Up/Down arrow keys, and select or unselect entries using the space bar.

2.2.8.5 A single entry can be selected or unselected from its *Entry edit* window (Figure 2-7) using the 

Key	Genus	Species	Strain number	1	2
G@Gel04@009	STANDARD				
G@Gel07@002	Ambiorix	sylvestris	52441		
G@Gel07@003	Ambiorix	aberrans	52449		
G@Gel07@004	Vercingetorix	palustris	42815		
G@Gel07@005	Ambiorix	sylvestris	52416		
G@Gel07@006	Ambiorix	sp.	52415		
G@Gel07@007	Ambiorix	sylvestris	52414		
G@Gel07@008	Ambiorix	sp.	52424		
G@Gel07@009	STANDARD				
G@Gel07@010	Vercingetorix	nemorosum	42816		
G@Gel07@011	Ambiorix	sylvestris	52418		
G@Gel07@012	Ambiorix	sylvestris	52417		


Figure 2-13. Database entries panel in the InfoQuest FP main window, showing selected entries (orange arrows).

button. When the entry is selected, this button shows as



All the entries from the database or from the current subset (for more information on subsets, see 2.2.11) can be selected using the keyboard shortcut CTRL+A or with *Edit > Select all* in the InfoQuest FP main window.


2.2.8.6 To make viewing of selected entries easier in a large database, you can bring all selected entries to the top of the list with *Edit > Bring selected entries to top* or use the keyboard shortcut CTRL+T for this utility.

2.2.8.7 Clear all selected entries with *Edit > Unselect all entries* (F4 key) or .

*NOTE: A very convenient command in combination with the manual selection functions is **Arrange entries by field**, which allows the database to be sorted according to the selected information field (see 2.2.6 for a detailed description).*

2.2.9 Automatic search and select functions

In addition to manually selecting entries from the database, entries can be searched and selected automatically using a simple and intuitive search function.

2.2.9.1 Select *Edit > Search entries* (F3) or . This pops up the *Entry search* dialog box (Figure 2-14).

You can enter a specific search string for each of the database fields defined in the database (left panel). Wildcards can be used to search for substrings: an **asterisk** * replaces any range of characters in the beginning or the end of a string, whereas a **question mark** ? replaces one single character.

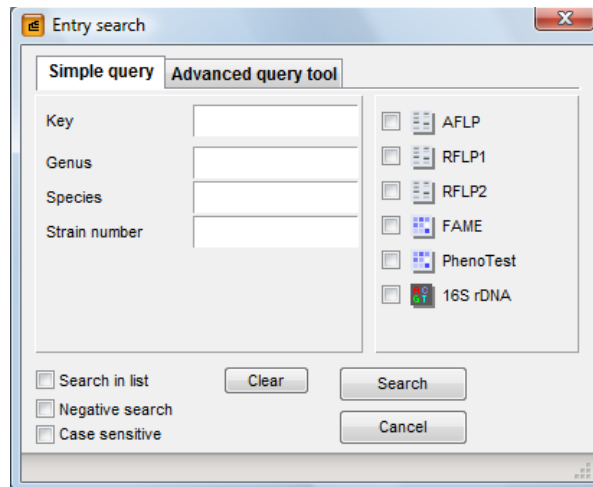


Figure 2-14. Entry search dialog box, Simple query tab.

It is also possible to search for all entries that contain a certain experiment (right panel). Both the string search and the experiment search can be combined.

Normally, successive searches are additive: new searches are added to the selection list. The *Search in list* checkbox allows you to refine the search within a list of selected entries.

With *Negative search*, all entries that do not match the specified criteria will be selected.

Case sensitive lets the program make a distinction between uppercase and lowercase.

The *<Clear>* button clears all entered search criteria.


2.2.9.2 As an example, enter *L* in the *Species* field.

2.2.9.3 Press *<Search>*. All entries having a L in their species name are selected: *Ambiorix sylvestris* and *Vercingetorix palustris*.

2.2.9.4 Call the *Entry search* dialog box again, and press the *<Clear>* button.

2.2.9.5 Enter **STANDARD** in the 'Genus' field, and check the *Negative search* checkbox.


2.2.9.6 Press *<Search>* to select all database entries, except the entries used as standard lanes in the RFLP and AFLP techniques.

2.2.9.7 Clear the selection with the F4 key or click the  button (*Edit > Unselect all entries*).

2.2.10 The advanced query tool

InfoQuest FP contains an advanced query tool that allows searches of any complexity to be made within the

database, based on information fields and experiment data.

2.2.10.1 Call the query tool again, by selecting *Edit* > *Search entries* or pressing .

The *Entry search* dialog box (Figure 2-14) contains an *Advanced query tool* tab.

2.2.10.2 Press <*Advanced query tool*>. The normal *Entry search* dialog box changes into the *Advanced query tool* (Figure 2-15).

The advanced query tool allows you to create individual *query components*, which can be combined with *logical operators*. The available targets for query components are *Database field*, *Database field range*, *Experiment presence*, *Fingerprint bands*, *Character value*, *Subsequence*, *Trend data parameter* and *Attachment*.

• Database field

Using this component button, you can enter a (sub)string to find in any database field (<*Any field*>) or in any specific field that exists in the database (Figure 2-16). Note that the wildcard characters * and ? are not used in the advanced query tool.

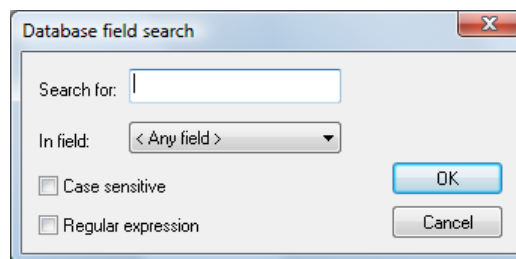


Figure 2-16. *Database field search component dialog box*.

The search component can be specified to be *Case sensitive* or not. In addition, a search string can be entered as a regular expression (see Section 7.2).

• Database field range

Using this component button, you can search for database field data within a specific range, which can be alphabetical or numerical. Specify a database field and enter the start and the end of the range in the respective input boxes (Figure 2-17). A range should be specified with the lower string or value first. Note that, when only one of both limits is entered, the program will accept all strings above or below that limit, depending on which limit was entered. For example, when only the first (lower) limit of the range is entered and the upper limit is left blank, all strings (values) *above* the specified string (value) will be accepted.

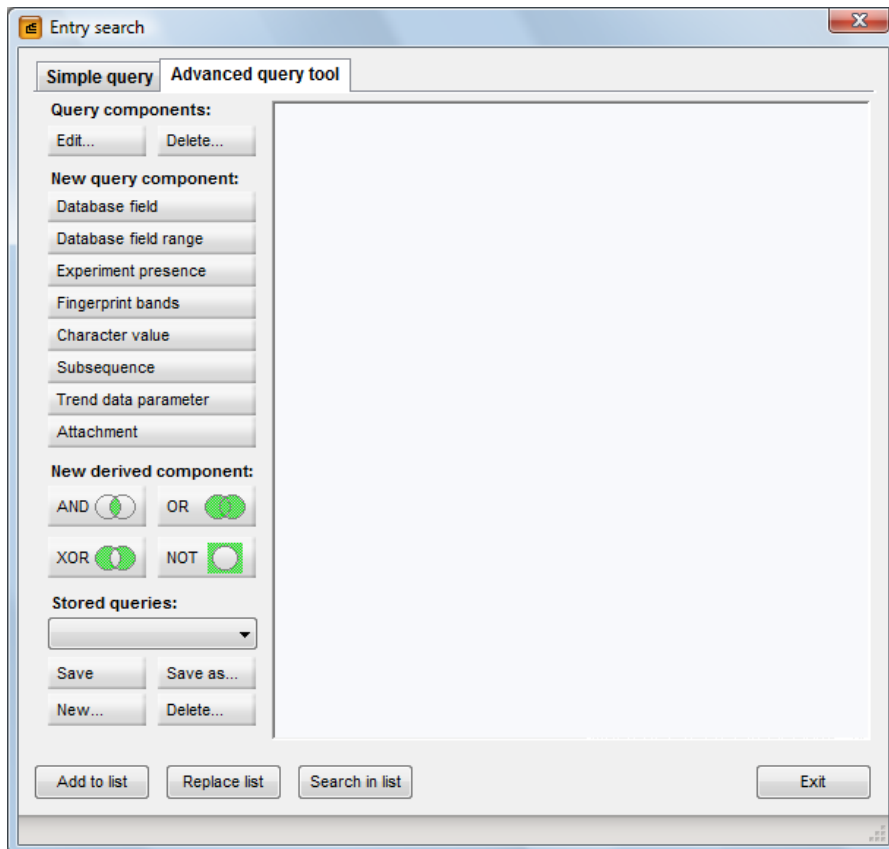


Figure 2-15. The advanced query tool.

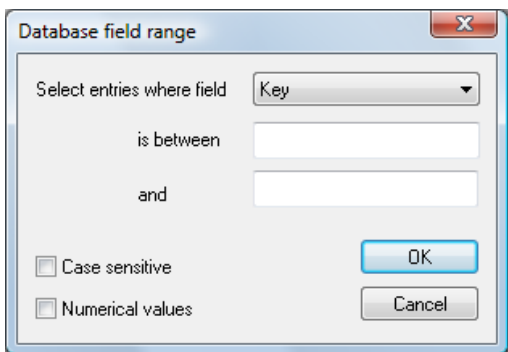


Figure 2-17. Database field range dialog box.

The search component can be specified to be *Case sensitive* or not. When *Numerical values* is checked, the search component will look only for numerical values and ignore any other characters.

• **Experiment presence**

With this search component, you can specify an experiment to be present in order for entries to be selected.

• **Fingerprint bands**

The *Fingerprint bands* search component allows specific combinations of bands to be found in the database entries. The dialog box that pops up (Figure 2-18) allows you to enter a *Fingerprint experiment*, and specify an *Intensity filter*, a *Target range*, and a *Number of bands present*.

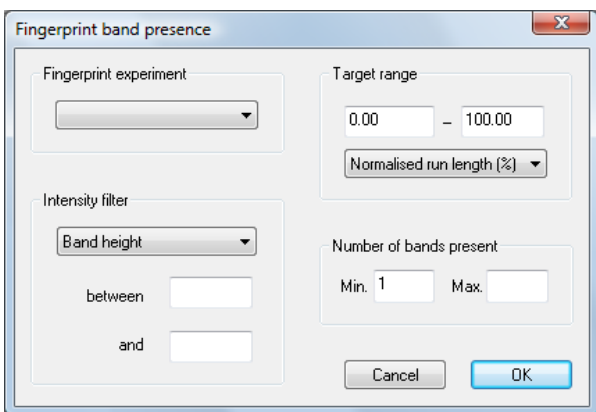


Figure 2-18. Fingerprint bands presence dialog box.

Under *Intensity filter*, you can choose which intensity parameter to be used: *Band height*, *Band surface* or *Relative band surface*. When a 2D quantification analysis is done, you can also choose *Volume*, *Relative volume* or *Concentration*. A range should always be specified with the lower value first. Note that, when only one of both limits is entered, the program will consider all bands above or below that limit, depending

on which limit was entered. For example, when only the first (lower) limit is entered and the upper limit is left blank, all bands *above* the specified intensity will be accepted. When both fields are left blank, no intensity range will be looked for, i.e. all bands will be considered.

Under **Target range**, you can search for bands with specific sizes, either entered as *Normalized run length (%)* or as *Metric values*. A target range should always be entered with the lower value first. Note that, when only one of both limits is entered, the program will consider all bands above or below that limit, depending on which limit was entered. For example, when only the first (lower) limit is entered and the upper limit is left blank, all bands *above* the specified size will be accepted. When both fields are left blank, no size range will be looked for, i.e. all bands will be considered.

Under *Number of bands present*, you can enter a minimum and a maximum number of bands the patterns should contain. Note that, when only one of both limits is entered, the program will consider all patterns with band numbers above or below that limit, depending on which limit was entered. For example, when only the first (lower) limit is entered and the upper limit is left blank, all patterns having *at least* the specified number of bands will be accepted. At least one of both limits must be entered.

• **Character value**

With the *Character value* component, you can search for characters within certain ranges. You should select a character type, specify a character or *<All>* characters, and enter a maximum and minimum value. A range should always be specified with the lower value first. Note that, when only one of both limits is entered, the program will consider all characters above or below that limit, depending on which limit was entered. For example, when only the first (lower) limit is entered and the upper limit is left blank, all characters with values *above* the specified value will be accepted.

• **Subsequence**

With the *Subsequence* component you can perform a search for a specific subsequence in a sequence type experiment (Figure 2-19). The sequence type experiment should be chosen, and a subsequence entered. A mismatch tolerance can be specified with *Maximum number of mismatches allowed*. The program can also search for sequences that have one or more gaps as compared to the search sequence, with the option *Allow gaps in sequence*. Similarly, the program can also find subsequences that match the search string with one or more gaps introduced with *Allow gaps in search string*. The gaps are counted with the mismatches, and the total number of mismatches and gaps together is defined by the parameter *Maximum number of mismatches allowed*. Unknown or partially unknown positions can

also be entered according to the IUPAC code, when *Accept IUPAC codes* is enabled.

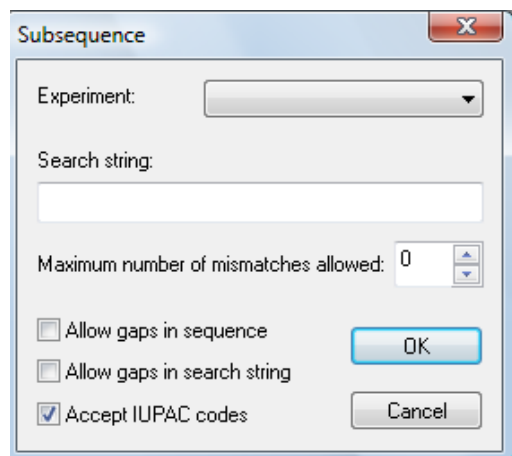


Figure 2-19. The *Subsequence search* dialog box.

• Trend data parameter

With the *Trend data parameter* component, you can search for trend data parameters within a specific range. You need to specify a trend data type experiment, a curve that belongs to it and a parameter defined for this experiment. Enter the start and the end of the range in the respective input boxes (Figure 2-20). A range should be specified with the lower value first. Note that, when only one of both limits is entered, the program will accept all values above or below that limit, depending on which limit was entered. For example, when only the first (lower) limit of the range is entered and the upper limit is left blank, all values *above* the specified value will be accepted.

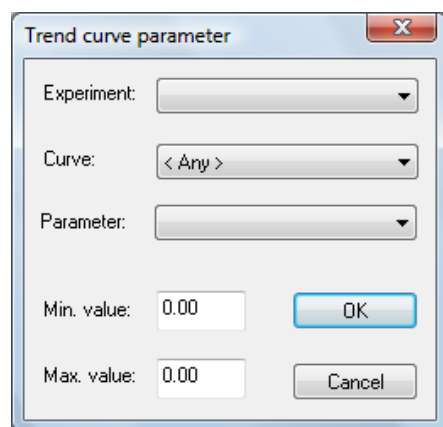


Figure 2-20. The *Trend curve parameter search* dialog box.

• Attachment

With the *Attachment* component, one can perform a search in attachments that are linked to database entries (Figure 2-21). With the pick list you can choose the type

of attachments to search in. One of the possibilities is **All**, i.e. to search within all attachment types. For all types of attachments it is possible to search in the *Description* field, and for text type attachments, it is also possible to search within the *Text*. The *Text* option does not apply to the other attachment types.

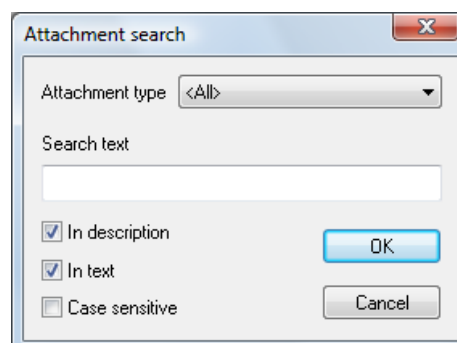






Figure 2-21. The *Attachment search* dialog box.

• Logical operators

NOT  **NOT**, operates on one component. When a component is combined with NOT, the condition of the component will be inverted.

AND  **AND**, combines two or more components. All conditions of the combined components should be fulfilled at the same time for an entry to be selected.

OR  **OR**, combines two or more components. The condition implied by at least one of the combined components should be fulfilled for an entry to be selected.

XOR  **XOR**, combines two or more components. Exactly one condition from the combined components should be fulfilled for an entry to be selected.

NOTE: The buttons for the logical operators contain a helpful Venn diagram icon that clearly explains the function of the operator.

To create a search component, you can select to search in the database fields, fingerprint bands, characters, sequences, and trend data parameters. As an example, we will select all entries from the genus *Ambiorix* that have no RFLP1 bands in the range 71-72 base pairs, and of which the 16S ribosomal RNA gene sequences contains a subsequence "TGGTGCATTG".

2.2.10.3 Press **<Database field>**. In the box that appears you can choose the genus field or leave **<Any field>** selected.

2.2.10.4 Enter “ambiorix” and press <OK>.

A query component now appears in the right panel, stating “Database field: Search ‘ambiorix’ in field ‘Genus’”.

2.2.10.5 Press <Fingerprint bands>. The *Fingerprint bands presence* dialog box appears (Figure 2-18).

2.2.10.6 Select **RFLP1** from the *Fingerprint experiment* pull-down list.

2.2.10.7 Under *Target range*, enter 71 - 72, and specify **Metric values**.


2.2.10.8 Press <OK>. A second component appears in the query window, saying “Fingerprint bands: ‘RFLP1’ has at least 1 bands in the range 71.00 - 72.00”.

2.2.10.9 Select this **Fingerprint bands** component by clicking on it (highlighted when selected), and press the

<NOT> button .

2.2.10.10 Select the first component by clicking on it.

2.2.10.11 Hold down the CTRL key and click on the **NOT** box resulting from the second component to select it together with the first one.

2.2.10.12 Press the <AND> button  to combine the created components with AND.

2.2.10.13 Press <Subsequence>. This box allows you to type or paste a sequence that will be searched for.

2.2.10.14 Type **TGGTGCATTG** in the input field, and select **16S rDNA**. Press <OK>.


A third query component appears in the right panel, stating “Subsequence: Search ‘TGGTGCATTG’ in the sequence ‘16S rDNA’”. We will now combine the resulting AND box from the first two components with this last component, using an AND operator, to restrict the selection to those sequences that fulfill the ‘Ambiorix’ and **RFLP1** conditions AND contain the specified subsequence.

2.2.10.15 Select the **AND** box by clicking on it.

2.2.10.16 Hold down the CTRL key and click on the **Subsequence** component to select it together with the **AND** box.

NOTE: Multiple components/operators can also be selected together by dragging the mouse over the boxes in the right panel.

As both components are now selected, we can combine them with a *logical operator*.

2.2.10.17 Press the <AND> button  to combine the created components with AND.

This is now shown graphically in the right panel (Figure 2-22).

2.2.10.18 To view the selected entries, press <Add to list>.

The entries that were found are highlighted with a colored arrow left from them.

The result of a logical operator as obtained in this example can be combined again with other components (or logical operators) to construct more complex queries.

Individual components can be re-edited at any time by double-clicking on the component or by selecting them and pressing <Edit>. Selected components can be deleted with <Delete>.

Queries can be saved with <Save> or <Save as>. Saved queries can be loaded using the pull-down listbox under **Stored queries**. Existing queries can be removed by loading them first and pressing <Delete>.

NOTES:

(1) In order to speed up the search function in case of large databases, it is important to know that searching through the database fields is extremely quick, while searching through sequences or large character sets can be much slower. Using the AND operator, it is always recommended to define the quickest search component as the first, since the searching algorithm will first screen this first component and subsequently screen for the second component on the subset that match the first component.

(2) When combined with a logical operator, query components contain a small node at the place where they are connected to the logical operator box (AND, OR, XOR). By dragging this node up or down, you can switch the order of the query components, thus making it possible to move the most efficient component to the top in AND combinations, as explained above.

(3) The Decision Networks in InfoQuest FP (see Section 5.3) provide an alternative to the advanced query tool: any selection that can be made in the advanced query tool can also be created using Decision Networks.

2.2.11 Subsets

A selection of entries from the database can be saved as a *subset*. Subsets can include a certain target group in a database, for example, a single genus in a database containing many species, or any selection of relevant strains for a certain purpose. Selecting the defined subset displays a view of the database containing only the entries of the subset. Search functions, copy and

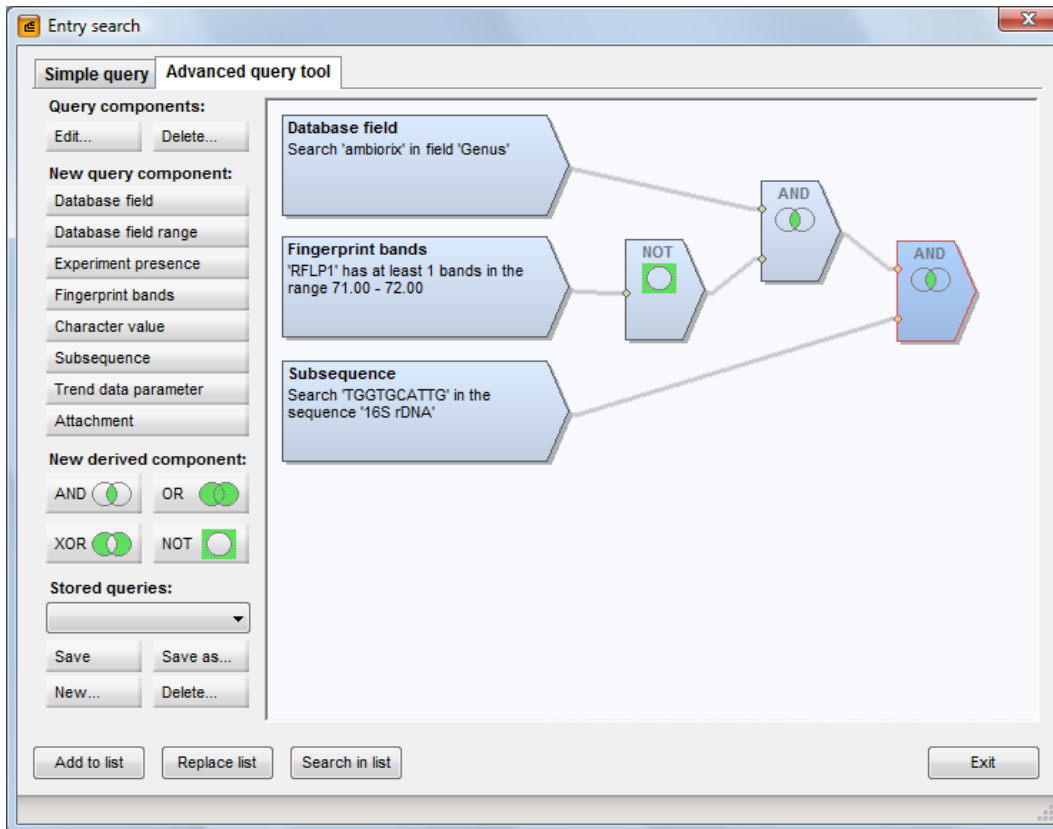


Figure 2-22. Combined query constructed in the *Advanced query tool* (see text for explanation).

select functions will be restricted only to the displayed subset, and new comparisons, when created, will only contain the selected entries from the subset.


2.2.11.1 In database **DemoBase**, make sure no entries are selected using *Edit > Unselect all entries* (F4 key) or

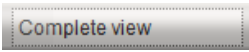


2.2.11.2 Selecting *Edit > Search entries* or press .

2.2.11.3 In the *Entry search* dialog box, enter "**Ambiorix**" under **Genus**, and press <OK>.

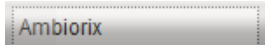
All *Ambiorix* entries are now selected. When we create a new subset, the selected entries will automatically be placed in the subset.


2.2.11.4 Select *Subsets > Create new* or press . Alternatively, you can click on the subset selector button

 which will drop down a list of currently defined subsets (initially empty), and an option <Create new subset>. Selecting this option has the same effect.

2.2.11.5 Enter a name for the subset, e.g. the name of the selected genus "**Ambiorix**".

The created subset is now displayed, and the name of the current subset is displayed in the subset selector


button .


2.2.11.6 Selecting the complete database or another subset, when available, can be done by pressing  and selecting *Complete view* or the other subset in the list.

Once a subset exists, it remains possible to add or remove entries, using the copy and paste functions. The following example will illustrate this.

2.2.11.7 Select subset *Ambiorix* from the subset selector button .


2.2.11.8 We want to remove all the "sp." entries from this subset. Clear any selected entries by pressing F4 on the keyboard and select the 3 "sp." entries by manual selection or using the search function.

2.2.11.9 Press  or select *Edit > Cut selection* to cut the selected entries from the current subset (keyboard shortcut CTRL+X).

2.2.11.10 We can place them in a new subset by pressing , entering a name, e.g. "**Unknowns**" and in this

new subset, pressing  or selecting *Edit > Paste selection* (keyboard shortcut CTRL+V).

2.2.11.11 If you want to copy entries from one subset to another, without removing them from the first subset,

there is also a command *Edit > Copy selection* or  (keyboard shortcut CTRL+C).


NOTES:

(1) A selection that is copied or cut from a subset or copied from the database is placed on the Windows clipboard as the keys of the selected entries, separated by line breaks. You can paste them in other software when desired.

(2) The commands *Cut selection*, *Paste selection* and *Delete selection* are not available in the *Complete view*.

2.2.11.12 When you want to remove entries from a subset without overwriting the contents of the clipboard, you can use the command *Edit > Delete selection* (keyboard shortcut DEL).

2.2.11.13 The current subset can be renamed using *Subsets > Rename current*.

2.2.11.14 The current subset can be deleted using *Subsets > Delete current* or .

2.2.12 Opening an additional database

Within InfoQuest FP, an additional InfoQuest FP database can be opened easily using the menu command *File > Open additional database* in the *InfoQuest FP main window*. A dialog box appears with the question "Do you want to open this database in a new instance of the software?". If you answer *<Yes>*, the additional database is opened in its own *InfoQuest FP main window*. If you select *<No>*, the additional database is opened in the same *InfoQuest FP main window*. When the additional database contains experiments and/or database fields that are not available in the already open database, InfoQuest FP will automatically create these components in order to be able to display them.

NOTE: Two connected databases cannot be opened simultaneously in the same instance of the software.

2.3 Connected databases

2.3.1 Advantages of a connected database

InfoQuest FP offers two possibilities to store its databases: the program's own local database engine (the *local database*) or an external ODBC compatible database engine. The latter solution is called a *connected database*. Currently supported database engines are Microsoft Access, Microsoft SQL Server, Oracle, PostgreSQL, MySQL, and DB2. Others may work as well but are not guaranteed to be fully compatible in a standard setup.

*NOTE: InfoQuest FP uses **Quoted Identifiers** to pass information to the connected database. Some database systems, for example MySQL, do not use this ANSI standard by default, but optionally. To use the database as a InfoQuest FP connected database, make sure that the use of Quoted Identifiers is enabled in the database setup.*

Connected databases are particularly useful in the following cases:

1. Environments where several users need to access the same database simultaneously. When the connected database engine is set up to support multi-user access, InfoQuest FP will allow multiple users to access and modify the database simultaneously. Note, in this respect, that InfoQuest FP takes a "snapshot" of the database when the program is launched. As such, changes to the database made by others while you have a InfoQuest FP session open will not be seen in your current session, until you reload the database during your session (see 2.3.8).
2. When sample information and/or experiment data is already stored in a relational database.
3. Laboratories where vast amounts of data are generated. In cases where many thousands of experiment files are accumulated, a powerful database structure such as PostgreSQL, Oracle or SQL Server will be faster and more efficient in use than InfoQuest FP's own local file-based database system.
4. When a more flexible database setup is to be achieved, for example with different access/permission settings for different users, and with built-in backup and restore tools.

In a connected database, InfoQuest FP will require a number of tables with specific columns to be available (see 7.1). InfoQuest FP can either construct its own tables and appropriate fields or link to existing tables and fields in the connected database. The latter option is particularly interesting to create a setup where InfoQuest FP hooks on to an existing database.

As soon as a valid connected database is defined, the user can start entering information in the connected database. InfoQuest FP writes and reads the information directly into and from the external database, without storing anything locally. Since every connected database has a local InfoQuest FP database associated with it, the user has the option to store and analyze local entries together with entries in the connected database. The information field *Location* displays either *Local* or *Shared* for locally or externally stored data, respectively. Although the use of connected databases and associated local databases is transparent, it is not recommended to store entries and experiments in a mixed way.

NOTE: A number of tables in a GelCompar II connected database deal with character types, sequence types, 2D gel types, and matrix types (InfoQuest FP). These tables are also required by GelCompar II, in order to assure compatibility with InfoQuest FP databases and to allow upgrading from GelCompar II to InfoQuest FP.

There is a function in the InfoQuest FP software that allows a local database to be converted into a connected database at any time (see 2.3.7). This process is irreversible: once a local database has been converted into a connected database, the local database is removed, and connected databases cannot be back-converted into local databases. However, using the available XML export and import scripts, it is also possible to export the contents of a local database as XML files (see 2.6.3 for more information about the XML Tools plugin) and import them in another connected database. In this way, the local database does not disappear. The same scripts also provide a means to convert connected database entries back into local database entries.

The combined use of local and connected databases is limited to avoid possible conflicts between the two database systems. In particular, the possibility that local and connected experiment types have the same name but different settings, should be avoided. Therefore, a few approved possibilities for working with connected databases are supported:

1. Creating a new database in InfoQuest FP, which is linked to a new connected database. InfoQuest FP is allowed to construct the database layout.
2. Creating a new connected database in InfoQuest FP, by linking to an existing database that has a table structure already in a InfoQuest FP compatible format (e.g. linking to an existing InfoQuest FP connected database).

3. Creating a new connected database, linking to an existing database which is not created using InfoQuest FP.
4. Converting a local database to a new connected database.

These possibilities are described in subsequent paragraphs.

2.3.2 Setting up a new connected database

InfoQuest FP can automatically create a new database in Microsoft Access. When you are using SQL Server, Oracle, or PostgreSQL, however, you will have to create a new blank database before proceeding with the following steps.

2.3.2.1 In the InfoQuest FP Startup program, click the



button to create a new database.

2.3.2.2 Enter **ConnectedBase** as database name.

2.3.2.3 In the next step, choose **<Yes>** to automatically create the required directories, since a local database associated with the connected database is required.

2.3.2.4 In the next step, click **<Yes>** to enable the creation of log files, and press **<Finish>**.

A new dialog box pops up, prompting for the type of database (see Figure 2-23):

- ***New connected database (automatically created)*** is the default setting and creates a new, empty Access database. This is recommended in most cases.
- ***New connected database (custom created)*** should be checked if a DBMS different from Microsoft Access is employed (e.g. SQL Server, PostgreSQL, or Oracle).

This option is especially useful when one expects to generate a very large database (>4 gigabyte) or when multi-level database protection tools are required.

- ***Existing connected database*** is to be selected when InfoQuest FP should be linked to an already existing database (see 2.3.5 and 2.3.6 for instructions).
- ***Local database (single user only)*** creates a local file-based InfoQuest FP database. This option is not recommended, since it has limited functionality in comparison with a connected database (see 2.3.1).

2.3.2.5 In most cases, an Access database will be sufficient, so you can leave the default setting ***New connected database (automatically created)*** and press **<Proceed>**. Next, continue with step 2.3.2.11.

NOTE: It is not required to have Microsoft Access installed on your computer to create an Access (.mdb) database in InfoQuest FP. InfoQuest FP uses instead the Microsoft Jet Engine, which comes with the Windows operating system.

2.3.2.6 If the database engine is SQL Server, PostgreSQL, or Oracle, select ***New connected database (custom created)***.

2.3.2.7 The checkbox ***Store fingerprints in database*** (enabled by default) offers the choice to store fingerprint files (TIFF images and .CRV curve files) in the connected database or in the sourcefiles directory (see 2.3.3). The checkbox can be left checked.

2.3.2.8 To connect InfoQuest FP to the database that you have created in the DBMS, you will need to build a ***Connection String*** using the **<Build>** button.

2.3.2.9 The dialog box that pops up now is generated by your Windows operating system and may differ depending on the Windows version installed. Therefore we refer to the Windows manual or help function to select or create a DSN file (ODBC Data Source) that specifies the ODBC driver, and to set up a connection to the database.

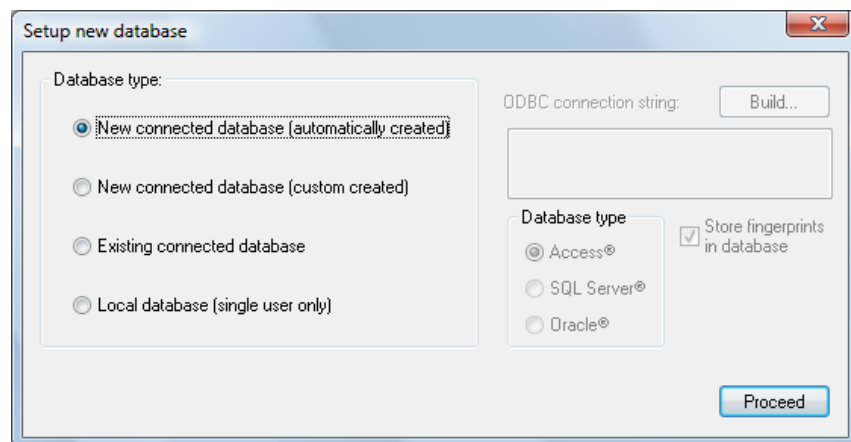


Figure 2-23. Database selection dialog box.

2.3.2.10 Once the database connection is properly configured, you can press <OK> to quit the database setup.

The *Plugin installation* window pops up, from which you can install the available plugins. For more information on the use of plugins, see Section 1.5.3.

2.3.2.11 Press <Proceed> in the *Plugin installation* window to open the *InfoQuest FP main* window with the newly created, blank database.

2.3.3 Configuring the connected database link in InfoQuest FP

In the *InfoQuest FP main* window, you can set up a connection to a connected database, or configure an existing connection. In case the program reports database linkage problems when opening the database, you will need to use this configuration to create the required tables in the database.

2.3.3.1 Select *Database > Connected databases*.

This opens a list of all currently defined connected databases for this InfoQuest FP database (normally just one; see Figure 2-24).

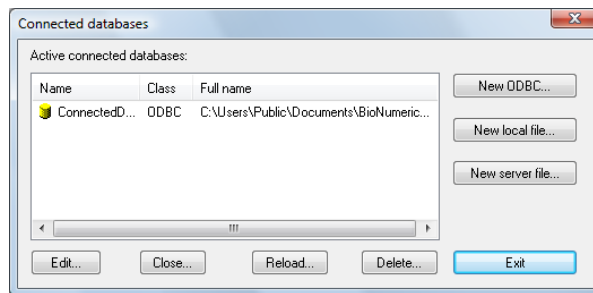


Figure 2-24. *Connected databases list window*.

2.3.3.2 Select the connected database of choice and click <Edit>, or double-click on the name.

This results in the *Connected database configuration* dialog box (Figure 2-25).

The upper left input field (*Connected database*) shows the name of the *connection description file*, which can be found in the local database directory. When InfoQuest FP has created a new connected database in the Startup program, the file is named **dbname*.xdb* by default. The default directory for the .xdb file is [HOMEDIR]*dbname*. The [HOMEDIR] tag thereby points to the home directory as defined in the Startup program (see 1.5.1). The .xdb file is a text file and can be edited in Notepad or any other text editor.

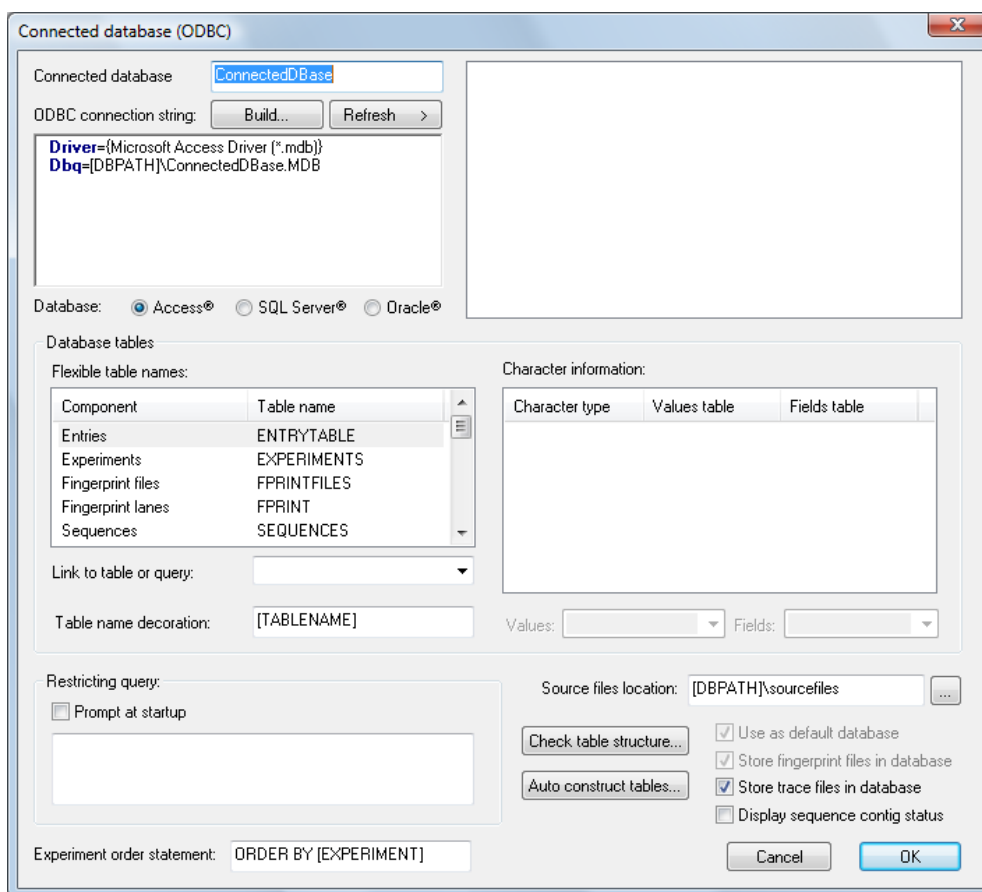


Figure 2-25. The *Connected database configuration* dialog box.

Under *ODBC connection string*, the ODBC connection string is defined. The same string can be found in the connection description file, under the tag [CONNECT].

2.3.3.3 The **<Build>** button allows a new connection string to be defined. This will call the Windows setup dialog box to create a new ODBC connection (see also 2.3.2.9).

2.3.3.4 By pressing the **<Refresh>** button, the connection between InfoQuest FP and the connected database is refreshed. A tree-like table structure view of the database is displayed in the upper right panel.

2.3.3.5 The database type can be selected under *Database (Access, SQL Server, and Oracle)*. This information is written under [DATABASETYPE] in the connection description file.


The second panel in the *Connected databases configuration* dialog box concerns the tables of the connected database. InfoQuest FP assumes a certain table structure to be able to store its different kinds of information. Each table should contain a set of columns with fixed names. This table structure is described in detail in . In a database setup where InfoQuest FP is connected to an existing database system, views can be created with table names that correspond to the required InfoQuest FP tables, and that have the required InfoQuest FP columns. To add flexibility, however, it is also possible to select different table names than the default ones. This allows one to create additional views, for example, where certain information is shown or hidden. These views can be saved under a different name, and specific views can be made visible to users with specific permissions.

The right panel relates to character types. Each character type is stored in two separate tables. One table, **<CharacterTypeName>FIELDS**, contains the field (i.e., character) names. When connected databases are used, characters can be described by more than one information field (see 3.3.4). The name field and the additional fields are stored in columns in this table. The second table, **<CharacterTypeName>**, contains the character values for the entries. Both tables can also be chosen under *Values* and *Fields*, respectively. The default names are **<CharacterTypeName>** and **<CharacterTypeName>FIELDS** (**<CharacterTypeName>** being the name of the character type).

Under *Restricting query*, there is a possibility to enter a query that restricts the number of entries in the database to those that fulfill a specific query. The use of restricting queries is explained further in 2.3.9.

The option *Prompt at startup* allows the user to enter or choose a restricting query at startup when the connected database is loaded. The program prompts with a user-friendly graphical query builder similar to the advanced query builder described in 2.2.10. The use of the *Prompt at startup* query builder is discussed in 2.3.9.

With the option *Experiment order statement*, it is possible to define a specific order for the experiments to show up in the *Experiments* panel in the *InfoQuest FP main* window. By default, the experiments are listed alphabetically, which is indicated by the default SQL string "ORDER BY [EXPERIMENT]". [EXPERIMENT] refers to the column EXPERIMENT in table EXPERIMENTS (see 7.1.15), which holds the names of the experiments. This means that the experiments will be sorted by their name. It is possible to add an extra column to this table, with information entered by the user, for example an index number. If this column is specified in the SQL string, the experiments will be ordered by the index.

In *Source file location*, the path for storing the source files (TIFF images and .CRV curve files) is entered. This is only used when this information is not stored in the database itself (see next paragraph). The path can be a local directory or a network path, for example on a server computer. To change the path, click  to browse through the computer or the network.

When a connected database is automatically created (see 2.3.2.5), fingerprint files (TIFF files, CRV files) are always saved in the connected database. In this case, the option *Store fingerprint files in database* is checked and grayed so it cannot be changed by the user (see Figure 2-25). For custom created connected databases (see 2.3.2.6 to 2.3.2.10), the user has the choice whether to store the fingerprint files into the connected databases (default) or in the sourcefiles directory. As opposed to earlier versions of InfoQuest FP, contig projects are always saved in the connected database. For the trace files from automated sequencers (four-channel sequence chromatogram files), the user has the choice between linking to the original path of the files or storing them in the database, using a checkbox *Store trace files in database* (enabled by default). The trace files are stored in column **DATA** of table **SEQTRACEFILES** (see 7.1.20). In case *Store traces in database* is not checked, the column **DATA** will hold a link to the original path they were loaded from.

Also for contigs associated with sequences in a connected database, it is possible to display the contig status by checking *Display sequence contig status*. When checked, the program shows the presence of a contig file as well as an Approved flag (see 3.4.3.77 to 3.4.3.78).

With the checkbox *Use as default database*, the database can be specified to be the default connected database or not. **Once a database is specified to be the default connected database, it cannot be disabled anymore!**

Two buttons, **<Check table structure>** and **<Auto construct tables>**, allow one to check if all required tables and fields are present in the connected database, and to automatically insert new tables and fields where necessary, respectively.

WARNING: When pressing *<Auto construct tables>*, InfoQuest FP will automatically create a new table for every required table that is not yet linked to an existing table in the database. For tables already linked, it will insert all required fields that do not yet exist in the database. In case you want to link InfoQuest FP to an existing database, this may cause a number of tables and fields to be created and cause irreversible database changes! Solutions to link InfoQuest FP to existing databases having different table structures are explained in 2.3.6.

2.3.4 Working in a connected database

Once a connected database is correctly set up, adding, processing and analyzing data is nearly identical to working in a local database. For entries stored in the connected database, the *Location* information field displays *Shared*.

NOTES:

(1) When entry information fields are obtained from a view (query) in the connected database, it will not be possible to define new information fields directly from InfoQuest FP. In that case, you will have to create the field in Oracle, SQL Server, PostgreSQL, or Access, add it to the view, and reload the InfoQuest FP database (see 2.3.7.14).

(2) Certain characters, for example a period, that are allowed in column names in a InfoQuest FP database, may not be allowed in the connected database. We refer to the manual of the database system for more information.

(3) Views with joined columns may be read-only and it may not be possible to add new records to the database that are seen through these views (e.g. entries, experiments). It is possible to bypass this in Oracle or SQL Server using triggers.

There are a few differences, however, concerning (1) adding new entries to the database, (2) the default way of storing images and contig information, and (3) the way log files are recorded and viewed.

- When adding new entries to the database using the menu command *Database > Add new entries*, the choice is offered to add the entries to the local database or to the connected database. When no connected database is the **default** database, you will be able to choose between these two possibilities. Once a connected database is specified as the default database, however, it will only be possible to add new entries to the connected database.
- In a standard connected database setup, images and contig projects are stored within the connected database itself. If you have specified not to save this information in the connected database (see 2.3.3), the files are saved in a common directory **Sourcefiles** under the local database directory. The default directory for such files is denoted in a relative way as

[DBPATH]\sourcefiles. [DBPATH] hereby refers to the database folder in the InfoQuest FP home

directory, as specified under Settings () in the Startup screen.

- Within **Sourcefiles**, there are three subdirectories: **contig**, **images** and **gel2d**. In case the fingerprint files are not saved in the connected database, the **images** subdirectory will contain the TIFF files for fingerprint types. The TIFF files are placed in this directory using the command *File > Add experiment file* in the *InfoQuest FP main* window. To make a gel TIFF file visible in the *Files* panel in a connected database, the file should be present in this directory (or stored in the connected database, which is the default option). Under **contig**, Assembler contig (sequence assembly) projects can be saved. This will only be the case if no column Contig is present in the connected database; otherwise (and by default), contigs are saved in the connected database (see 7.1.21). The source file directory can be modified as described in 2.3.3. The path can be a network path, for example on a server computer. The **gel2d** subdirectory contains the TIFF files for the 2D gel types.
- Log files are stored in a different way in a connected database. The log events are stored in a database table called EVENTLOG. Different events are stored under different categories: **Database** concerns all actions affecting the database (adding, changing or removing information fields, adding experiment types, adding entries, changing entry information, etc.). Furthermore, there is a category **EXPER_<ExperimentName>** (<ExperimentName> being the name of the experiment), relating to changes made to the experiment type (i.e. normalization settings in case of a fingerprint type, adding, removing, or renaming characters in a character type, etc.). A third category reports on changes made to the data in a certain experiment type. In this category, components have the name of the experiment type.
- The *Event log* window (Figure 2-6) called from the *InfoQuest FP main* window offers the possibility to view the log file for a connected database or the local database under **Database**. Under **Component**, you can choose to view a specific component, e.g. Database, an experiment type, or data belonging to an experiment type. With **All**, you can view all components together, listed chronologically. The components can only be selected when a connected database is viewed.

2.3.5 Linking to an existing database with standard InfoQuest FP table structure

Any computer running InfoQuest FP can link up to an existing InfoQuest FP connected database at any time. When this connected database has its table structure in the standard InfoQuest FP format (see Section 7.1), this can be done very easily in the Startup program.

2.3.5.1 In the InfoQuest FP Startup program, click the



button to create a new database.

2.3.5.2 Enter a name for the connected database (this can be a different name on different computers).

2.3.5.3 In the next step, choose **<Yes>** to automatically create the required directories, since a local database associated with the connected database is required.

2.3.5.4 In the next step, choose to whether or not create log files, and press **<Finish>**.

A new dialog box pops up, prompting for the type of database: *New connected database (automatically created)*, *New connected database (custom created)*, *Existing connected database*, or *Local database (single user only)* (Figure 2-23).

2.3.5.5 Select *Existing connected database* and press **<Build>** to establish the connection to the database.

2.3.5.6 The dialog box that pops up now is generated by your Windows operating system and may differ depending on the Windows version installed. Therefore we refer to the Windows manual or help function to select or create a DSN file (ODBC Data Source) that specifies the ODBC driver, and to set up a connection to the database.

2.3.5.7 Once the database connection is defined, you can press **<OK>** to quit the database setup.

The *Plugin installation* window pops up, from which you can install the available plugins. For more information on the use of plugins, see 1.5.3.

2.3.5.8 Press **<Proceed>** in the *Plugin installation* window to open the *InfoQuest FP main* window with the newly created, blank database.

The connected database will now be the default database. If the connected database contains the standard table structure for InfoQuest FP (see 7.1), no error message is produced and you can start working immediately. InfoQuest FP will automatically recognize the existing information fields, experiment types, subsets, entries and data. If the table structure is not in standard InfoQuest FP format however, a dialog box appears, warning for several errors that have occurred while trying to open specific tables in the connected database that were not found. See 2.3.6.10 and further to assign the correct tables or views from the database.

2.3.6 Linking to an existing database with table structure not in InfoQuest FP format

This paragraph describes the situation where an Oracle, SQL Server, PostgreSQL, or Access database, containing

descriptive information on organisms (entries) and/or experiment data is already present and InfoQuest FP should be hooked up to that database in order to read and write experiment data and information fields.

Before proceeding with the configuration of the database connection, it will be necessary to make the database compatible with the InfoQuest FP table structure. In a typical case, a number of information fields and/or experiment fields from the connected database will need to be linked to InfoQuest FP. However, these fields will occur in different tables having different field names. The obvious method in this case is to create *views* (or, in Access, *queries*) in the database.

- For those InfoQuest FP tables for which the connected database contains fields to be used, a view (query) should be constructed in the database. Within that view (query), those database fields that contain information to be used by InfoQuest FP should be linked to the appropriate field.
- InfoQuest FP tables for which the connected database contains no fields can be created automatically by InfoQuest FP.
- Finally, the database should be configured in such a way that the InfoQuest FP tables that contain fields already present in the database, be present either as table or as view, with all the recognized field names as outlined in . The names for the tables or views, however, can be freely chosen.
- Additional tables required by InfoQuest FP for which there are no fields available in the database can be created automatically by InfoQuest FP.

NOTE: When views are created in the database, to match the required InfoQuest FP tables, it is recommended to name the views using the standard InfoQuest FP names for the required tables. This will allow new users to log on to an existing connected database in the easiest way, by just defining the connection in the Startup program (2.3.5). By using different names, new users will have to specify the table/view names manually in the Connected database configuration window (Figure 2-25) after defining the connected database. Using different names for the views is only useful if it is the intention to assign different permissions to different users; in this way, views can be created showing only restricted information, while other views show full information, etc.

2.3.6.1 In the InfoQuest FP Startup program, click the



button to create a new database.

2.3.6.2 Enter a name for the new database.

2.3.6.3 In the next step, choose **<Yes>** to automatically create the required directories, since a local database associated with the connected database is required.

2.3.6.4 In the next step, choose whether or not to create log files, and press **<Finish>**.

A new dialog box pops up, prompting for the type of database: *New connected database (automatically created)*, *New connected database (custom created)*, *Existing connected database*, or *Local database (single user only)* (Figure 2-23).

2.3.6.5 Select *Existing connected database* and press **<Build>** to establish the connection to the database.

2.3.6.6 The dialog box that pops up now is generated by your Windows operating system and may differ depending on the Windows version installed. Therefore we refer to the Windows manual or help function to select or create a DSN file (ODBC Data Source) that specifies the ODBC driver, and to set up a connection to the database.

2.3.6.7 Once the database connection is specified, you can press **<OK>** to quit the database setup.

The *Plugin installation* window pops up, from which you can install the available plugins. For more information on the installation of plugins, see 1.5.3.

2.3.6.8 Press **<Proceed>** in the *Plugin installation* window to open the *InfoQuest FP main* window with the connected database.

Since the connected database does not contain the standard table structure for InfoQuest FP (see 7.1), a dialog box now appears, warning for several errors that have occurred while trying to open specific tables in the connected database that were not found.

2.3.6.9 Press **<OK>** to close the message(s). The *InfoQuest FP main* window shows a blank database.

In the *InfoQuest FP main* window, you can now configure the database connection as described in 2.3.3:

2.3.6.10 Select **Database > Connected databases**.

This opens a list of all currently defined connected databases for this InfoQuest FP database (normally just one; see Figure 2-24).

2.3.6.11 Select the connected database of choice and click **<Edit>**, or double-click on the name.

This opens the *Connected database configuration* dialog box (Figure 2-25). This dialog box shows the default suggested table names for the required database components under **Database tables** (see 2.3.3). Some, or all, of these tables do not correspond to the tables of the database.

2.3.6.12 Press the **<Refresh>** button. The upper right panel now lists the tables and views in the connected database, as it exists.

2.3.6.13 You can expand each table/view to display its fields by clicking on the “+” sign on the tree.

2.3.6.14 Under **Database tables**, select the corresponding table or view for each component.

2.3.6.15 When this is finished, check the correspondence by pressing **<Check table structure>**.

When required, you can further configure the database, leaving the *Connected database configuration* dialog box open. As soon as the new configuration is done, press **<Refresh>** and check the table structure again.

2.3.6.16 Finally, when all links to existing database tables/views are made correctly, you can allow InfoQuest FP to create additional tables for which there are no fields available in the external database, by pressing **<Auto construct tables>**. InfoQuest FP will now only construct tables that are not yet linked, and fields that are not yet present in the connected tables.


NOTE: It will not be possible for InfoQuest FP to create new fields within a view/query. In that case, you will have to create the field in Oracle, SQL Server or Access, add it to the view, and reload the InfoQuest FP database.

2.3.7 Converting a local database to a connected database

In order to take full advantage of all features available in InfoQuest FP, the user may want to convert a previously created local database to a connected database. There are two options available for this conversion:

1. Exporting all entries from the local database to XML files and importing these XML files in a new connected database
2. Setting up an ODBC connection and converting the local data to the connected database via a function available in InfoQuest FP.

The first procedure is the safest way of working and is therefore recommended. It does, however, require the *Database sharing tools* module to be present. To check whether you have the *Database sharing tools* module, open any database and select **File > About** (see 1.1.5).

• **Option 1: Using the XML Tools (requires the Database sharing tools module ).**

2.3.7.1 Open the local database that you want to convert. In the *InfoQuest FP main* window, select **File > Install/Remove plugins** and install the XML Tools plugin (see 1.5.3 on the installation of plugins).

2.3.7.2 In the *Database entries* panel of the *InfoQuest FP main* window, click on the first database entry and, while holding the SHIFT key, click on the last entry to select all database entries. Alternatively, press CTRL+A on the keyboard.

2.3.7.3 Select *File > Export selection as XML*. The *Export data to XML* dialog box appears (see Figure 2-26).

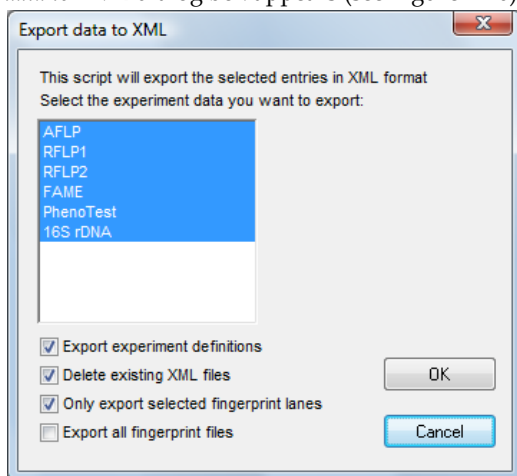


Figure 2-26. The *Export data to XML* dialog box from the XML Tools plugin.

2.3.7.4 Leave all experiment types selected, uncheck *Only export selected fingerprint lanes* and check *Export all fingerprint files*. Next, press **<OK>** to start the creation of the XML files.

The complete database information is now exported to XML files. These XML files are stored in the subfolder **Export** of the database folder.

NOTE: This procedure also allows the user to convert only a part of the local database information to a new connected database, by selecting a subset of the database entries in step 2.3.7.2 and/or selecting a subset of the available experiment types in step 2.3.7.4.

In case the database contains a fingerprint type based on two-dimensional gels, the original TIFF files also need to be exported.

2.3.7.5 With all database entries still selected, select *File > Export TIFF files for selected entries*. The *Export TIFF files* dialog box pops up (see Figure 2-27).

2.3.7.6 Leave all TIFF files selected and press **<OK>**.

2.3.7.7 Close the database.

2.3.7.8 In the Startup screen, create a new, empty connected database as described in to . You can leave all settings default.

2.3.7.9 From the *Plugin installation* toolbox that appears, install the XML Tools plugin in the newly created database (see 1.5.3 on the installation of plugins).

2.3.7.10 Select *File > Import selection as XML* and browse for the **Export** folder of the exported database. In the *XML import* dialog box (see Figure 2-28), leave all settings default and press **<OK>**.

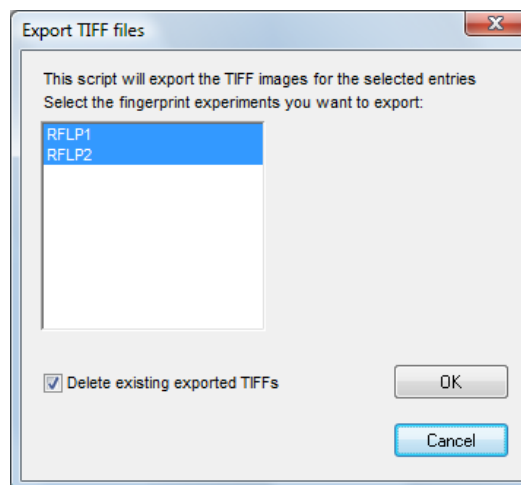


Figure 2-27. The *Export TIFF files* dialog box from the XML Tools plugin.

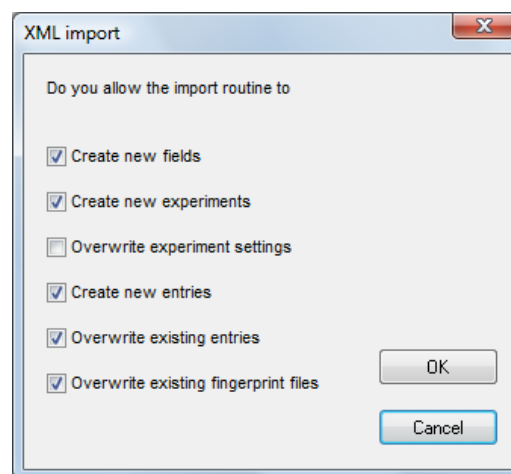


Figure 2-28. The *XML import* dialog box from the XML Tools plugin.

The database information and experiment type information, is now copied from the XML files to the connected database.

In case the exported database contained a fingerprint type based on two-dimensional gels, the TIFF files still need to be imported:

2.3.7.11 Select *File > Import TIFF files* and browse for the **Export** folder of the exported database. Select all TIFF files and press **<Open>** to import them in the connected database.

• **Option 2: Using the conversion function after setting up an ODBC connection.**

InfoQuest FP also offers the possibility to convert an entire local database at once to a new connected database, without the need for the *Database sharing tools* module. This is **an irreversible operation, which causes the local database to be removed once the conversion is done**. It is therefore strongly recommended to make a

backup copy of the local database before carrying out a conversion to a connected database.

NOTE: This procedure is not recommended to convert a local database into an existing connected database that already contains data, since experiment types with the same name would be overwritten. Converting a local database into a connected database using the XML Tools plugin as described in 2.3.7.1 to 2.3.7.11 is a better option in this case.

To convert a local database into a new connected database, proceed as follows:

2.3.7.12 Create a new empty database in Oracle, SQL Server or Access.

2.3.7.13 Open the local database in the InfoQuest FP main program.

2.3.7.14 In the *InfoQuest FP main window*, select **Database > Connected databases**.

This opens a list of all currently defined connected databases for this InfoQuest FP database (normally empty at this stage; see Figure 2-24).

2.3.7.15 Click **<New ODBC>**.

2.3.7.16 In the *Connected databases configuration* dialog box (see Figure 2-25) that appears, click **<Build>**.

2.3.7.17 The dialog box that pops up now is generated by your Windows operating system and may differ depending on the Windows version installed. Therefore we refer to the Windows manual or help function to select or create a DSN file (ODBC Data Source) that specifies the ODBC driver, and to set up a connection to the database.

2.3.7.18 Make sure the connected database is checked as the default database; otherwise, the conversion cannot be executed.

2.3.7.19 Check the table structure of the database, if it does not contain the required tables and fields, press **<Auto construct tables>** to allow InfoQuest FP to construct its tables.

2.3.7.20 Once the connection is defined correctly, press **<OK>** to close the *Connected databases configuration* dialog box.


2.3.7.21 Close the *Connected databases list* window.

2.3.7.22 In the *InfoQuest FP main window*, select **Database > Convert local data to connected database**.

An important warning message is displayed. If you are converting the local database to a NEW connected database, and if you have made a backup of the data before starting this conversion (see 2.7.1), you can safely click **<OK>** to start the conversion.

Depending on the size of the database, the conversion can take seconds to hours. Fingerprint image files take most time to convert. When the conversion is finished successfully, InfoQuest FP will automatically restart with the connected database, and the contents of the local database will be removed.

NOTE: If some information fields are not displayed in the Database entries panel after the conversion, they can be shown by clicking on the column properties

button  in the database information fields header and selecting them from the pull-down menu.

2.3.8 Opening and closing database connections

• Connecting to multiple connected databases

It is possible to connect to other connected databases in addition to the default connected database.

2.3.8.1 In the *InfoQuest FP main window*, select **Database > Connected databases**.

This opens a list of all currently defined connected databases for this InfoQuest FP database (normally just one; see Figure 2-24).

2.3.8.2 Click **<New ODBC>**.

In the *Connected databases configuration* dialog box that appears, click **<Build>**.

2.3.8.3 The dialog box that pops up now is generated by your Windows operating system and may differ depending on the Windows version installed. Therefore we refer to the Windows manual or help function to select or create a DSN file (ODBC Data Source) that specifies the ODBC driver, and to set up a connection to the database.

2.3.8.4 In the *Connected databases configuration* dialog box, enter a name for the connected database definition file (upper left input field, *Connected database*). This name should be **different** from the names of any of the existing connected databases.

2.3.8.5 Under **Source files location**, select the directory where the source files can be found for this connected database. This directory should always be different from the **Source files** directory of the default connected database.

2.3.8.6 Once the connection is defined correctly, press **<OK>** to close the *Connected databases configuration* dialog box.

NOTE: When the two connected databases have the same Source files directory associated, an error message is produced at this time: "Another connected database is already associated with this source files directory." It will not be possible to save this new

connection until the source files directories are different.

The new connected database is listed in the *Connected databases list* window. When you open the main program, the contents of the two databases are seen together.

•Closing or deleting a connected database

2.3.8.7 In the *Connected databases list* window (**Database > Connected databases**), select the connected database you want to close, and press **<Close>**.

2.3.8.8 Confirm with **<Yes>**. The database disappears from the list, and the contents of the closed database disappears from the *InfoQuest FP main* window.

Closing a connected database is temporary. When it is closed, it will automatically be reopened the next time the InfoQuest FP main program is started up with the same database.

To delete a connection to a database, press **<Delete>** in the *Connected databases list* window. The connected database will never reappear until you build the connection again.

•Reloading a connected database

Suppose you have modified the connected database directly in Oracle, SQL Server or Access, you can use the function **<Reload>** in the *Connected databases list* window. Any columns that were added, for example as information fields, or any entries or data that were added externally after InfoQuest FP was started up will be updated in the *InfoQuest FP main* window.

Reloading a connected database can also be useful in case several persons are working in the database simultaneously. Any entries added by other persons will not be seen in your session until you reload the database.

2.3.9 Restricting queries

When massive databases are generated, loading the full database into InfoQuest FP might become quite time-consuming and unnecessary for most purposes. To that end, it is possible to load a connected database in InfoQuest FP using a *restricting query* (see also 2.3.3). A restricting query is an SQL query that is used to load only those database entries that comply with the query statement. There are two possibilities of using restricting queries, each serving a more or less different purpose:

1. An automatic query specified in the *Connected databases configuration* dialog box, which will apply each time InfoQuest FP is started up.
2. An interactive one which prompts the user to build a query when InfoQuest FP is started up. Such queries

can be saved in query templates and modified or reused at any time. The queries can be built either using a user-friendly graphical query builder, or by typing an SQL query directly, or by combining both.

•Automatic restricting queries

This type of restricting query is particularly useful if you want to work with only one specific group or taxon from the database. For example, if you have a database with a number of species, of which you want to work with only one, you can use the field "Species" to apply a restricting query "Species=...". As a result, only those entries having the specified string in their Species field are loaded. In addition, when new entries are created, they will automatically have the species field filled in. This can save time, help avoid typing errors and restrict users to specific groups of the database.

To specify an automatic restricting query, a restricting query is entered in the input field **Restricting query** of the *Connected databases configuration* dialog box (Figure 2-25). A restricting query is of the general format **FieldName=String**. **FieldName** is the name of the field that the restriction is applied to, and **String** is the restricting string. As a result, when the InfoQuest FP main program is opened with the connected database, only those entries having **String** filled in the field **FieldName** will be seen in the database.

In addition, when new entries are added to the database, they will automatically have their field **FieldName** filled with **String**.

To try out this feature, you can e.g. install the **DemoBase_SQL** database, as described in . A restricting query to visualize only *Ambiorix* can be entered as follows.:

2.3.9.1 In the *Connected databases configuration* dialog box under Restricting query, type:

```
GENUS=Ambiorix
```

2.3.9.2 Press **<OK>** to confirm the changes. The *InfoQuest FP main* window now only shows *Ambiorix*.

2.3.9.3 Add a new entry with **Database > Add new entries**. The new entry is automatically called *Ambiorix* in its **Genus** field.

Restricting queries can be combined by separating them with semicolons. For example, if you want to visualize only *Ambiorix sylvestris* entries, enter the following as a restricting query:

```
GENUS=Ambiorix;SPECIES=sylvestris
```

The result is a database that only shows *Ambiorix sylvestris*. New entries will automatically be added as *Ambiorix sylvestris*.

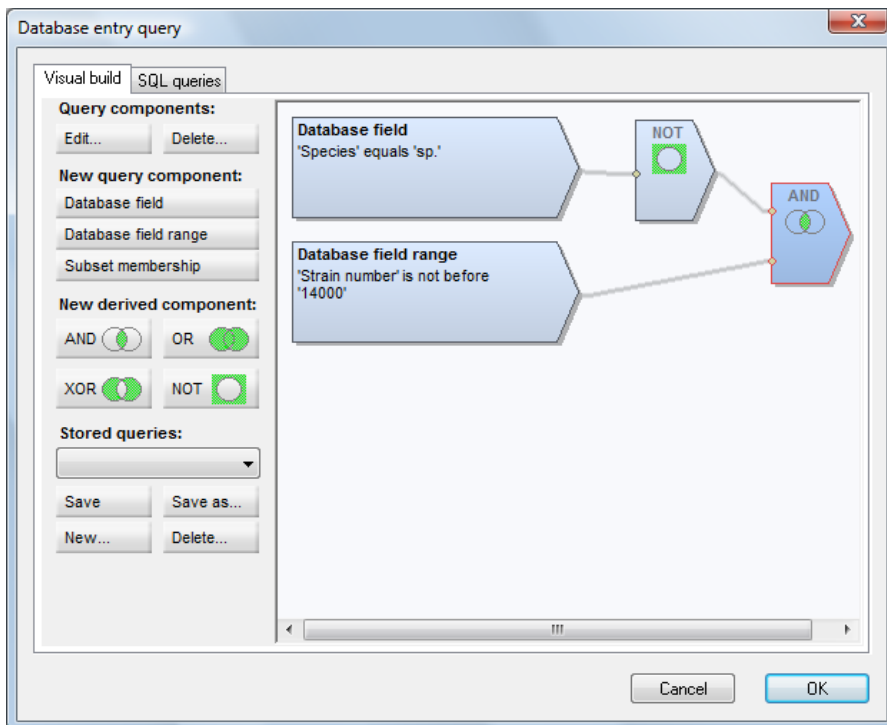


Figure 2-29. The Interactive query builder, prompting at startup.

NOTES:

(1) Do not use spaces in a restricting query.

(2) If you open a database using a default restricting query, you may not be able to work with gel files, comparisons, libraries, or subsets that contain entries which are not loaded by the restricted query. The program will generate an error message that one or more keys are not present or not loaded in the database. Using the interactive restricting queries, however, missing entries can be loaded during the session if requested (see below).

•Interactive restricting queries

The aim of this type of startup queries is to be able to restrict database loading in a flexible way each time the program is started up. Another source of flexibility in this option is the fact that the software can load additional entries dynamically whenever a gel file, a comparison, a subset, or a library is opened that contains entries which were not loaded by the restricting query used.

2.3.9.4 The interactive queries can be activated by checking **Prompt at startup** in the *Connected databases configuration* dialog box (Figure 2-25). As a result, each time the program is started up, an interactive graphical query builder pops up (Figure 2-29).

2.3.9.5 By pressing <OK> without entering any restricting query, the complete database is loaded.

The interactive restricting query tool is very similar to the *Advanced query tool* described in 2.2.10. It allows you to create individual *query components*, which can be combined with *logical operators*. The available targets for

query components are *Database field*, *Database field range*, and *Subset membership*.

•Database field

Using this component button, you can enter a (sub)string to find in any specific field that exists in the database (Figure 2-30). Note that wildcard characters are not used in this query tool and that the string entered has to match completely with the field contents. The queries are not case sensitive.

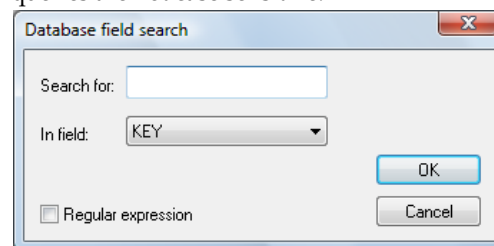


Figure 2-30. Database field search component dialog box.

A search string can also be entered as a regular expression (see 7.2).

•Database field range

Using this component button, you can search for database field data within a specific range, which can be alphabetical or numerical. Specify a database field, and enter the start and the end of the range in the respective input boxes (2.3.7). A range should be specified with the lower string or value first. Note that, when only one of both limits is entered, the program will accept all strings above or below that limit, depending on which limit was

entered. For example, when only the first (lower) limit of the range is entered and the upper limit is left blank, all strings (values) *above* the specified string (value) will be accepted.

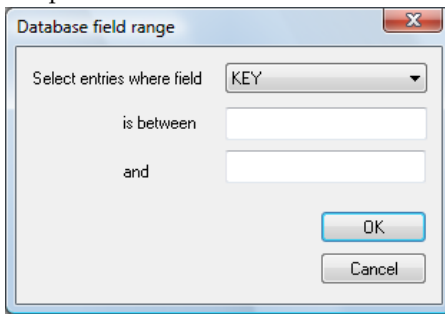


Figure 2-31. Database field range component dialog box.

•Subset membership

With this search component, you can specify that only entries belonging to a certain subset should be loaded (Figure 2-32). This option offers additional flexibility as subsets can be composed of any selection of database entries and are not necessarily bound to global query statements.

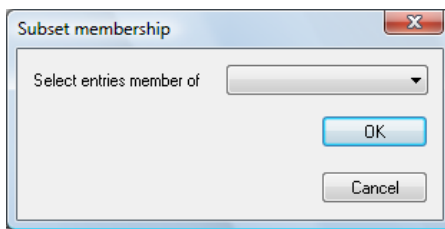


Figure 2-32. Subset member component dialog box.

•Logical operators



NOT, operates on one component. When a component is combined with NOT, the condition of the component will be inverted.



AND, combines two or more components. All conditions of the combined components should be fulfilled at the same time for an entry to be selected.



OR, combines two or more components. The condition implied by at least one of the combined components should be fulfilled for an entry to be selected.



XOR, combines two or more components. Exactly one condition from the combined components should be fulfilled for an entry to be selected.

NOTES:

(1) The buttons for the logical operators contain a helpful Venn diagram icon that clearly explains the function of the operator.

(2) An example on the use of the logical operators is given in 2.2.10 for the graphical query builder.

Note that:

- Individual components can be re-edited at any time by double-clicking on the component or by selecting them and pressing **<Edit>**.
- Selected components can be deleted with **<Delete>**.
- The result of a logical operator as obtained in this example can be combined again with other components (or logical operators) to construct more complex queries.
- Queries can be saved with **<Save>** or **<Save as>**.
- Saved queries can be loaded using the pull-down listbox under **Stored queries**.
- Existing queries can be removed with loading them first and pressing **<Delete>**.

2.3.9.6 To view the selected entries, press **<Add to list>**.

The entries that were found are highlighted with a colored arrow left from them.

NOTES:

(1) When combined with a logical operator, query components contain a small node at the place where they are connected to the logical operator box (AND, OR, XOR). By dragging this node up or down, you can switch the order of the query components, thus making it possible to move the most efficient component to the top in AND combinations, as explained above.

(2) Multiple components/operators can also be selected together by dragging the mouse over the boxes in the right panel.

2.3.9.7 The second tab of the interactive restricting query builder, **SQL queries**, contains the actual SQL query statements translated from the active query (Figure 2-33). These SQL statements are passed on to the database to obtain the restricted view.

In principle, the user can compose queries or make changes directly in these fields. This is however not recommended unless you are very familiar with both the SQL language and the InfoQuest FP database table structure. Incorrect SQL query inputs can lead to information partially not being downloaded from the database and might eventually cause the database to become corrupted in case attempts are made to save changes.

When a database is opened with a restricting query, it may occur that an analysis is done which contains entries that are not loaded in the current view. This can happen with gel files, comparisons, subsets, or library units. If such a situation occurs, the program will first

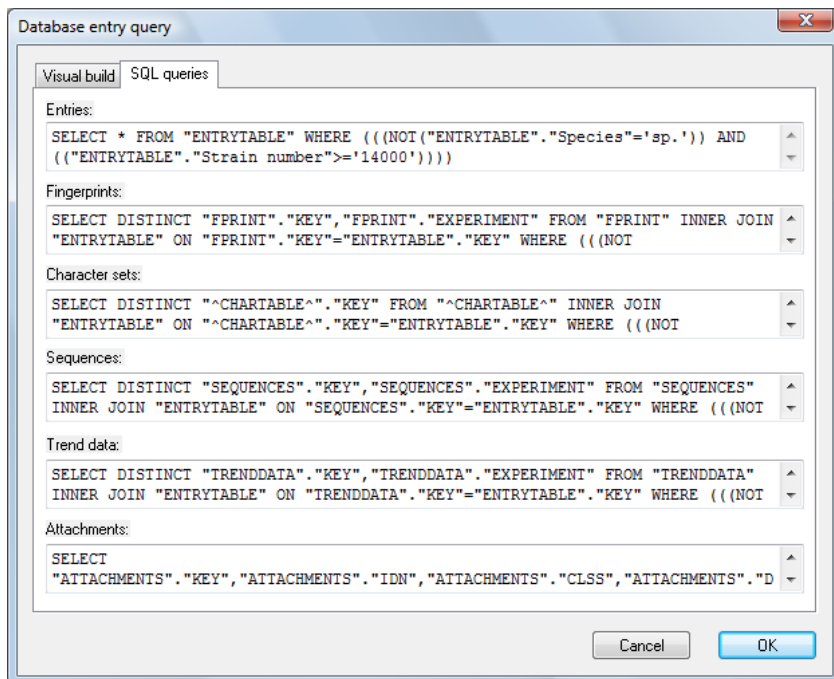


Figure 2-33. SQL query statements translated from a visual query build.

generate an error message that one or more keys are not present or not loaded in the database. Next, the program will propose to try to fetch the entries from the database. If you answer <Yes> the entries will be loaded dynamically from the connected database. In case of a gel file or a subset, this can technically be achieved very quickly. In case of a comparison, however, the operation requires an SQL command to be launched for each additional entry to download. In case of large numbers of additional entries, but also depending on the size of the database and several other factors, this may take considerable time. Therefore, the number of entries to fetch is indicated in the confirmation box (Figure 2-34).

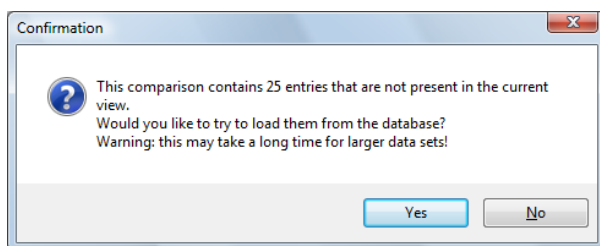


Figure 2-34. Confirmation box to download additional database entries into a comparison.

In case the download time to complete a comparison becomes critical, there is a simple workaround by creating or opening a subset and using the feature *File > Add entries to current subset* in the *Comparison* window (see also 4.1.7). As soon as this command is executed, the entries from the comparison are added to the current subset and the program automatically retrieves the non-loaded entries. It will prompt to load the entries into the database, and the comparison will be complete at once.

Since library units (see 5.2.1) have physically the same structure as comparisons, the same constraints apply. However, a library unit will never consist of thousands of entries as can be the case with comparisons.

2.3.10 Protecting connected databases with a password

Connected databases can be protected by the use of a password.

Access database:

2.3.10.1 Open MS Access and select the 'Open' command in the menu of Access.

2.3.10.2 In the *Open file* dialog box, navigate to the connected database. Click the arrow to the right of the Open button and choose the option '*Open Exclusive*' (see Figure 2-35).

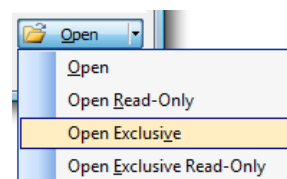


Figure 2-35. Open database for exclusive use.

2.3.10.3 If you are using MS Access 2000 or 2003, go to the *Tools* menu, and select *Security option > Set database password*. If MS Access 2007 is installed on your

computer, select the *Database Tools* tab and select *Set Database Password*.

2.3.10.4 A dialog box pops up, asking you to enter and confirm your password (see Figure 2-36).

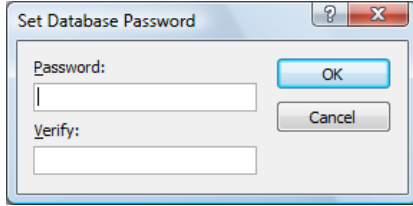


Figure 2-36. *Set Database Password* dialog box.

If you close the database in Access and open the database in InfoQuest FP, the program will prompt you for the specified password before loading the database.

Other connected databases:

For all other connected databases (SQL Server, Oracle, ...), a username and password are required upon creation of the database. The reason why InfoQuest FP does not prompt for it when loading the database, is because the password is saved in the ODBC string.

2.3.10.5 Open the database and select *Database > Connected databases*.

The line "**PWD=*password***" holds the password (see Figure 2-37).

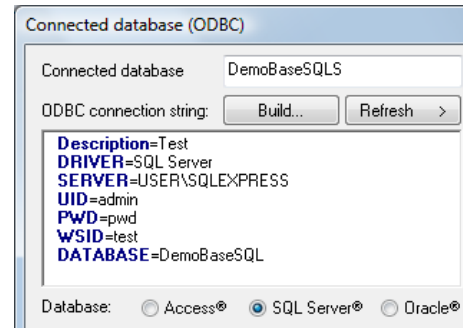


Figure 2-37. The ODBC connection string.

If you want InfoQuest FP to prompt for a username and a password each time you open the database, delete the line "**PWD=*password***" in the ODBC connection string.

If you want a specific username to be filled in the username box, change the username after "**UID=**".

2.4 Levels and relations in a database

2.4.1 Introduction

In a default InfoQuest FP database, all entries belong to the same level or category. Some applications, however, require a more advanced database structure, where entries belong to different hierarchies or levels. In a connected database, InfoQuest FP offers this possibility by introducing the *Levels*. Levels are hierarchical layers in the database, with the purpose to store and represent entries of different categories in a better organized way.

NOTE: The features described in this section are only available in a connected database; they cannot be used in a local database (see 2.1.1).

The meaning and utility of *Levels* can best be explained with the following example: In a clinical lab, samples are regularly obtained from patients (e.g. blood, skin,...). From these samples, fingerprints (profiles) are generated using MALDI as technique. In this context, there are three levels of entries to which information fields and data can be assigned: the *patients*, the *samples*, and the *profiles*. In a flat InfoQuest FP database setup, one would create a new entry for each profile generated and enter patient and sample-specific information in dedicated information fields. For example, the patient could be described in a set of fields "Patient name", "Patient age", "Patient gender" etc. The same can be done for sample-specific information. It is clear that, for a number of reasons, this is not the most elegant approach for building a levelled database.

- Patient information is unnecessarily duplicated over all samples/profiles;
- If patient information is to be added or changed, it has to be added for all samples/profiles;
- There is no formal way of linking profiles to patients/samples, except by filling in an information field. In case of a typing error, the link is lost.
- There is no framework to deal with duplicate runs (e.g. averaging, standard deviations).

In addition to the concept of *Levels*, InfoQuest FP also introduces the concept of *Relations*. As the name tells it, a *Relation* can define the relation between entries belonging to different levels. Using the same example of patients, samples and profiles, the interaction between levels and relations is illustrated in Figure 2-38. The database consists of 3 levels, Patients, Samples and Profiles. Each *Level* has specific information fields associated, e.g. for Patients: *Name*, *Gender*, *Birth date*. Multiple samples can be obtained from one patient, as illustrated

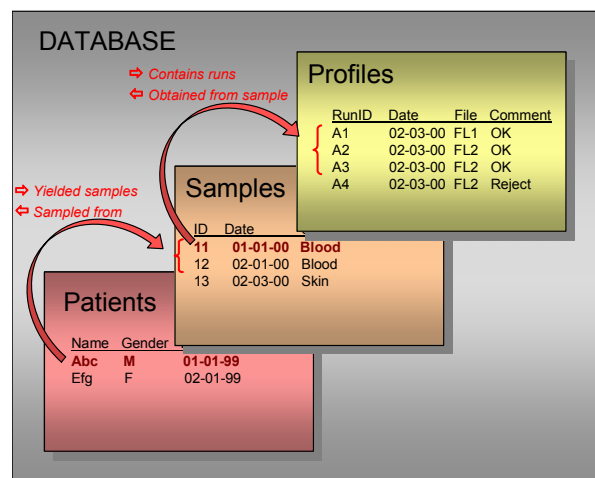


Figure 2-38. Scheme illustrating the use of *Levels* and *Relations* in an InfoQuest FP database.

in Figure 2-38: samples 11 and 12 are both obtained from patient Abc (red). The link between the two samples and the patient is provided by a *Relation*. The relation has a *one-to-many* forward description (from patients to samples) called "Yielded samples", and a *many-to-one* reverse description (from samples to patients) called "Sampled from". Similarly, 3 profiles were obtained from sample 11 (red) in the figure. The link between the sample and the profiles is provided by a second *Relation*, which has a forward description (from samples to profiles) called "Contains runs", and a reverse description (from profiles to samples) "Obtained from sample".

In this way, samples are unambiguously assigned to patients, and profiles to samples, without information being duplicated. However, there is more to be achieved with the Levels/Relations construction. It becomes for example possible to calculate average MALDI profiles based upon all MALDI profiles that belong to the same sample. Or, one can also calculate and fill in the experiment type for MALDI profiles at the level of Samples, also by averaging the profiles obtained for each sample.

Most of the functionality provided by levels and relations is to be filled in using scripts and is therefore either provided as plugin tools by Bio-Rad, or to be programmed by an experienced InfoQuest FP user. However, it is possible to create levels and define relations using the InfoQuest FP window menus. We will illustrate a generic implementation of levels and relations by creating a database similar to the above example.

2.4.2 Creating levels

2.4.2.1 Create a new connected database as described in to . Call the database **HumanTyping**, for example.

In the *InfoQuest FP* main window, a single *All levels* tab is displayed at the bottom of the window (see Figure 1-15).

2.4.2.2 Either by clicking the right mouse button inside the *Levels* tab bar or by choosing the menu **Database > Levels**, you can add a new level with **Add new level**.

It is recommended to add new levels in the order of hierarchy, e.g. the deepest level first. In this example, we will add levels *Patients*, *Samples* and *MALDI profiles*, in this order.

2.4.2.3 Enter **Patients** as level name and press **<OK>**.

2.4.2.4 Repeat 2.4.2.2 and 2.4.2.3 for two more levels: **Samples** and **MALDI profiles**.

NOTE: Maximally four levels can be created, in addition to "All levels".

When finished, the three levels are displayed as tabs (see Figure 2-39).

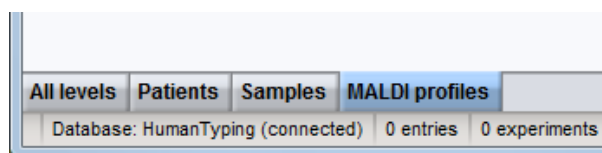


Figure 2-39. Levels shown as tabs in the database.

2.4.2.5 Click on the *Patients* tab to display the *Patients* level.

2.4.2.6 Add a few patient-specific database fields, such as *Patient name*, *Gender*, *Birth date*, *City*, *ZIP code*. See 2.2.2 for instructions on how to create new database fields.

2.4.2.7 Click on the *Samples* tab to display the *Samples* level.


Note that the information fields defined under *Patients* are not shown in this level.

2.4.2.8 Add a few sample-specific fields, for example *SampleID*, *Sample Source*, *Sample Type*, *Sampling Date*.

2.4.2.9 Click on the *MALDI profiles* tab. Again, no information fields are shown for this level.

2.4.2.10 Enter some MALDI run specific fields, such as *ProfileID*, *RunDate*, *FileName*, *MicroplatePos*.

The database levels are now configured; we will now add a few entries.

NOTE: Physically, the entries from different levels are stored in the same table, and as such, have the same information fields. By default, however, only information fields created within a level are shown in that level. If you want to display or hide fields in a certain level, press the  button and select the fields to switch on or off.

2.4.2.11 Click on the *Patients* tab to display the *Patients* level and select **Database > Add new entries** or press the



button. Enter the number of entries you want to create, e.g. 2, and press **<OK>**. For more information on adding database entries, see 2.2.1.

2.4.2.12 Enter some fictitious information, e.g. such as displayed in Figure 2-40. See 2.2.3 for available options to enter database information.

Patients

Key	Patient name	Gender	Birth date	City	ZIP code
HumanTypin...	Rosa Bloum	F	1982-10-03	BigCity	80000
HumanTypin...	Tiss Prutz	M	1974-01-20	SmallTown	67840

Samples

Key	SampleID	Sample Source	Sample Type	Sampling Date
HumanTypin...	LVM-3A	Sputum	50574	2007-05-02
HumanTypin...	LVM-4A	Sputum	E17_6	2007-05-02
HumanTypin...	LVK-2	Skin	50574	2007-04-29
HumanTypin...	LVK-3	Skin	E17_6	2007-04-29

MALDI profiles

Key	ProfileID	RunDate	FileName	MicroplatePos
HumanTypin...	LVM-3A-001	2007-05-03	LVM-3A-001.rec	B2
HumanTypin...	LVM-3A-002	2007-05-03	LVM-3A-002.rec	B3
HumanTypin...	LVM-3A-003	2007-05-03	LVM-3A-003.rec	B4
HumanTypin...	LVM-4A-001	2007-05-04	LVM-4A-001.rec	D8
HumanTypin...	LVM-4A-002	2007-05-04	LVM-4A-002.rec	D9

Figure 2-40. Example of some fictitious entries to explain levels and relations.

Repeat steps 2.4.2.11 to 2.4.2.12 to create four new entries in the *Samples* level and five new entries in the *MALDI profiles* level. Each time, the fictitious information in Figure 2-40 can be entered.


2.4.3 Creating new relation types

In the default configuration of the *InfoQuest FP* main window, the *Entry relations* panel is available as a second tab in the *Experiments* panel (see Figure 1-15). If this is not the case, you can either restore the default configuration with **Window > Restore default configuration**, or locate the *Entry relations* panel in your current configuration. If it is not available, you can visualize it with

Window > Show/hide panels and enabling *Entry relations*. The *Entry relations* panel looks as in Figure 2-41.



Figure 2-41. The *Entry relations* panel.

2.4.3.1 To create a new relation type, press the  button or select *Database > Relations > Add new relation type*.

The *Entry relation* dialog box (Figure 2-42) allows you to define the **Relation type** and the **Relation description**.

2.4.3.2 To define a relation type between patients and samples, select **Patients** in the left pull-down box, and **Samples** in the right pull-down box.

2.4.3.3 Since one patient can provide multiple samples, define the relation as **One** to **Many**.

You can fill in a *Forward description* and a *Reverse description*. Although not required, such descriptions will help understand the relation types when working with the database.

2.4.3.4 As a forward description, enter for example “*To Samples*”, and as a reverse description, e.g. “*From Patient*”.

NOTE: One could also use more descriptive forward and reverse descriptions, for example “Yielded samples” and “Sampled from patient”.

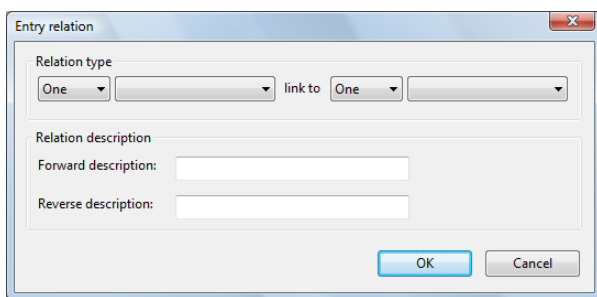


Figure 2-42. The *Entry relations* dialog box.

2.4.3.5 Likewise, create a second relation type from **Samples** to **MALDI profiles**, again as **One** to **Many**. As forward and reverse descriptions, you can enter “*To MALDI profiles*” and “*From Sample*”, respectively.

When finished creating these two relation types, the *Entry relations* panel looks as in Figure 2-43.

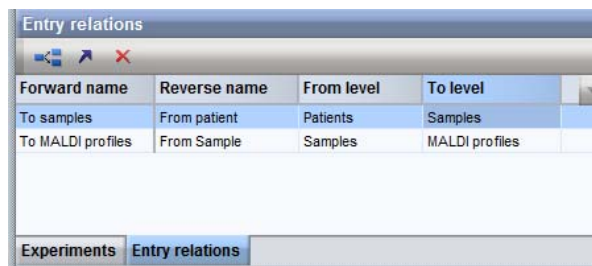


Figure 2-43. The *Entry relations* panel, two relation types created.

2.4.4 Defining relations between entries

Establishing a relation between entries happens by first selecting the entry or entries you want to link to, and then opening the entry you want to link from. In a *One to Many* relation, it is therefore recommended to select the entries that are at the *Many* side of the relation, and then open the entry that is at the *One* side.


As an example, we will link two samples to a patient.

2.4.4.1 In the **Patients** level, open the *Entry edit* window for the first patient by double-clicking on the entry record. See 2.2.3 for explanation on the *Entry edit* window.

In the default configuration of the *Entry edit* window, the *Relations* panel is available as a second tab in the *Database fields* panel (see Figure 1-15). It can also be undocked as a separate panel in the *Entry edit* window (see 1.6.4 for an explanation on the display of panels). The *Relations* panel looks as in Figure 2-45.

2.4.4.2 Leave the *Entry edit* window for the patient open and switch to the **Samples** level.

2.4.4.3 In the **Samples** level, select a few entries (samples) you want to link to the patient. See 2.2.8 on how to select entries in the database.

2.4.4.4 To link the selected samples to the currently opened patient entry, press the  button or select *Edit > Relations > Link currently selected entries* in the *Entry edit* window.

A dialog box pops up, prompting for a relation type to choose. Since there is only one relation type that pertains to the **Patients** level, only *To Samples* is available as choice (see Figure 2-46).

2.4.4.5 Press <OK> to establish the relation.

The *Entry edit* window for the patient now lists two selected **Sample** type entries linked to it. The relation

Key	Patient name	Gender	Birth date	City	ZIP code
HumanTypin...	Rosa Bloum	F	1982-10-03	BigCity	80000
HumanTypin...	Tiss Prutz	M	1974-01-20	SmallTown	67840

Key	SampleID	Sample Source	Sample Type	Sampling Date
HumanTypin...	LVM-3A	Sputum	50574	2007-05-02
HumanTypin...	LVM-4A	Sputum	E17_6	2007-05-02
HumanTypin...	LVK-2	Skin	50574	2007-04-29
HumanTypin...	LVK-3	Skin	E17_6	2007-04-29

Key	ProfileID	RunDate	FileName	MicroplatePos
HumanTypin...	LVM-3A-001	2007-05-03	LVM-3A-001.rec	B2
HumanTypin...	LVM-3A-002	2007-05-03	LVM-3A-002.rec	B3
HumanTypin...	LVM-3A-003	2007-05-03	LVM-3A-003.rec	B4
HumanTypin...	LVM-4A-001	2007-05-04	LVM-4A-001.rec	D8
HumanTypin...	LVM-4A-002	2007-05-04	LVM-4A-002.rec	D9

Figure 2-44. In this example, two samples are obtained from the first patient, and three MALDI profiles are obtained from the second sample.

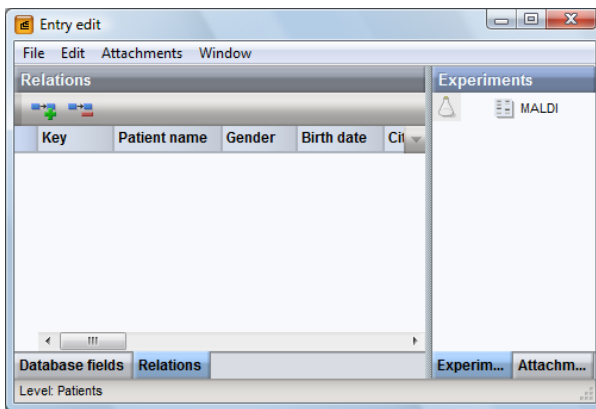


Figure 2-45. The *Relations* panel in the *Entry edit* window.

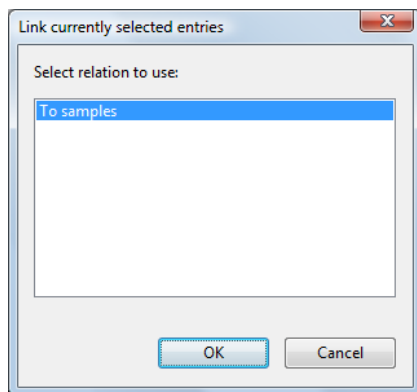


Figure 2-46. Dialog box prompting for relation type to use.

from the patient to the samples is shown as TO SAMPLES, as we have defined it earlier in the relation type (Figure 2-47).

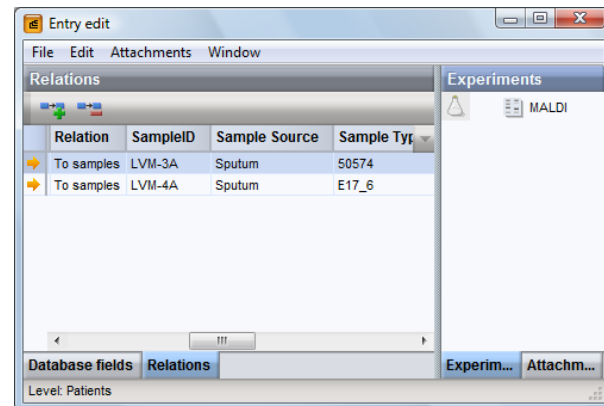



Figure 2-47. Relations panel in the *Entry edit* window, showing two linked entries and description of the relation.

2.4.4.6 Close the *Entry edit* window and unselect any entries by pressing F4 on the keyboard.

Further in this example, we will link some MALDI profiles to a sample.

2.4.4.7 In the **Samples** level, open the *Entry edit* window for one of the sample entries, and click the **Relations** tab.

2.4.4.8 Switch to the **MALDI profiles** level and select a few MALDI entries.

2.4.4.9 To link the selected profiles to the currently opened sample entry, press the  button or select

Edit > Relations > Link currently selected entries in the *Entry edit* window.

A dialog box pops up, prompting for a relation type to choose. Since there are two relation types that pertain to the **Samples** level, the list contains two choices: FROM PATIENT and TO MALDI PROFILES (see Figure 2-48).

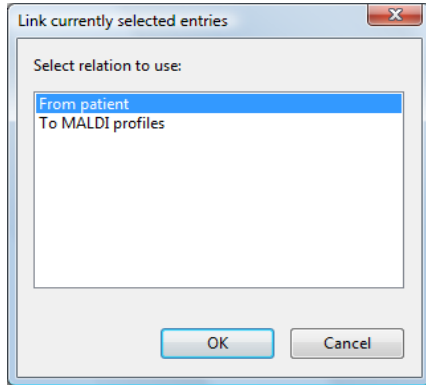


Figure 2-48. Dialog box prompting for relation type to use.

2.4.4.10 Select TO MALDI PROFILES and press <OK> to establish the relation.

The *Entry edit* window for the sample now lists three **MALDI profiles** type entries linked to it. The relation from the patient to the samples is shown as TO SAMPLES, as we have defined it earlier in the relation type (Figure 2-47).

Linked entries from other levels can easily be selected or unselected from the *Entry edit* window.

2.4.5 Relations and scripts

The InfoQuest FP software offers a number of script functions allowing scripts to be written that act on rela-

tions. Besides creating new relations or modifying relations, script functions can also query for relations between entries, and if relations exist, perform specific actions.

One useful example in the above database would be to calculate average MALDI profiles for the samples. The script would look for all MALDI level entries that belong to a sample, average them, and fill in the MALDI experiment for that sample with the average profile.

2.4.6 Different relation types

Although in most applications, relations will be established between entries from different levels, it is also possible to define relations between entries from the same level. For example, in the database we created in this chapter, one could define a relation type within the level of MALDI profiles, to calculate average MALDI profiles from a set of repeats. One MALDI profile would then be related to a number of others (One to Many) with forward and reverse descriptions “Average of” and “To average”, respectively (see Figure 2-50).

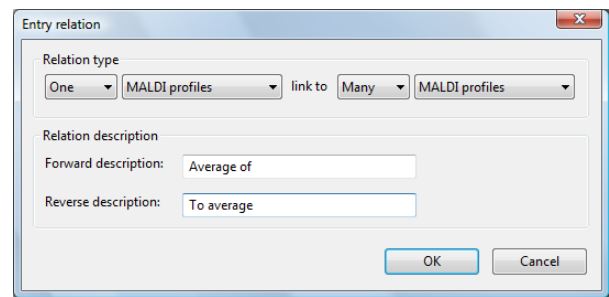


Figure 2-50. Example of a relation type within the same level.

A script, such as described in 2.4.5, would then automatically calculate the average MALDI profiles.

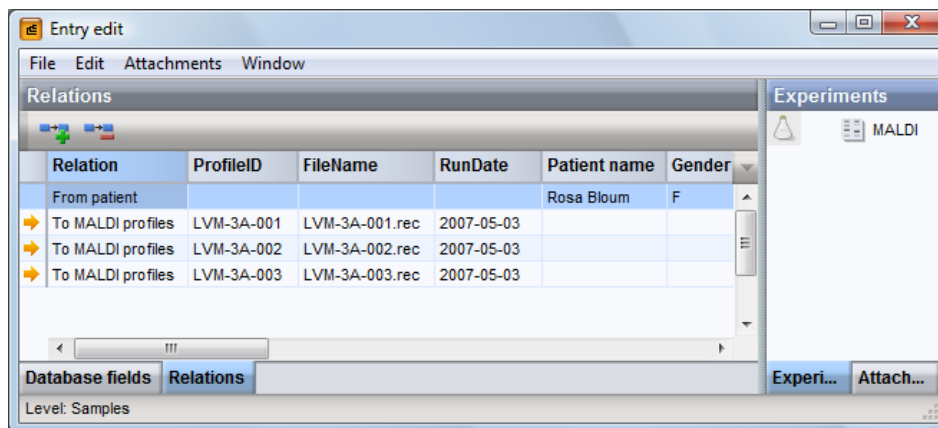


Figure 2-49. *Relations* panel in the *Entry edit* window, showing one linked higher level entry, three linked lower level entries and description of the relation.

2.5 Importing data in a InfoQuest FP database

2.5.1 Importing data using the Import plugin

Various options for importing experiment data are described in , which deals with the different experiment types in InfoQuest FP. Many types of data can also be imported using the **Import** plugin, which is installed automatically with the InfoQuest FP software. To activate the Import plugin, select *File > Install / remove plugins* in the *InfoQuest FP main window* (see also 1.5.3 on how to install plugins).

The Import plugin allows the following data to be imported:

- **Information field data.** InfoQuest FP database information fields can be imported from text files (tab, comma or semicolon separated) or from other databases (e.g. Access, Excel) via an ODBC link.
- **Characters and numerical data.** Character type data can be imported from text files (tab, comma or semicolon separated) or from other databases (e.g. Access, Excel) via an ODBC link. Different formats are supported.
- **Sequence data.** Nucleic acid sequences, e.g. from publicly available sequence repositories, can be imported from text files in different formats (EMBL, GenBank, Fasta). See 3.4.2 for more information on the import of sequence data.
- **Fingerprint data from automated sequencers.** Typing techniques for which the electrophoresis step is performed on an automated sequencer (e.g. AFLP, t-RLFP, etc.) can be imported as densitometric curves from the raw chromatogram files. The different file formats from commercially available sequencers (Applied BioSystems, Beckman and Amersham) are supported. See 3.2.12 for more information.
- **Genemapper peak files.** The Genemapper (Applied BioSystems) text files can be imported as fingerprint type. See 3.2.13 for more information.
- **Trend data.** Trend data can be imported from text files. See 3.5.3 for detailed information on the import of trend data.

NOTES:

(1) Some import routines (e.g. automated import of fingerprint files from AB sequencers) are exclusively for data import in connected databases and cannot be used for local databases.

(2) Database information field data and experiment data that are linked to it can be imported directly from

another InfoQuest FP database using the XML Tools plugin (see 2.6.3).

For detailed instructions on the use of the Import plugin, we refer to the separate Import plugin manual. A pdf version of this manual becomes available when you click on *<Manual>* in the *Plugin installation toolbox* ().

2.5.2 Importing data via an ODBC link

In a local database, InfoQuest FP allows one to establish a link with an external relational database using the *Open Database Connectivity* (ODBC) protocol. This protocol is supported by almost any commercial relational database: Access, Excel, FoxPro, Dbase, Oracle, SQL server, etc... By establishing such a link between InfoQuest FP and an external data source, the user can import data in a completely transparent way into InfoQuest FP. Moreover, the InfoQuest FP local database can be brought up to date using the external data source by performing automatic downloads.

NOTE: The option to configure an external ODBC link is only offered for a local database. However, the use of a connected database is recommended, since there data are stored in a single location (data normalization) and the ODBC link is permanent. This way of working avoids any possible updating conflicts and ensures the data stay up to date. For more information on connected databases, see Section 2.3.

The database records in the external database are mapped into InfoQuest FP entries by making use of the *database key*. The user should specify a unique field of the external database that corresponds to this key, and then the software is able to automatically determine which external record corresponds to which local InfoQuest FP database entry.

2.5.2.1 Setting up the ODBC link

Use the menu item *Database > ODBC link > Configure external database link* in the *InfoQuest FP main window* to call the *ODBC configuration* dialog box (see Figure 2-51) This dialog box contains two information fields, which are to be filled in:

- **The ODBC data source.** This field is to be filled in with a string that defines the external database that will be linked using ODBC. If you are familiar with ODBC, you can specify a string manually. Alternatively, you can press the button *<Select>*. This action pops up the standard Windows dialog box that allows one to select an ODBC data source. In this

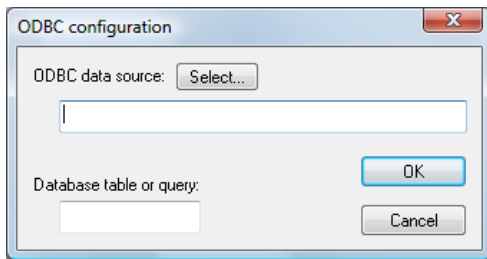


Figure 2-51. The ODBC configuration dialog box.

dialog box, double-click on the name of the appropriate available database software and select the database file on the hard disk.

- **The database table or query.** In this field, you should fill in the name of the table or query in the external database that you want to use to import data from. If you are importing data from a spreadsheet program (e.g. Microsoft Excel), you should first create a “table” in the spreadsheet. This can be done by selecting a range of cells that you want to export and assign a name to this selection (read the documentation of the spreadsheet software on how to export data using an ODBC link).

Pressing <OK> creates the *ODBC database import* dialog box (see Figure 2-52). This dialog box allows the user to specify how each field in the external database should be mapped to a particular field in the InfoQuest FP database. On the left side, the InfoQuest FP fields are listed, while on the right side the external database fields are shown. Initially, all fields are unlinked. You can link two fields by selecting the local InfoQuest FP field from the left column and the external field from the right column, and pressing the <Link> button. At this time, both fields are displayed at the same height, and a green arrow indicates the established link. You can remove any existing link by selecting it and pressing <Unlink>.

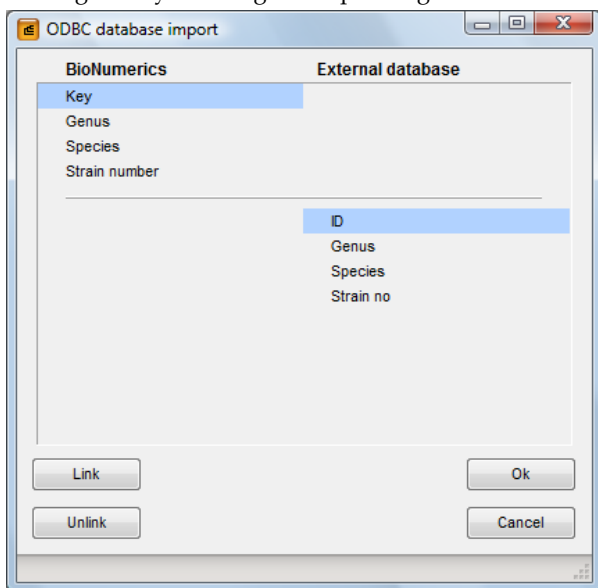


Figure 2-52. The ODBC database import dialog box.

Before you will be able to perform any exchange action, you should make sure that the InfoQuest FP “Key” field, which corresponds to the local database keys, is linked to a field from the external database. This link is obligatory, because the software needs to know which record in the external database corresponds to which entry in InfoQuest FP.

If the necessary links are established between external and local database fields, press <OK> to validate the ODBC link configuration. At this moment, InfoQuest FP is ready to download information from the external data source.

2.5.2.2 Import of database fields using ODBC

• Update all InfoQuest FP database entries from the external data source

It is possible to automatically update all the information fields from each InfoQuest FP entry, using the data provided by the external database. To this end, select *Database > ODBC link > Copy from external database* in the *InfoQuest FP main* window. After confirmation, the software downloads, for each entry, all the database fields that have been linked to the external data source. If the external data source contains records that do not have a corresponding entry in the InfoQuest FP software, the program automatically creates new entries in the InfoQuest FP database (after confirmation by the user). In this way, information fields of existing InfoQuest FP database entries are updated and new entries are automatically added.

• Download a database field from the external data source


It is possible to temporarily download an extra database field from the external data source, into an empty database field of InfoQuest FP. To this end, select the empty database field in the *InfoQuest FP main* window (or create a new one), and use the menu command *Database > ODBC link > Download field from external database*. A dialog box pops up, showing all the fields present in the external database. Select the appropriate field and press <OK> to download the information in the local field. Note that the downloaded information is only held temporarily and not stored on disk. The next time you re-open the same database in InfoQuest FP, the field will be again in its initial state.


• Selection of a list using a query in the external data source

The software allows you to perform a query in the external database, and to visualize the result as a selection list in InfoQuest FP. In the *InfoQuest FP main* window, use the command *Database > ODBC link > Select list from external database*. In the dialog box, you can specify a table that should be used to search in (alternatively, you can specify the name of a pre-defined query that is present in the external database). In the next field, you can write an SQL WHERE clause that should be used to build the selection. A complete description of the possible variants is beyond the scope

of the manual, and can be found in books on the SQL language. Some possibilities are: "GENUS='Ambiorix'" or "GENUS like 'Amb%'". The WHERE clause is applied to the records of the external database, and the resulting selection is visualized as a selection of the corresponding entries in the InfoQuest FP database (assuming that they are present in the local database).

• Getting a detailed report of the external database record

For each entry in the InfoQuest FP database, you can obtain a complete list of all information present in the external data source. To this end, you should first open the *Entry edit* window, e.g. by double-clicking on the name in the entry list. Then use the button  to create a new window that shows a list of all information fields that are present for this entry in the external database. Note that there is no limit to the number of fields that can be viewed and edited in this way, and that each field may consist of several lines and can contain up to 5000 characters.

Moreover, you can change some of these fields, and upload these changes to the external database using the  button.

2.5.2.3 Import of character data using ODBC

One can use an ODBC link to an external database for importing character data into the local InfoQuest FP database. Open the character type that you want to import by double-clicking on its name in the *Experiments*

panel in the *InfoQuest FP main* window. Then use the command *File > Import from external database*. A dialog box pops up, showing a complete list of all the database fields that are present in the external data source. The program determines automatically if any of the characters in the character type corresponds to a database field in the external database. If so, the field is written in boldface, and the character will be filled with the values from this field during the import. You can add new characters to the character type by selecting an unmatched field and pressing *<Create character>*. Groups of characters to add can be selected using the SHIFT key. To import the data, press *<OK>*. For every local entry that has a matching key in the external database, the corresponding characters of this character type will be filled with information from the external database.

2.5.2.4 Import of sequence data using ODBC

The ODBC link to an external database can also be used to import sequence data into the local InfoQuest FP database. Open the sequence type that you want to import by double-clicking on its name in the *Experiments* panel in the *InfoQuest FP main* window. Then use the command *File > Import from external database*. A dialog box pops up, showing a complete list of all the database fields that are present in the external data source. Select the database field that contains the required sequence information and press *<OK>* to import the data. For every local entry that has a matching key in the external database, the corresponding sequence will be filled with the data contained in the selected external database field.

2.6 Database exchange tools

2.6.1 Solutions for data exchange: bundles and XML files

InfoQuest FP offers two simple and powerful solutions to exchange database information between research sites on a peer-to-peer basis: via *bundles* or *XML files*.

A **bundle** contains selected information (e.g. experiment types, information fields) for a selection of database entries and is the original tool for exchanging InfoQuest FP database information. It is a compact data package contained in a single file, which can be sent to other research sites over the internet. The receiver can open the bundle directly in InfoQuest FP and compare the entries contained in it with the own database. However, the information in a bundle is “as is”, and cannot be modified or re-analysed by the receiver.


Exporting InfoQuest FP database information as **XML files** and importing these again in another database is another available exchange tool. Like bundles, selected information can be included for a selection of database entries. When the XML files are imported in a database, the database entries that were contained in the XML files behave just like other database entries.

Which database exchange tool is to be preferred (bundles or XML files), depends on the specific case and will be a trade-off between compactness and flexibility of analysis.

2.6.2 Using bundles in InfoQuest FP

We will illustrate the use of bundles in the **DemoBase** database, by creating a bundle for all entries belonging to the genus *Vercingetorix*.

2.6.2.1 In the *InfoQuest FP* main window with **DemoBase** loaded, select all entries belonging to *Vercingetorix* (see 2.2.8 on how to select database entries).

2.6.2.2 Select **File > Create new bundle** or .

The *Create new bundle* dialog box (Figure 2-53) lists the available database information fields in the left panel and all available experiment types in the right panel.

You can check each of the database information fields and experiment types to be incorporated in the bundle. For fingerprint types, the fingerprint images, band information, and densitometric curves can be incorporated separately.

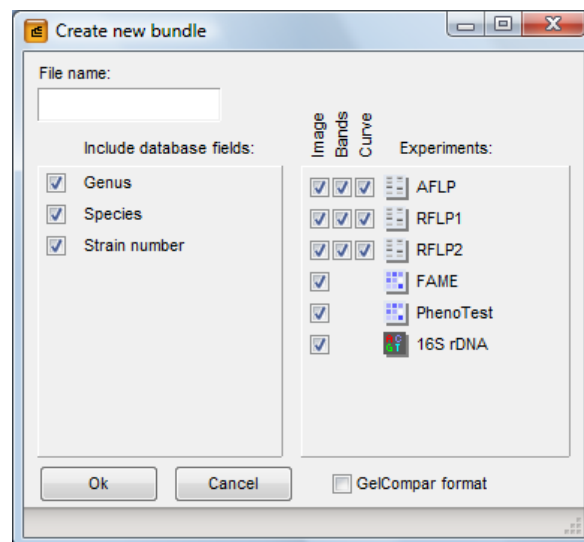


Figure 2-53. The *Create new bundle* dialog box.

2.6.2.3 Leave all checkboxes checked.

NOTE: With the checkbox **GelCompar format**, one can save bundles in the format of *GelCompar* versions 4.1 and 4.2 and *Molecular Analyst Fingerprinting* versions 1.12 through 1.60. Only *Fingerprint information* can be saved in this format. *InfoQuest FP* also recognizes and reads *GelCompar* and *Molecular Analyst Fingerprinting* bundles.

2.6.2.4 Enter a name for the bundle, for example **Vercingetorix**, and press <Ok> to create the bundle.

A bundle file **Vercingetorix.bdl** is created in the **Bundles** directory of **DemoBase** (see Figure 2-3 for the directory structure).

Besides the numerical information of the experiments, a bundle contains all the information of the experiment type, so that *InfoQuest FP* can check whether the experiment types contained in the bundle are compatible with those of the receiver's database. If an experiment type in a bundle is not compatible, this experiment type will automatically be created in the receiver's database. If the bundle contains a database information field which is not defined for the database, this information field will be added to the database.

In case of fingerprint types, the bundle holds the complete information about the reference system used and the molecular weight regression, so that *InfoQuest FP* can automatically remap the bundle fingerprints to be compatible with the database fingerprints.

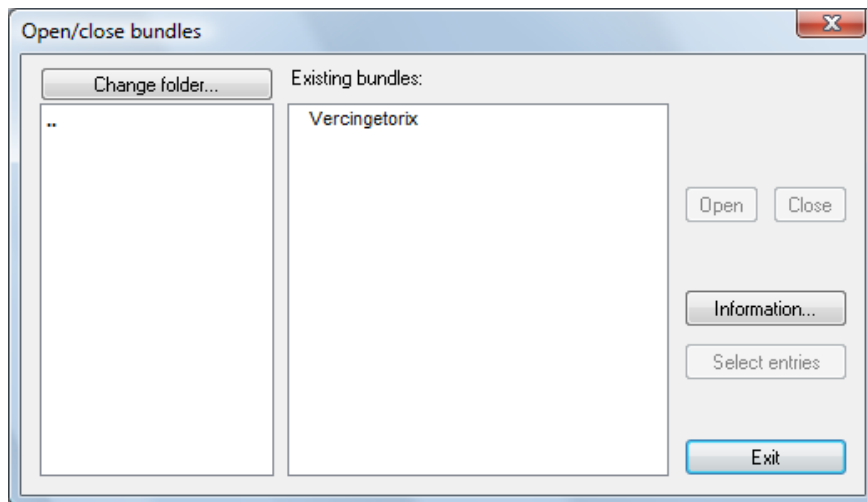



Figure 2-54. The *Open/close bundles* dialog box.

2.6.2.5 As an example for database exchange, copy **Vercingetorix.BDL** to the **Bundles** directory of database **Example**.

2.6.2.6 Close InfoQuest FP and restart the main program under database **Example**.

2.6.2.7 In database **Example**, select **File > Open bundle** or press the  button.

In the *Open/close bundles* dialog box (Figure 2-54), you can browse to the local or network path where the bundle files can be found with the *<Change folder>* button. The default path is the **Bundles** subdirectory of the current database. In the right panel, you can select a bundle in the list of available bundles in the specified path.

2.6.2.8 Select **Vercingetorix** and press the *<Information>* button.

This opens the *Bundle information* dialog box for the selected bundle (Figure 2-55). It shows the available information fields in the bundle, as well as the experiment types contained in it. If an information field or an experiment type is recognized as one of the fields or experiment types in the database, a green dot is shown left from it. If not, a red dot is shown left from it. As soon as the bundle is opened, the missing information fields and experiment types are automatically added to the database.

For example, in the **Example** database, we have created an information field **Strain no**. This clearly corresponds to the information field **Strain number** in the bundle, but since the names are different, InfoQuest FP would add a new information field to the database. To avoid this, you can rename the information fields in the bundle.

2.6.2.9 Select **Strain number** and press the *<Rename>* button under the information fields panel.

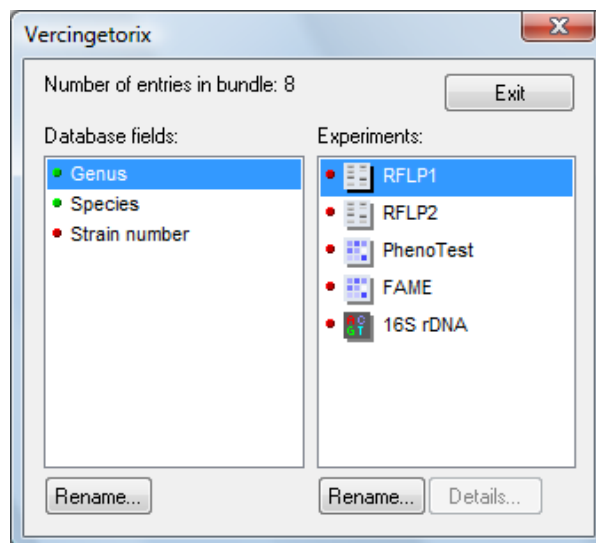


Figure 2-55. The *Bundle information* dialog box.

2.6.2.10 Enter **Strain no** and press **<OK>**.

The information field **Strain no** now has a green dot left from it, indicating that it corresponds to the information field in the database.

A similar problem can happen for the experiment types: another user may have given a different name to the same technique, and this would InfoQuest FP cause to consider the techniques as different experiment types. If you know a technique in a bundle is the same as one of the experiment types defined in the database, you can also rename it using the *<Rename>* button under the experiments panel.


In addition, in character types, the characters may have received different names from other users. For example, institution 1 may have named a character "**Alpha-Glucosidase**", and institution 2 "**a-Glucosidase**". Obviously, InfoQuest FP will consider these different names as different tests. To avoid this, you can select the char-

acter type and press the **<Details>** button. A list of all characters in the experiment type is shown, and those corresponding to characters in the database's experiment type are marked with a green dot; the characters not recognized in the database's experiment type are marked with a red dot. You can rename such characters with the **<Rename>** button.

2.6.2.11 **<Exit>** the *Bundle information* dialog box, and press **<Open>** to load the bundle into the database.

If a bundle is loaded, it is marked with a "+" in the *Open/close bundle* dialog box.

In the database, entries from a bundle are recognized by the name "Bundle" in the *Location* information field. If the *Location* information field is not displayed in the database, it can be shown by clicking on the column

properties button  in the database information fields header and selecting it from the pull-down menu. For all functions, they behave like normal database entries. If you exit InfoQuest FP, they are not automatically loaded when you run the software again. If you know a saved comparison contains bundle entries, you should load the bundles before opening the comparison, in order to avoid an error message.

2.6.2.12 You can select all entries from an opened bundle by pressing the **<Select entries>** button in the *Open/close bundle* dialog box.

2.6.2.13 To close a loaded bundle, select it in the list and press the **<Close>** button.

2.6.2.14 Press **<Exit>** to close the *Open/close bundle* dialog box.

NOTE: If you want a bundle to be always opened with the database when InfoQuest FP is started up, you should rename it to contain the prefix @_ before its name and the .bdl bundle file should be placed in the Bundles folder of the corresponding database.

2.6.3 Export and import using XML files



The tools to export and import database entries as XML files are available as a plugin. To activate the XML Tools plugin, select **File > Install/remove plugins** in the *InfoQuest FP main* window (see also 1.5.3 on how to install plugins). A detailed description on how to use the XML Tools can be found in the XML Tools plugin manual.

2.7 Taking backups from a InfoQuest FP database

In many cases, InfoQuest FP will be used to construct large databases of information that has been collected over a long time span. Obviously, the user should pay attention to protect such databases from accidental data losses, e.g. due to hard disk crashes, power interruptions, etc. and take backups on regular intervals.

The location where the data is stored - and therefore the directories to backup - is different whether a local or a connected database is used. For more details on how information is stored in local and connected databases, see Section 2.1.

2.7.1 Backing up a local database

In a local database setup, all data files that belong to a particular InfoQuest FP database are stored on the hard disk in subdirectories of a single top directory that has the database name (see also Figure 2-3). If InfoQuest FP is opened with this database, this directory is indicated in the status bar on the bottom of the *InfoQuest FP main* window. Alternatively, the corresponding directories of all databases can also be displayed at once in the InfoQuest FP Startup screen, by clicking on the column properties button () in the information fields header and selecting *Path* from the drop-down list. [HOMEDIR] in the path refers to the home directory as specified in the settings. To find out what the current home directory is, or to modify the home directory, press the Settings button () and select *Change home directory*. A dialog box appears which shows the currently selected home directory. For more information about the InfoQuest FP home directory, see 1.1.4.

Since all important information concerning a database is stored inside this top directory, one only needs to back up this complete directory (including subdirectories) to have a complete copy of all data. When the database needs to be restored later on, this top directory can be copied back to the right place on the hard disk.

NOTE: Backups restored from CD or DVD may be read-only. In this case you will have to specify the files to be write-accessible before you run InfoQuest FP with the restored database.

It is possible to create a duplicate of a local database in a similar way. Copy the entire contents of the database's top directory to a new directory. In the InfoQuest FP Startup screen, select *<New>* to create a new database. When the *Database creation* wizard pops up, fill in a name of the duplicate database and click *<Next>*. In the

next tab, click *<Browse>* to change the database top directory into the name of the duplicate directory. In addition, specify *<No>* to the question "Do you want to automatically create the required directories?" In the *Setup new database* dialog box, select *Local database (single user only)* to finish the creation of the database.

2.7.2 Backing up a connected database

In a connected database setup, the actual data may be stored outside the InfoQuest FP data folder (see Figure 2-3). The location of the connected database and associated source files can be found as follows:

2.7.2.1 Open the database in InfoQuest FP and select *Database > Connected database* in the *InfoQuest FP main* window.

2.7.2.2 In the *Connected databases* dialog box, select the currently defined connected database and click *<Edit>*.

The ODBC connection string in the *Connected database configuration* dialog box (top left panel, see Figure 2-37) contains the database name and location. The source files is shown in the bottom right of the same dialog box.

To ensure completeness, both the connected database and the source files should be backed up.

In case of an **automatically created connected database** (default setting when creating a new database), the connected database (.mdb) and the source files folder are located in the top directory that has the database name. This is indicated in the *Connected database configuration* dialog box as [DBPATH]. [DBPATH] refers to the database folder in the InfoQuest FP home directory as

specified under Settings () in the Startup screen.

Therefore, backing up this complete directory (including subdirectories) is sufficient to have a complete copy of all data. When the database needs to be restored later on, this top directory can be copied back to the right place on the hard disk.

In case of **custom created databases or when InfoQuest FP was connected to an already existing database**, the user needs to check the connected database and source file location in the *Connected database configuration* dialog box and back them up separately.

Professional DBMS such as SQL Server, Oracle, MySQL, etc. can be configured to take automatic backups on regular time intervals. We refer to the DBMS documentation for the setup of such automatic backups.

3. EXPERIMENTS

3.1 Experiment types available in InfoQuest FP

In InfoQuest FP, experiments are divided in six classes: *fingerprint types*, *character types*, *sequence types*, *2D gel types*, *trend data types*, and *matrix types*. Additionally, a “container” experiment type called *composite data set* is available.

The **fingerprint types** include any densitometric record seen as a profile of peaks or bands. Examples are electrophoresis patterns, gas chromatography or HPLC profiles, spectrophotometric curves, etc. For example, within the fingerprint types, you can create a Pulsed Field Gel Electrophoresis (PFGE) experiment type with specific settings such as reference marker, MW regression, stain, band matching tolerance, similarity coefficient, clustering method, etc. Fingerprint types can be derived from TIFF or bitmap files as well, which are two-dimensional bitmaps. The condition is that one must be able to translate the patterns into densitometric curves.

With the **character types**, it is possible to define any array of named characters, binary or continuous, with fixed or undefined length. The main difference between character types and electrophoresis types is that in the character types, each character has a well-determined name, whereas in the electrophoresis types, the bands, peaks or densitometric values are unnamed (a molecular size is NOT a well-determined name!). Examples of character types are antibiotics resistance profiles, fatty acid profiles (if the fatty acids are known), metabolic assimilation or enzyme activity test panels such as API, Biolog, and Vitek, etc. Single characters such as Gram stain, length, etc. also fall within this category.

Within the **sequence types**, the user can enter sequences of nucleic acids (DNA and RNA) and amino acids. InfoQuest FP recognizes widely used sequence file formats such as EMBL, GenBank, and Fasta with import of user-selected header tags as information fields, and optional storage of headers. Other sequence formats can be imported easily.

The **2D gel types** include any two-dimensional bitmap image seen as a profile spots or defined labelled structures. Examples are e.g. 2D protein gel electrophoresis patterns, 2D DNA electrophoresis profiles, 2D thin layer chromatograms, or even images from radioactively labelled cryosections or short half-life radiotracers.

The **trend data types** are measurements that register a trend of a condition in function of a parameter. Examples are the kinetic analysis of metabolic and enzymatic activity, real-time PCR, or time-course experiments using microarrays. Although multiple readings per experiment are mostly done in function of time, they can also depend on an other factor, for example readings in function of different concentrations.

With the **matrix types**, it is possible to import external similarity matrices, providing similarity between entries revealed directly by the technique, or by other software. These matrices can be linked to the database entries in InfoQuest FP and they are used together with other information to obtain classifications and identifications. An example of a matrix type is a matrix of DNA homology values. DNA homology between organisms can only be expressed as pairwise similarity, not as character data.

Composite data sets do not necessarily correspond to an actual experiment, but are character tables derived from one or more of physical experiments. They provide a convenient way to analyse the combined results of several character type experiments and offer additional analysis tools for other experiment types.


The user can create more than one experiment of the same type. For example, one can create two different fingerprint type experiments, to analyze PFGE gels obtained with two different restriction enzymes. The setup of the different experiment types is described in the chapters that follow. The 2D gel types are described separately in Chapter 6.

3.2 Setting up fingerprint type experiments

3.2.1 Defining a new fingerprint type

The steps involved in data processing of fingerprint types will be illustrated with an example TIFF file from the **Sample and Tutorial data\Sample gel image file** folder on the CD-ROM. This directory contains a gel **Gel_01.tif**. The same gel file is also available from the download page of the website (www.bio-rad.com/softwaredownloads).

3.2.1.1 Create a new database (see 1.5.2).

3.2.1.2 In the *InfoQuest FP* main window, select **Experiments > Create new fingerprint type** from the main menu, or press the  button from the *Experiments* panel toolbar and select **New fingerprint type**.

3.2.1.3 The *New fingerprint type* wizard prompts you to enter a name for the new type. Enter a name, for example **RFLP**.


3.2.1.4 Press **<Next>** and check the type of the fingerprint data files. The default settings correspond to the most common case, i.e. two-dimensional TIFF files with 8-bit OD depth (256 gray values).


3.2.1.5 After pressing **<Next>** again, the wizard asks whether the fingerprints have inverted densitometric values. This is the case when you are using ethidium bromide stained gels, photographed under UV light (such as the example provided). The bands then appear as fluorescent lighting on a black background. Since *InfoQuest FP* recognizes the darkness as the intensity of a band, you should answer **Yes**, to allow the program to automatically invert the values when converting the images to densitometric curves. Furthermore, the wizard allows you to adjust the color of the background and the bands to match the reality. The red, green and blue components can be adjusted individually for both the background color and the band color. Usually, you can leave the colors unaltered.

3.2.1.6 In the next step, you are prompted to allow a **Background subtraction**, and to enter the size of the disk, as a percentage of the track length. The default disk size of 10% will suit for most fingerprint types. For high resolution fingerprints (e.g. AFLP and sequencer-generated patterns) you can try a smaller disk size. Later, we will see how we can have the program propose the optimal background subtraction settings automatically. At this time, we leave the background subtraction disabled.

3.2.1.7 Press **<Finish>** to complete the creation of the new fingerprint type.

NOTE: You will be able to adjust any of these parameters later on.

The *Experiments* panel () now lists **RFLP** as a fingerprint type.

3.2.1.8 Click the  button in the *Experiment files* panel or select **File > Add new experiment file** in the *InfoQuest FP* main window.

3.2.1.9 Select the file **Gel_01.tif** from the **Sample and Tutorial data\Sample gel image file** folder on the CD-ROM or from the downloaded and unzipped folder.

3.2.1.10 The software now asks "Do you want to edit the image before adding it to the database". Answer **<Yes>** to open the *Image import* editor.


The selected file is opened in the *Fingerprint image import* editor, an editor which allows the user to perform a number of preprocessing functions on the image (Figure 3-1). These functions include flipping, rotating and mirroring the image, inverting the image color, converting color images to grayscale, and cropping the image to defined areas.


NOTES:

(1) It is possible to skip the *Fingerprint image import* editor and copy the file directly to the database, by answering **<No>** to the question in 3.2.1.10. In case you skip this step, make sure the file is an uncompressed grayscale TIFF file, which is the only format recognized by the *InfoQuest FP* database. Continue with paragraph 3.2.2.

(2) The *Fingerprint import image editor* supports most known file types such as JPEG, GIF, PNG and compressed TIFF files in gray scale or RGB color. For the conversion to an uncompressed grayscale TIFF file see 3.2.1.12 (**Image > Convert to gray scale**).

The *Fingerprint image import* window consists of three tabs: **Original**, **Processed**, and **Cropped**.

3.2.1.11 In the **Original** tab, the unprocessed image is shown. In the **Original** view, you can zoom in ()

or zoom out () , and save the image to the database

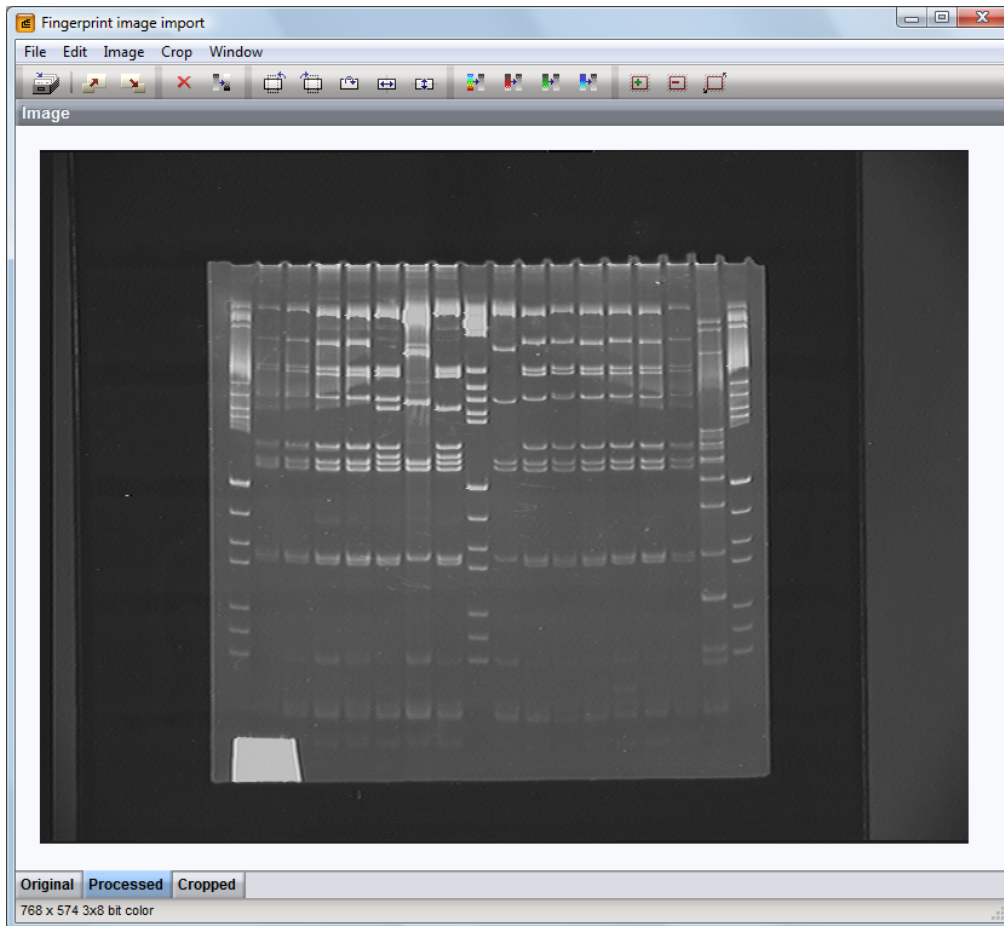









Figure 3-1. The *Fingerprint image import* window.





( or *File > Add image to database*). The image can only be saved when it is in gray scale mode (see below).

3.2.1.12 In the **Processed** tab, the same options are available as in the **Original** tab, plus a number of image editing tools. These include:


- Inverting the color ( or *Image > Invert*) to invert images that have a black background, for example gels that were stained with ethidium bromide.


- Rotating the image 90° left ( or *Image > Rotate > 90° left*), 90° right ( or *Image > Rotate > 90° right*), or 180° ( or *Image > Rotate > 180°*).

- Mirroring the image horizontally ( or *Image > Mirror > Horizontal*) or vertically (, *Image > Mirror > Vertical*).


- Average RGB colors to gray scale ( or *Image > Convert to gray scale > Averaged*), or convert a single channel to gray scale, either red ( or *Image > Convert to gray scale > Red channel*), green ( or *Image > Convert to gray scale > Green channel*) or blue ( or *Image > Convert to gray scale > Blue channel*).

3.2.1.13 The editor also allows you to crop the image to a selected area, to which the following functions are available:


- *Crop > Add new crop* or , to add a new crop mask to the image. The crop mask can be moved by clicking anywhere inside the rectangle and dragging it to another position, or resized by clicking and dragging the bottom right corner of the rectangle.


- *Crop > Rotate selected crop* or , to rotate the crop mask over a defined angle. Rotating the crop mask over an angle different from 90° or 180° will cause the program to recalculate densitometric values based upon interpolation, which means that the quality of the image


may slightly decrease. This action is therefore not recommended unless it is inevitable.

- **Crop > Delete selected crop** or  is to delete the crop mask that is currently selected. Note in this respect that the program allows multiple crop masks to be defined for a single image. The final image that will be saved to the database, will be composed of all cropped areas aligned horizontally next to each other.

3.2.1.14 With **Image > Expand intensity range**, it is possible to recalculate the pixel values of the image so that they cover the entire range within the OD depth of the file, e.g. 8-bit = 256 gray levels, 16-bit = 65,536 gray levels.

3.2.1.15 The image can be reset to its original state with **Image > Load from original** or by pressing .

3.2.1.16 To edit this gel, convert it to gray scale by averaging the 3 channels () and define a crop mask within the gel borders, excluding the black area at the left bottom, but including the full patterns.

3.2.1.17 The third tab, *Cropped*, displays the result of the image as defined by the crop mask(s). When you are satisfied with the result of the preprocessing, you can save the image to the database using the  button.

3.2.1.18 Give the gel a name and exit the *Fingerprint image import* window with **File > Exit**.

One gel becomes available in the *Experiment files* panel, Gel_01. The file is marked with N, which means that it has not been edited yet (see Figure 3-2).

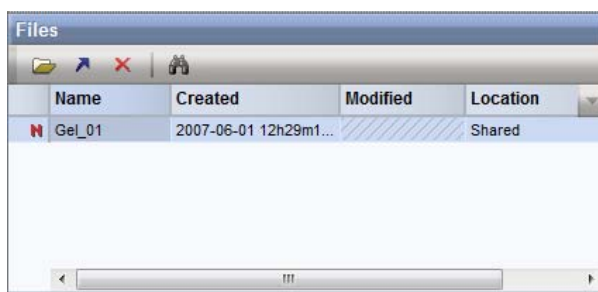




Figure 3-2. The Experiment files panel after import of a gel.

Any other gel TIFF file you want to process can be imported in the same way in the current database. The program will list these TIFF files in the *Experiment files* panel.

*NOTE: Experiment files added to the Files panel can be deleted by selecting the file and choosing **File > Delete experiment file** from the main menu or clicking on*

*the  icon in the Files panel toolbar. Deleted experiment files are struck through (red line) but are not actually deleted until you exit the program. So long, you can undo the deletion of the file by selecting **File >***

***Delete experiment file** or clicking on the  icon again.*

3.2.2 Processing gels

An experiment file is edited in two steps: in a first step, the data are entered or edited, and in a second step, the data is assigned to the database entries.



3.2.2.1 Click on Gel_01 in the *Files* panel (see Figure 3-2), and then select **File > Open experiment file (data)** in the main menu.


Since the gel is new (unprocessed), InfoQuest FP does not know what fingerprint type it belongs to. Therefore, a list box is first shown, listing all available fingerprint types, and allowing you to select one of them, or to create a new fingerprint type with **<Create new>**. In this case, there is only one fingerprint type available, **RFLP**.

3.2.2.2 Select **RFLP** and press **<OK>**.

The gel file is being loaded, which may take some time, depending on the size of the image. The *Fingerprint data editor* window appears (Figure 3-3), showing the image of the gel.

*NOTE: In a local database, the gel can be mirrored with **File > Tools > Vertical mirror of TIFF image** or **File > Tools > Horizontal mirror of TIFF image**. These commands are equivalent to the commands available in the Image import editor (see 3.2.1.12).*

The whole process of lane finding, normalization, band finding and band quantification is contained in a wizard, allowing the user to move back and forth through the process and make changes easily in whichever step of the process. The  and  buttons in the toolbar are to move back and forth, respectively. The process involves the following steps, shown in the tabs in the bottom left corner of the window: **1. Strips** (defining lanes), **2. Curves** (defining densitometric curves), **3. Normalization**, and **4. Bands** (defining bands and quantification). The tabs themselves can be used for navigation between the different steps and allow you to 'skip' steps, e.g. to return in one click from **Normalization** to **Strips** when it turns out a lane was not properly defined. When processing a new gel image, however, it is not recommended to skip any steps in the process.

Within each of these four steps, there is an undo/redo function. To undo one or more actions, you can use the undo button , or **Edit > Undo** (CTRL+ Z) from the

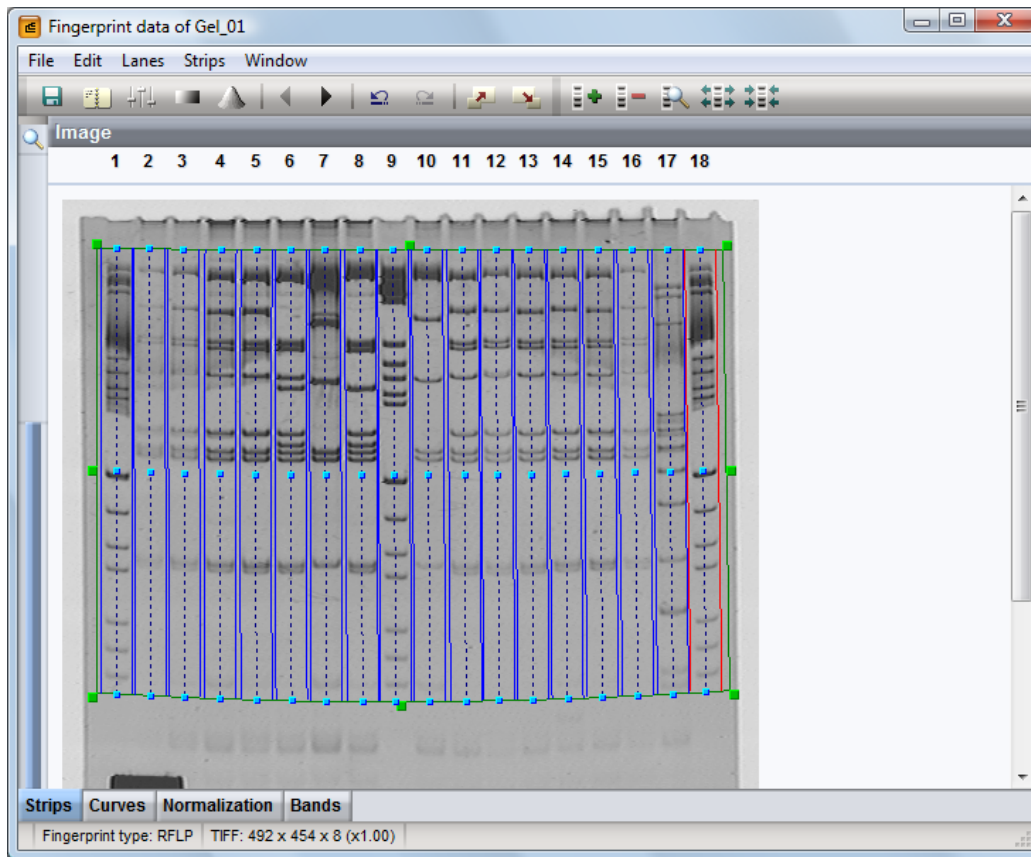





Figure 3-3. The *Fingerprint data editor* window. Step 1: defining pattern strips.


menu. To redo one or more actions, use the redo button , or *Edit > Redo* (CTRL+Y) from the menu. Once you have moved from one step to another, the undo/redo function within that step is lost.

3.2.3 Defining pattern strips on the gel

3.2.3.1 At the start, the image is shown in original size (x 1.00, see status bar of the window). You can zoom in and zoom out with *Edit > Zoom in* and *Edit > Zoom out*, or using the  and  buttons, respectively. Shortcuts are CTRL+PageUp and CTRL+PageDown on the keyboard. The zoom slider (left of the *Image* panel in default configuration) offers a convenient alternative for zooming in and out on the gel image. See 1.6.7 for a detailed description of the zoom slider functions.

3.2.3.2 When a large image is loaded, a *Navigator* window can be popped up to focus on a region of the image. To call the navigator, double-click on the image, press the space bar or right-click and select *Navigator* from the floating menu.

3.2.3.3 You can change the brightness and contrast of the image with *Edit > Change brightness & contrast* or with

. This pops up the *Image brightness & contrast* dialog box (Figure 3-4).

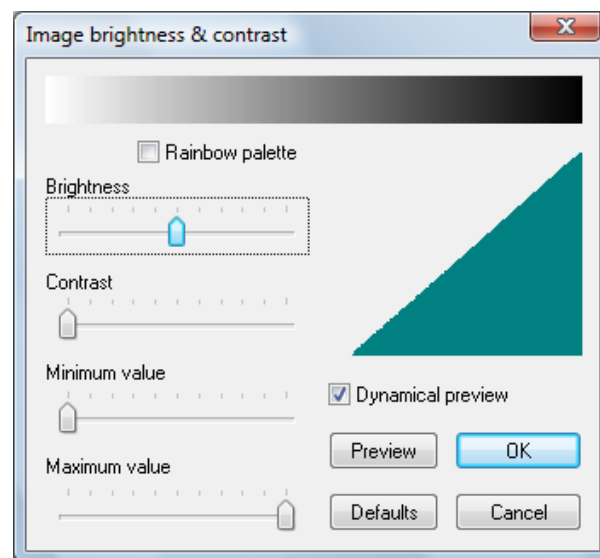


Figure 3-4. *Image brightness & contrast* dialog box.

3.2.3.4 In the *Image brightness & contrast* dialog box, click **Dynamical preview** to have the image directly updated with changes you make.


3.2.3.5 Use the *Minimum value* slide bar to reduce background if the background of the whole image is too high.

3.2.3.6 Use the *Maximum value* slide bar to darken the image if the darkest bands are too weak.

The option *Rainbow palette* can be used to reveal even more visual information in areas of poor contrast (weak and oversaturated areas) by using a palette that exists of multiple color transitions.

3.2.3.7 If you press <OK>, the changes made to the image appearance are saved along with the fingerprint type.

NOTE: The brightness and contrast settings are saved along with the fingerprint type, but are not specific for a particular gel. The Gel tone curve editor, as explained further, is a more powerful image enhancement tool for which the settings are saved for each particular gel.

3.2.3.8 With **File > Show 3D view** or , a three dimensional view of the gel image can be obtained in a separate 3D view window (Figure 3-5).

3.2.3.9 In the 3D view window, you can use the **Left**, **Right**, **Up** and **Down** arrows keys on the keyboard, to


turn the position of the image in all directions. The image can also be rotated horizontally and vertically by dragging the image left/right or up/down using the mouse.

3.2.3.10 You can change the zoom factor using **View > Zoom in** (PgDn) or **View > Zoom out** (PgUp).

3.2.3.11 You can also change the vertical zoom (Z-axis showing the peak height) with **View > Higher peaks** (INS) or **View > Lower peaks** (DEL).

NOTE: In the three further steps of the Fingerprint data editor window (2. Curves, 3. Normalization, and 4. Bands), the 3D view window can also be popped up, showing only the selected lane image rather than the entire gel image.

3.2.3.12 Close the 3D view window with **File > Exit**.

3.2.3.13 To save the work done at any stage of the process, you can select **File > Save**, press CTRL + S, the F2 key, or the  button. In case you work with complex gels, it is advisable to save the work at regular times.

When you save the gel file with **File > Save**, the program may prompt you with the following question: "*The resolution of this gel differs considerably from the normalized track resolution. Do you wish to update the normalized track resolution?*". The gel resolution is

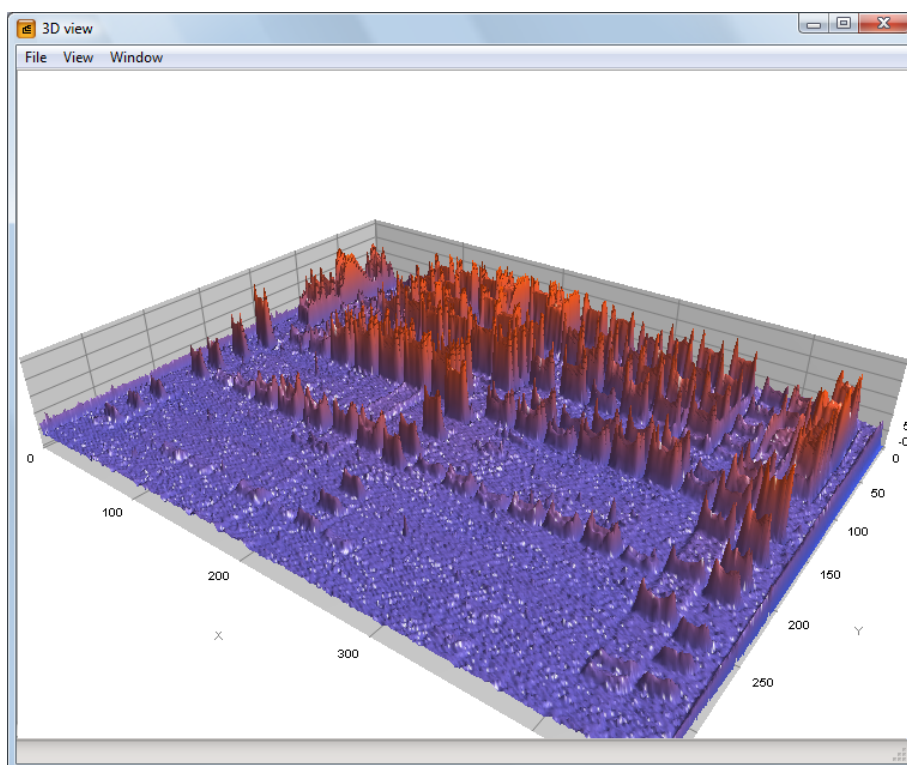


Figure 3-5. The 3D view window.

explained further (see 3.2.9.2). If the question appears (not the case for the example gel), answer <Yes>.

The green rectangle is the *bounding box*, which delimits the region of interest of the gel: tracks and gelstrips will be extracted within the bounding box.

3.2.3.14 To move the bounding box as a whole, hold down the CTRL key while dragging it in any of the green squares (*distortion nodes*).

3.2.3.15 Adjust the box by dragging the distortion nodes as necessary: corner nodes can be used to resize the box in two directions, whereas inside nodes can only be used to resize one side of the box.

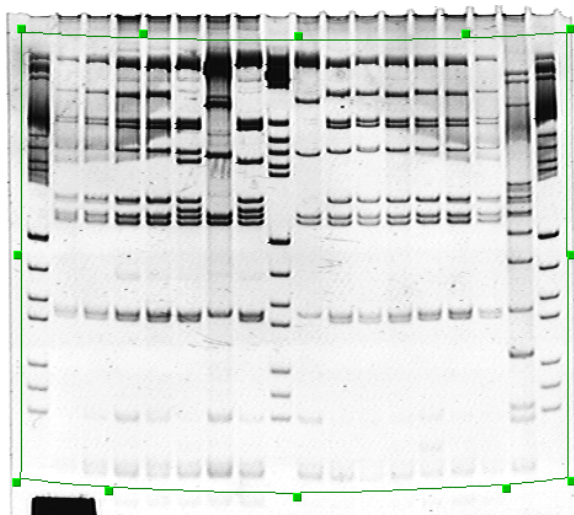


Figure 3-6. Defining the bounding box to follow contours of distorted gel.

3.2.3.16 By using the SHIFT key, one can even distort the sides of the rectangle. Holding the SHIFT key while dragging the corner nodes will change the rectangle into a non-rectangular quadrangle (parallelepiped).

3.2.3.17 A curvature can be assigned to the sides of the bounding box by holding the SHIFT key while dragging one of the inside nodes in any direction (see Figure 3-6, top and bottom sides).

3.2.3.18 On the top and bottom sides of the bounding box, more nodes can be added using *Lanes > Add bounding box node*. While holding down the SHIFT key, a node can be dragged to the left or to the right using the mouse.


3.2.3.19 A node can be deleted from the bounding box using *Lanes > Delete bounding box node*.

NOTES:

(1) Following the curvature of a distorted gel is not crucial, as this is normally corrected in the normalization step (see further, 3.2.5) in case there are sufficient reference lanes on the gel. However, as it will provide a first rough normalization, it can aid the


automatic or manual assignment of bands as explained in 3.2.5. Also, the software allows the bounding box curvature to be used for rectifying sloping or “smiling” lanes (e.g. Figure 3-6, outer lanes), if this option is enabled (see the Fingerprint conversion settings dialog box, Figure 3-7, and explanation below).

(2) If you are running an upgrade from an older InfoQuest FP version (prior to 4.0) and using a connected database, the column BOUNDINGBOX in the connected database may not be long enough to hold an increased number of nodes. To resolve this, perform <Auto construct tables> in the Connected database setup window (see Figure 2-23).

3.2.3.20 Select *Lanes > Auto search lanes* or  to let the program find the patterns automatically. A dialog box asks you to enter the approximate number of tracks on the gel.

3.2.3.21 Enter 18 as the number of tracks in Gel_01 press <OK>.

Each lane found on the image is represented by a *strip*: a small image that is extracted from the complete file to represent a particular pattern. The borders of these strips are represented as blue lines, or red for the selected lane (see Figure 3-8). By default, the strip thickness is 31 points, which is too wide in this example.

3.2.3.22 Call the *Fingerprint conversion settings* dialog box with *Edit > Edit settings* or . This dialog box consists of four tabs, of which the tab corresponding to the current stage of the processing is automatically selected. Since we are now in the first step (defining strips), the *Raw data* tab is selected (see Figure 3-7).

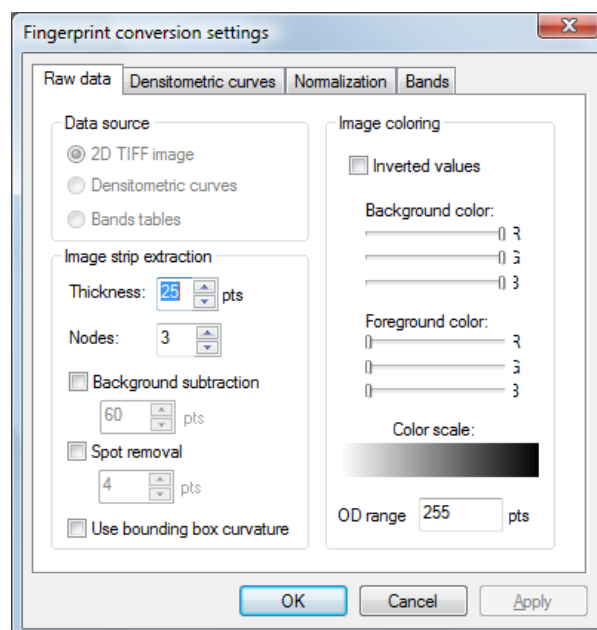


Figure 3-7. The *Fingerprint conversion settings* dialog box, *Raw data* tab.

3.2.3.23 Adjust the *Thickness* of the image strips so that the blue lines enclose the complete patterns (blue lines of neighboring patterns should nearly touch each other). See Figure 3-8 for an optimally adjusted example.

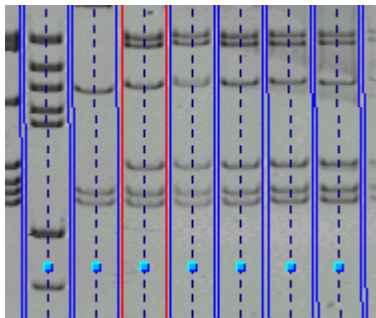


Figure 3-8. Optimal strip thickness settings, detail.

3.2.3.24 If necessary, increase the number of distortion nodes. These nodes allow you to bend the strips locally. Usually, three nodes should be fine.

Two more options, *Background subtraction* and *Spot removal* allow gel scans with irregular background and spots or artifacts to be cleaned up to a certain extent. It should be emphasized that the options *Background subtraction* and *Spot removal* have an influence on *gelstrips* in all further processes of the program: *gelstrips* will always be shown with background subtracted and with spots removed. In addition, when two-dimensional quantification is done, the *gelstrips* with background subtracted and spots removed are used. Hence, we recommend **NOT** to use these options unless (1) the image has a strong irregular background, for example by non-homogeneous illumination of the gel, so that the *gelstrips* would not look appropriate for presentation or publication; (2) the gel contains numerous spots that would influence the densitometric curves extracted from the *gelstrips* (spots on the image are seen as peaks on a densitometric curve, and hence have a strong impact on correlation coefficients, band searching etc.).

The *Background subtraction* is based on the “rolling ball” principle, and the size of the ball in pixels of the image can be entered. The larger the size of the ball, the less background will be subtracted.

The *Spot removal* is a similar mechanism as the rolling ball, but an ellipse is used instead, in order to separate bands from spots. The size of the ellipse can be entered in pixels. Unlike the background subtraction, the size of the ellipse should be kept as small as possible in order not to erase bands.

NOTES:

(1) The spot removal mechanism inevitably causes some distortion on the patterns. The smaller the size of the spot removal, the less the distortion.


(2) If background subtraction on the *gelstrips* is applied, it is not necessary anymore to perform background subtraction on the densitometric curves, since this is doing exactly the same but on one-dimensional patterns.


The effect of background subtraction and spot removal on *gelstrips* is only seen in the next step, when the *gelstrips* are shown. Since the example gels do not require these features, we will not further discuss them.

Using the option *Use bounding box curvature*, it is possible to have the program correct smiling or sloping bands due to distortion in the gel. The bands will be rectified according to the bounding box curvatures defined (3.2.3.17). An example is given in Figure 3-6, where the bounding box has been assigned a curvature to follow the distortions in the outer lanes. The result of enabling the correction for bounding box curvature is shown in Figure 3-10, where it can be clearly seen that the bands of the outer lanes have been straightened.



3.2.3.25 Click <OK> to validate the changes.

3.2.3.26 Adjust the position of each spline as necessary by grabbing the nodes using the mouse. Use the SHIFT key to bend a spline locally in one node.

3.2.3.27 Add lanes with *Lanes > Add new lane* or the ENTER key or . A new lane is placed right from the selected one.

3.2.3.28 Remove a selected lane with *Lanes > Delete selected lane* or DEL or  if necessary.


3.2.3.29 If one lane is more distorted than the number of nodes can follow, you can increase the number of nodes in that lane by selecting it and *Strips > Increase number of nodes*.

3.2.3.30 If the lanes are not equally thick, you can increase or decrease the thickness of each individual strip with *Strips > Make larger* and *Strips > Make smaller* (F7 and F8, or  and , respectively.

Once the lanes are defined on the gel, a powerful tool to edit the appearance of the image is the *Gel tone curve* editor. While the *Image brightness and contrast* settings act at the screen (monitor) level, i.e. after the TIFF grayscale information is converted into 8-bit grayscale, the *Gel tone curve* editor acts at the original TIFF information level. This means that, in case a gel image is scanned as 16-bit TIFF file, the tone curve settings are applied to the full 16-bit (65536) grayscale information which allows much

more information to be magnified in particular areas of darkness. The advantages are:

- Weak bands are much better enhanced resulting in a smoother and more reliable picture.
- The tone curve acts at a level below the brightness and contrast settings and can be saved along with a particular gel. In all further imaging tools of the program, the tone curve for the particular gel is applied. Brightness and contrast settings are not specific to a particular gel.
- The user can fine-tune the tone curve to obtain optimal results. This will be explained below.

3.2.3.31 Select the *Image brightness and contrast* box with *Edit > Change brightness & contrast* or with , and press *<Defaults>* to restore the defaults.

3.2.3.32 In the *Fingerprint data editor* window menu, select *Edit > Edit tone curve*. The *Gel tone curve* editor appears as in Figure 3-9.

The upper panel is a distribution plot of the densitometric values in the TIFF file over the available range. The right two windows are a part of the image *Before correction* and *After correction*, respectively.

3.2.3.33 You can scroll through the preview images by left-clicking and moving the mouse while keeping the mouse button pressed.

3.2.3.34 Select a part of the preview images which contains both very weak and dark bands.

Left, there are two buttons, *<Linear>* and *<Logarithmic>*. Both functions introduce a number of distortion points on the tone curve, and reposition the tone curve so that it begins at the grayscale level where the first densitometric values are found, and ends at its maximum where the darkest densitometric values are found. This is a simple optimization function that rescales the used grayscale interval optimally within the available display range. The difference between linear and logarithmic is whether a linear or a logarithmic curve is used.

3.2.3.35 In case of 8-bit gels, a linear curve is the best starting point, so press *<Linear>*. The interval is now optimized between minimum and maximum available values, and the preview *After correction* looks a little bit brighter.

There are six other buttons that are more or less self-explaining: *<Decrease zero level>* and *<Increase zero level>* are to decrease and increase the starting point of the curve, respectively.

<Enhance weak bands> and *<Enhance dark bands>* are also complementary to each other, the first making the curve more logarithmic so that more contrast is revealed in the left part of the curve (bright area), and the second making the curve more exponential so that more contrast is revealed in the right part of the curve (dark area).

<Reduce contrast> and *<Increase contrast>* make the curve more sigmoid so that the total contrast of the image is reduced or enhanced, respectively.

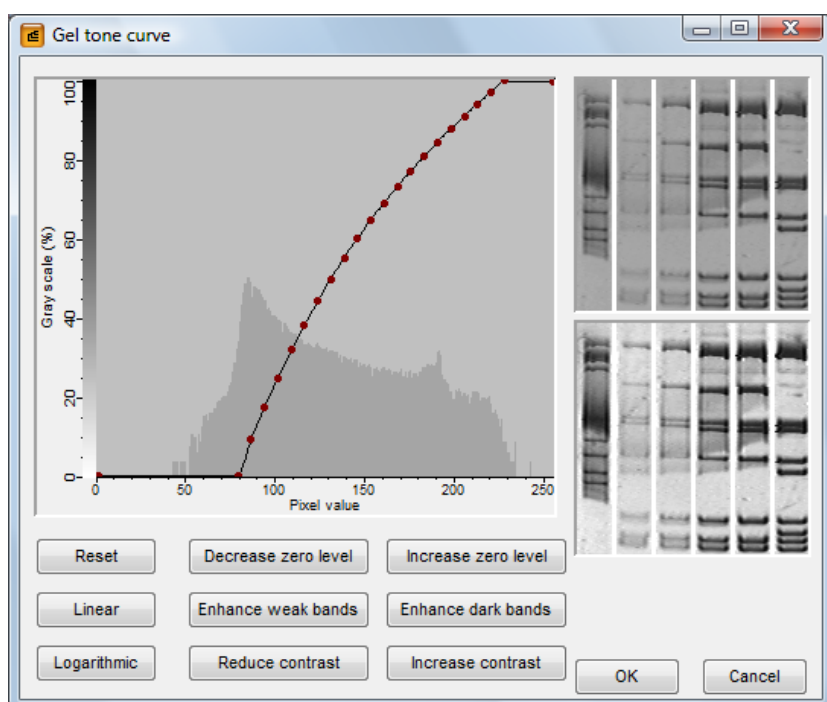



Figure 3-9. The *Gel tone curve* editor.

3.2.3.36 For the image loaded, pressing three times **<Enhance weak bands>** and subsequently 10 times **<Increase zero level>** provides a clear, sharp and contrastive picture.

3.2.3.37 Press **<OK>** to save these tone curve settings.

*NOTE: It is also possible to edit the tone curve manually; nodes can be added by double-clicking on the curve in the Tone curve window, or can be deleted by selecting them and pressing the DEL key. The curve can be edited in each node by left-clicking on the node and moving it. There is a **<Reset>** button to restore the original linear zero-to-100% curve.*

3.2.3.38 Press  to go to the next step: defining densitometric curves.

3.2.4 Defining densitometric curves


In this step, the window is divided in two panels (Figure 3-10): the left panel shows the strips extracted from the image file and the right panel shows the densitometric curve of the selected pattern, extracted from the image file.

3.2.4.1 You can move the separator between both panels to the left or to the right to allow more space for the strips or for the curves.

The program has automatically defined the densitometric curves using the information of the lane strips you entered in the previous step. Normally, you will not have to change the positions of the densitometric curves anymore, except when you want to avoid a distorted region within a pattern, e.g. due to an air bubble within the gel.

3.2.4.2 If necessary, adjust the position of a spline by grabbing the nodes using the mouse. Use the SHIFT key to bend the spline locally in one node.

The blue lines represent the width of the area within which the curve will be averaged. The default value is 7 points. In most cases, you will have to optimize this value for a given type of gel images.

3.2.4.3 Call the *Fingerprint conversion settings* dialog box with **Edit > Edit settings** or . This time, the *Densitometric curves* tab is displayed (Figure 3-11).

3.2.4.4 Change the *Averaging thickness* for curve extraction. For the example, enter 11. Ideally, the thickness should be chosen as broad as possible. However, smiling and distortion at the edges of the bands should be excluded (see Figure 3-12).

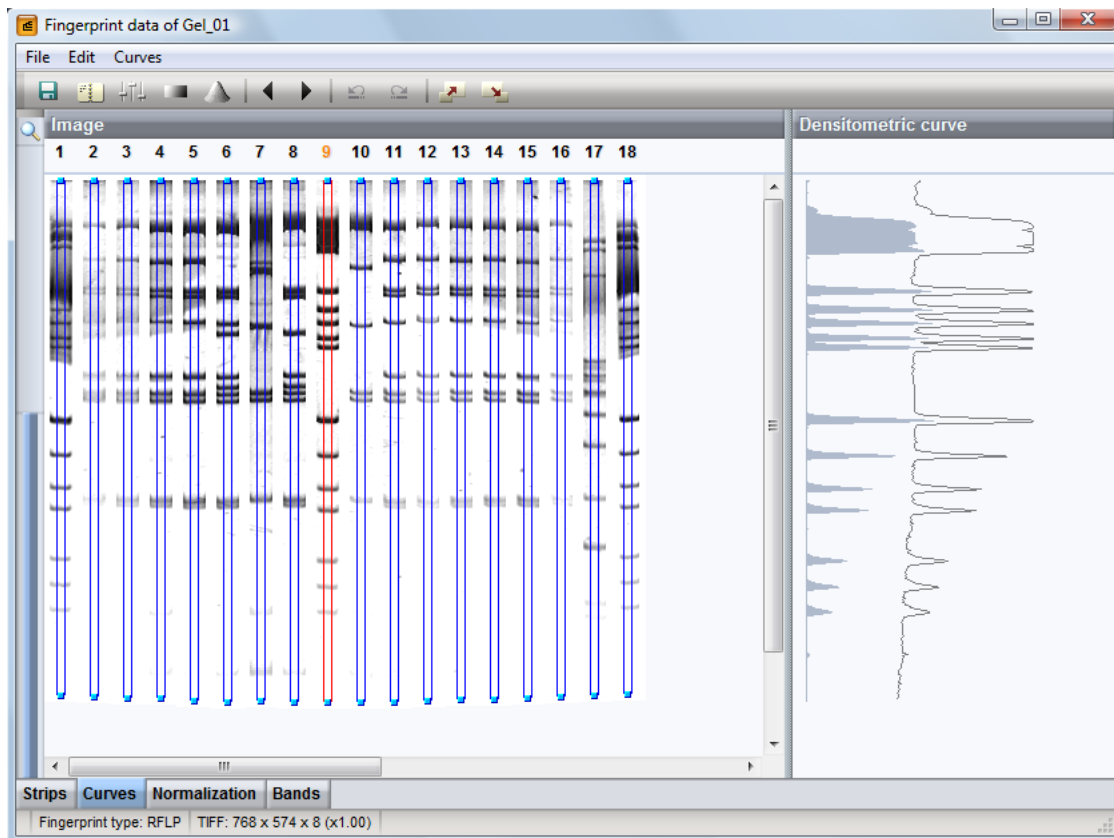


Figure 3-10. The *Fingerprint data editor* window. Step 2: defining densitometric curves.

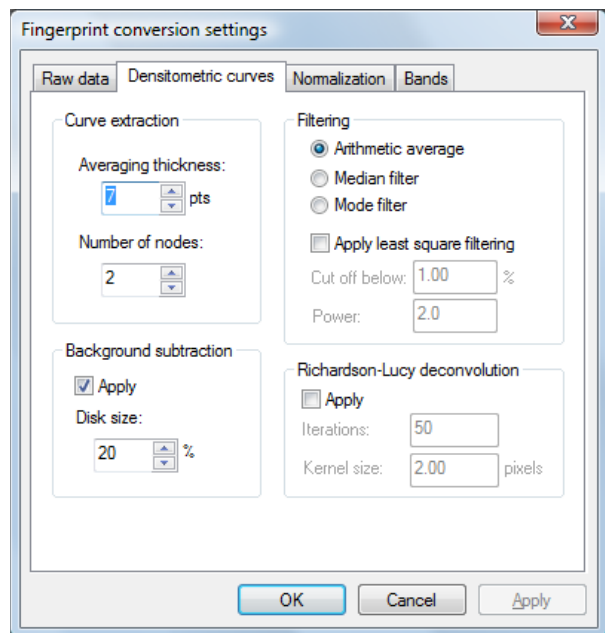


Figure 3-11. The *Fingerprint conversion settings* dialog box, *Densitometric curves* tab.

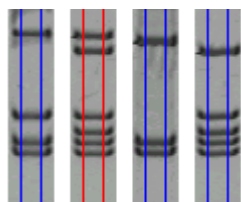


Figure 3-12. Optimal settings for curve averaging thickness.

3.2.4.5 Select *Edit > Edit settings* again to specify other settings.

The curve extraction settings include other important parameters which apply to the background removal and smoothing.

When we defined the fingerprint type, we left the *Background subtraction* disabled (see 3.2.1.6), because we will see how we can have the program propose the optimal settings.

Filtering is a method to make an average of the values within a specified thickness. Simple averaging is obtained with *Arithmetic average*, whereas *Median filter* and *Mode filter* are more sophisticated methods to reduce peak-like artifacts caused by spots on the patterns. Figure 3-13 illustrates the effect of the Median filter on a small spot. The latter two filters, however, reduce less noise on the curves (particularly the Mode filter). Only in case your gels contain hampering spots, you should use the Mode filter.

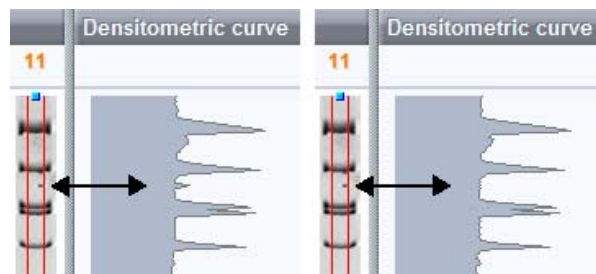


Figure 3-13. Result of *Arithmetic average filtering* (left) and *Median filtering* (right).

3.2.4.6 Select *Median filter*.

The *Least square filtering* applies to the smoothing of the profiles. This filter will remove background noise, seen as small irregular peaks, from the profile of real (broader) peaks. Like for background subtraction, the program can predict the optimal settings for least square filtering, if necessary. For now, we leave this parameter disabled.

Richardson-Lucy deconvolution is a method to *deblur* (sharpen) one-dimensional and two-dimensional arrays. This function sharpens and enhances the contrast of peaks in the densitometric curves. While peaks will become sharper, noise also will increase. Deconvolution actually does the opposite of least-square filtering. Since the method is iterative, the number of *Iterations* can be set (default 50). The more iterations, the stronger deconvolution will be obtained. The *Kernel size* (default 2.00) determines the resolution of the deconvolution: the smaller this value is set, the more shoulders will be split into separate peaks.

3.2.4.7 Press *<OK>* to save the settings.

We will now determine the optimal settings for background and filtering settings using *spectral (Fourier) analysis*.

3.2.4.8 Select *Curves > Spectral analysis*. This shows the *Spectral analysis* window (Figure 3-14).

The black line is the spectral analysis of the curves in function of the frequency in number of points (logarithmic scale). Ideally, the curve should show a flat background line at the right hand side, and then slowly raise further to the left. This indicates that the scanning resolution is high enough. Another parameter which indicates the quality is the *Signal/noise ratio*, which should be above 50 if possible. The example gel is only of moderate resolution.

The *Wiener cut-off scale* determines the optimal setting for the least square filtering. Figure 3-14 shows an optimal setting of 0.89%.

The *Background scale* is an estimation of the disk size for background subtraction. The figure shows a setting of 11%.

3.2.4.9 Call *Edit > Edit settings* again and specify the background subtraction and the least square filtering.

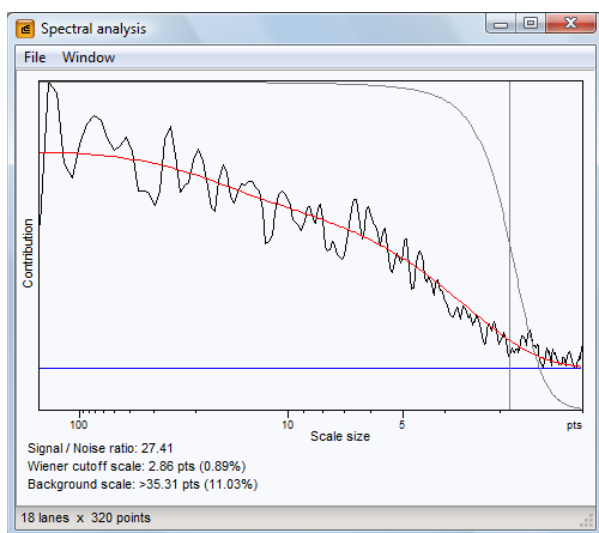



Figure 3-14. Spectral analysis of the patterns of a gel.

3.2.4.10 If you want to have a better look at the curves (right panel) you can rescale them with *Edit > Rescale curves*. This will rescale the gray processed curves (background subtracted and filtering applied) to fit within the available window space. The raw curves (lines) may then fall beyond the window.

3.2.4.11 With the command *File > Print report* or *File > Export report*, you can generate a printed or text report of the non-normalized densitometric curves, respectively.

3.2.4.12 Press  to enter the next phase: normalization of the patterns.

3.2.5 Normalizing a gel

In the Normalization step, the *Fingerprint data editor* window consists of three panels (Figure 3-15): left the *Reference system* panel, which will show the *reference positions*, and the *standard pattern*; the center panel shows the pattern strips; and the right panel shows the densitometric curve of the selected pattern.

When setting up a new database, the normalization process of the first gel involves the following steps. The underlined steps are the ones that will be followed for all subsequent gels.

- Marking the reference patterns (reference patterns are identical samples loaded at different positions on the gel for normalization purposes);
- Showing the gel in normalized view;

- Identifying a suitable reference pattern on which we will define bands as *reference positions*. Reference positions are bands that will be used to align the corresponding bands on all reference patterns from the same and from other gels.


- Defining the *reference positions*;


- Assigning the bands on the reference patterns to the corresponding reference positions;

- Updating the normalization;

- Defining a standard (optional).

We proceed as follows:

3.2.5.1 Select the first reference pattern (lane 1 on the example) and *Reference > Use as reference lane* or  (keyboard shortcut CTRL+R). Repeat this action for all other reference lanes (lanes 9 and 18 on the example).

3.2.5.2 Select *Normalization > Show normalized view* or press .

3.2.5.3 Choose the most suitable reference pattern to serve as standard: lane 9.

3.2.5.4 Select a suitable band on the destined standard pattern and *References > Add external reference position*.

You are prompted to enter a name for the band. You can enter any name, or if possible, the molecular weight of the band. In the latter case, the program will be able to determine the molecular weight regression from the sizes entered at this stage.

3.2.5.5 Use the provided scheme (see Figure 3-16) to enter all reference positions on the example gel.

Within a fingerprint type, the set of reference positions as defined, and their names, together form a *reference system*. Once a gel is normalized using the defined reference positions and saved, the reference system is saved as well. As soon as you change anything in the reference system, a position or a name, a new reference system will automatically be created in addition to the original reference system. Once a reference system has been used in one or more gels however, the program will produce a warning if you want to change anything to the reference positions.

If more than one reference system exists, one of them is the *active reference system*, i.e. the reference system to which all new gels will be normalized. Without intervention of the user, the first created reference system will always remain the default. Later, we will see how we can set the active reference system and delete unused reference systems (3.2.15).

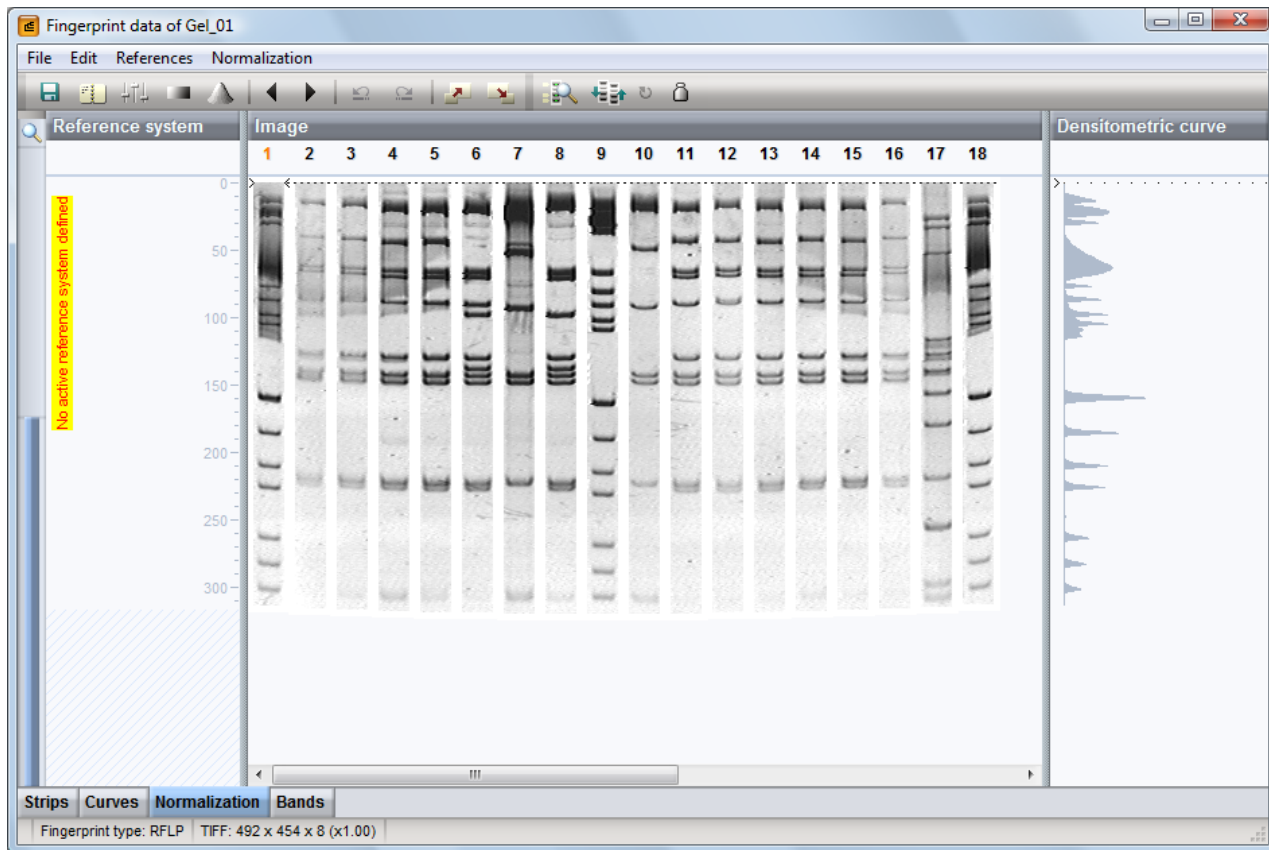


Figure 3-15. The *Fingerprint data editor* window. Step 3: normalization.

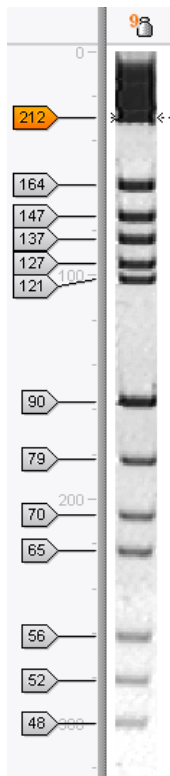


Figure 3-16. Band sizes of the reference positions on the example gel.

NOTE: Our current gel shows “No active reference system defined” in the left panel (see Figure 3-15). This message is displayed because we are processing the first gel of this fingerprint type. We already have created the reference system, but it is not saved to disk yet. Once a second gel is normalized, this message will not be displayed anymore.

The normalization is done in two steps: first are the reference bands assigned to the corresponding reference positions, and then is the display updated according to the assignments made. The last step is optional, but is useful to facilitate the correctness of the alignments made.

Assign bands manually as follows:

3.2.5.6 Click on a label of a reference position, or wherever on the gel at the height of the reference position.


3.2.5.7 Then, hold the CTRL key and click on the reference band you want to assign to that reference position.

3.2.5.8 Repeat this action for all other reference bands you want to assign to the same reference position.

3.2.5.9 Repeat actions 3.2.5.6 to 3.2.5.8 until all reference bands are assigned to their corresponding reference positions.

NOTE: The cursor automatically jumps to the closest peak; to avoid this, hold down the TAB key while clicking on a band.

3.2.5.10 With *Normalization* > *Show normalized view*,

or the  button, the gel will be shown in *normalized view*, i.e. the gelstrips will be stretched or shrunk so that assigned bands on the reference patterns match with their corresponding reference positions.


To show how the automated assignment works, we will undo the manual normalization:

3.2.5.11 Show the gel back in original view by pressing

the  button again.

3.2.5.12 Remove all the manual assignments by *Normalization* > *Delete all assignments*.

To let the program assign the bands and reference positions automatically, select *Normalization* > *Auto assign*

or . This will open the *Auto assign reference bands* dialog box (Figure 3-17). Under *Search method*, two options are available: *Using bands* and *Using densitometric curve*.

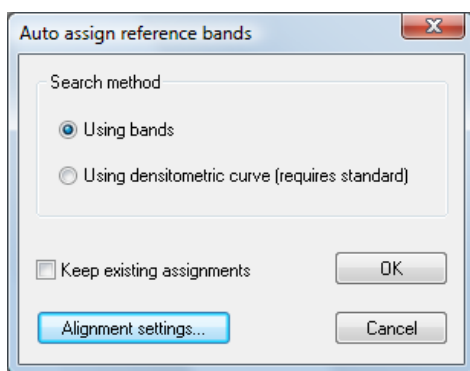


Figure 3-17. The *Auto assign reference bands* dialog box.

In the *Using bands* option, the program searches for bands on the reference patterns and tries to match them optimally with the defined reference positions. This method is always applicable, even for the very first gel, when no standard is defined.

In the *Using densitometric curve* option, a different algorithm is used, which matches the densitometric curve of standard pattern with the curves of the reference patterns. Obviously, the option requires a standard to be defined (see 3.2.9.3 to 3.2.9.7 on how to define a standard). This method employs a pattern matching algorithm that works best for complex banding patterns, but is less suitable for simple patterns such as molecular weight ladders.

An option independent of the search method is *Keep existing assignments*. When this option is chosen, any assignments made previously are preserved. This option allows the user to assign a few bands manually and let the program automatically assign the remaining bands on the reference patterns. This way of working is useful to provide some initial help to the algorithm in case of very distorted or difficult gels.

Pressing <*Alignment settings*> opens the *Alignment settings* dialog box (Figure 3-18). Parameters can be adjusted for the peak detection, global alignment and local alignment algorithms:

The **Peak detection parameters** determine what is recognized by the program as a peak.

- **Threshold** is the minimal height, expressed as a percentage of the highest peak in the profile, for which an elevation in the profile is still considered to be a peak. The default value is 2%.
- The **Valley depth** is important for peak separation: it is the minimal depth of the depression between two subsequent maxima, for which the program divides a single peak into two separate peaks. In case one maximum is higher than the other, the height between the lowest maximum and the minimum is used. Similar to the threshold, the valley depth is expressed as a percentage of the highest peak in the profile. The default value is 2%.

In a **Global alignment**, the profile as a whole is expanded (stretched) or compressed (shrunk) and displaced (shifted) to give the best possible fit with the reference positions. Depending on the status of the corresponding checkbox, a global alignment is performed or not. Not performing a global alignment can be useful in case individual reference patterns show only a minor shift, e.g. in case of fingerprints run on an automated sequencer. Regardless of the status of the checkbox, when *Keep existing alignments* is checked in the *Auto assign reference bands* dialog box (Figure 3-17), a global alignment is not performed. Instead, the program uses the distances as obtained after the first band assignment.

- The slider for **allowed expansion/compression** lets the user determine the maximally allowed expansion or compression of the profile, expressed as a percentage of the total profile length. The default value is 20%.
- The slider for **allowed displacement** lets the user determine the maximally allowed displacement (shift) of the profile, expressed as a percentage of the total profile length. The default value is 20%.

In a **Local alignment**, the profile is locally expanded (stretched) or compressed (shrunk) to match optimally with the reference positions. Using the corresponding checkbox, you can either perform or not perform a local alignment.

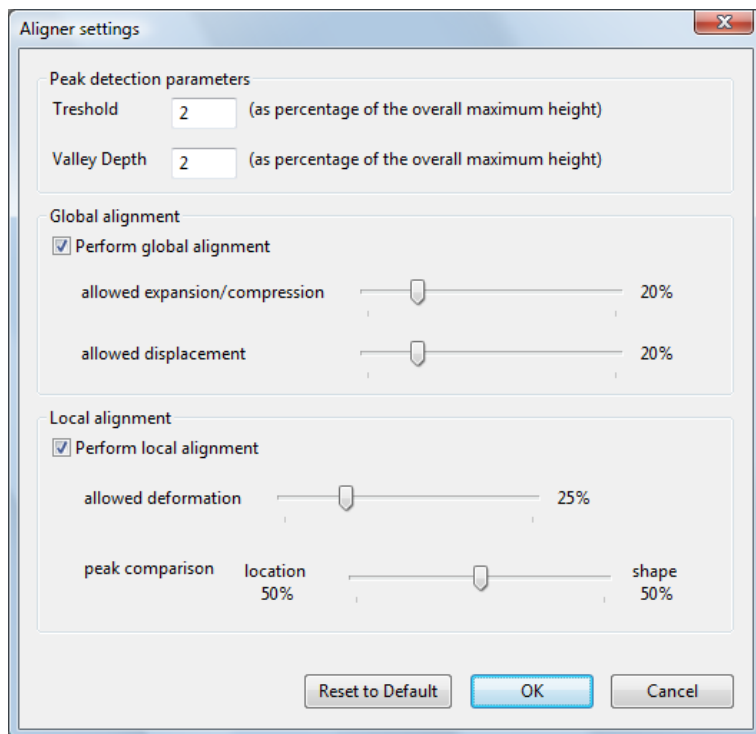



Figure 3-18. Alignment settings dialog box.

- The *allowed deformation* is the maximally allowed deformation, expressed as a percentage of the profile length. The default value is 25%.
- The *peak comparison* parameter allows the user to assign more weight on the peak **location** (position) or on the peak **shape**. The shape parameter is calculated based on a curve regression and a peak size parameter. By default, both location and shape are accounted for evenly (50%). The peak comparison parameter is only considered when *Using densitometric curves* was checked in the *Auto assign reference bands* dialog box (Figure 3-17).

Generally, the default settings perform well with most fingerprint types. Default settings can be restored by pressing **<Reset to default>**. Close the *Alignment settings* dialog box with **<OK>**.

3.2.5.13 Select **Using bands** in the *Auto assign reference bands* dialog box and press **<OK>**. Carefully inspect the assignments made, and if some are incorrect, correct them manually, as explained in 3.2.5.6 to 3.2.5.8.


3.2.5.14 Finally, when all assignments are made correctly, select **Normalization > Show normalized view**, or .

NOTE: In case most or all of the patterns on a gel contain one or more identical bands, such bands can be used for internal alignment of the gel. The software therefore creates an internal reference position which is saved with the gel but is not part of the reference system. An internal reference position can be created


with References > Add internal reference position, or right-clicking on the band and Add internal reference position. The program then asks "Do you want to automatically search for this reference band?". If you answer <Yes>, it will try to find all the correct assignments, but you can change or delete assignments afterwards.

When the gel is in normalized view, a reliable way to reveal remaining mismatches is by showing the distortion bars: these bars indicate local deviations with respect to the general shift of a reference pattern compared to the reference positions. A too strong shift is seen as a zone ranging from yellow over red to black, whereas a too weak shift is indicated by a zone ranging from bright blue over dark blue to black.

3.2.5.15 Show the distortion bars with **Normalization > Show distortion bars**.

Slight transitions from bright yellow to bright blue are normal, as long as the color does not change abruptly. In the latter case, a wrong assignment was made. You can correct the misalignment by assigning the correct band manually and **Normalization > Update normalization** or . Alternatively, you can show back the original view (3.2.5.11), assign the correct band manually, and show the normalized view again (3.2.5.14). The **Show distortion bars** setting (on or off) is stored along with the fingerprint type.

NOTE: If the program has difficulties in assigning the bands correctly, you can first make a few assignments manually (for example, the first and the last band of the

reference patterns), then display the normalized view with **Normalization > Show normalized view**, or the  button and then have the program find the assignments automatically with the option **Keep existing assignments checked**.

3.2.5.16 Save the normalized gel with **File > Save (F2)** or





3.2.5.17 It is possible to generate a text file or a printout of the complete alignment of the gel, by selecting the command **File > Export report** or **File > Print report**, respectively.

The file lists all the reference bands defined in the reference system with their relative positions, and the corresponding bands on each reference pattern, with the absolute occurrence on the pattern in distance from the start.

If you are going to use band-matching coefficients to compare the patterns, you should read the next paragraph (3.2.6), corresponding to the fourth phase in the processing of a gel (see 3.2.2). If you are going to use a curve-based coefficient, you can skip paragraph 3.2.6 and continue with 3.2.10.

3.2.6 Defining bands and quantification

In step 3 (**Normalization**), press , which brings you to the fourth step: **Defining bands and quantification**. This is the last step in processing a gel, which involves defining bands and quantifying band areas and/or volumes (see Figure 3-22).

3.2.6.1 Call the *Fingerprint conversion settings* dialog box with **Edit > Edit settings** or . The fourth tab, **Bands** is shown, which allows you to enter the **Band search filters** and the **Quantification units** (Figure 3-19).

The *Band search filters* involve a **Minimum profiling** which is the elevation of the band with respect to the surrounding background, also as percentage. The minimal profiling is dependent on the OD range you specified under **Raw data** (same dialog box, first tab). If, for example, you increase the OD range, peaks will look smaller on the densitometric profiles, and a smaller minimum profiling will need to be set in order to find the same number of bands. However, when **Rel. to max. val** is checked, the minimal profiling, i.e. the minimal height of the bands will be taken relative to the highest band on that pattern. When patterns with different intensities occur on the same gel, it is recommended to enable this option. Along with the minimum profiling, it is possible to specify a **"Gray zone"**, also as a height percentage. This gray zone specifies bands that will be

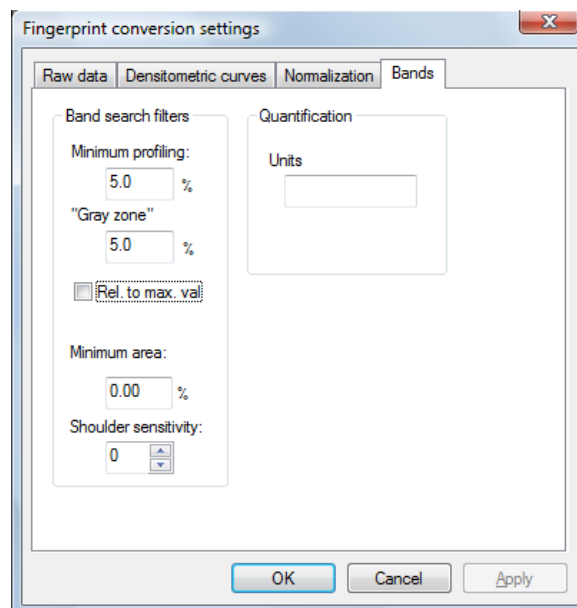


Figure 3-19. The *Fingerprint conversion settings* dialog box, **Bands** tab.

marked *uncertain*. In comparing two patterns, the software will ignore all the positions in which one of the patterns has an uncertain band. The percentage value for the gray zone is added to the minimum profiling value. To take the example of Figure 3-19, all bands with a profiling of less than 5% are excluded; bands with a profiling between 5% and 10% are marked uncertain, and all bands with a profiling of more than 10% are selected (see Figure 3-20).

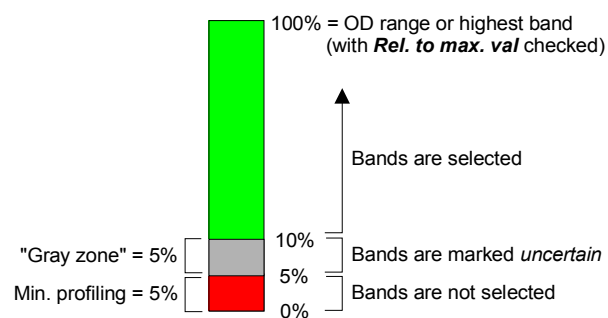



Figure 3-20. Understanding the meaning of the **"gray zone"** of uncertain bands in relation to the **minimum profiling**.

A **Minimum area** can also be specified, as percentage of the total area of the pattern.

A more advanced tool based on deconvolution algorithms, **Shoulder sensitivity**, allows shoulders without a local maximum as well as doublets of bands with one maximum to be found. If you want to use the shoulder sensitivity feature, we recommend to start with a sensitivity of 5, but optimal parameters may depend on the type of gels analyzed.

3.2.6.2 Change *Minimal profiling* to adjust the minimal peak height (in percent of the highest peak of the pattern), and/or *Minimal area* to adjust the minimal area, in percent of the total area of the pattern. Usually, setting 5% minimal profiling will be convenient, whereas the minimal area can be left zero in most cases. The present example however, requires a higher minimal profiling (e.g. 10%). Optionally, you can enter a percentage for uncertain bands (gray zone). As an example to see what happens, enter 5%. Click *Relative to max. value of lane*. Specify a *Shoulder sensitivity* only if you want to allow the program to find band doublets and bands on shoulders (sensitivity of 5 should be fine for most gels).

3.2.6.3 Press <OK> to accept the settings.

3.2.6.4 Select *Bands > Auto search bands* or  to find bands on all the patterns.

Before actually defining the bands on the patterns, the software displays a preview window (Figure 3-21). This preview shows the first pattern on the gel with its curve and gelstrip. Press the <Preview> button to see what bands the program finds using the current settings. A pink mask shows the threshold level based upon both the minimal profiling and the minimal area (if set). Only bands that exceed the threshold will be selected. If inappropriate, the settings can be changed in this preview window. The sensitivity of this search depends on the *band search settings*: if too many (false) peaks are found, or if real bands are undetected, you can change the search sensitivity using the band search filters as described above.

In addition, a blue mask shows the threshold level for bands that will be found as uncertain (gray zone). All bands exceeding the pink mask but not exceeding the blue mask will become uncertain bands.

In the *Band search preview* window, the currently selected pattern is shown and indicated in the status bar (bottom). To scroll through other patterns in the preview, press the < or > button (left and right from the curve).

You can search for bands on an individual lane by pressing <Search on this lane>, or on all lanes of the gel at once by pressing <Search on all lanes>.

3.2.6.5 Press <Search on all lanes> to start the search on the full gel.

NOTE: If bands were already defined on the gel, the program will now ask "There are already some bands defined on the gel. Do you want to keep existing bands?". If you answer <No>, the existing bands will be deleted before the program starts a new search. By answering <Yes>, you can change the search settings and start a new search while any work done previously is preserved.

Bands that were found are marked with a green horizontal line, whereas uncertain bands are marked with a small green ellipse (see magnification in Figure 3-22).

3.2.6.6 Add a band with *Bands > Add new band*, the ENTER key, or CTRL + left-click.

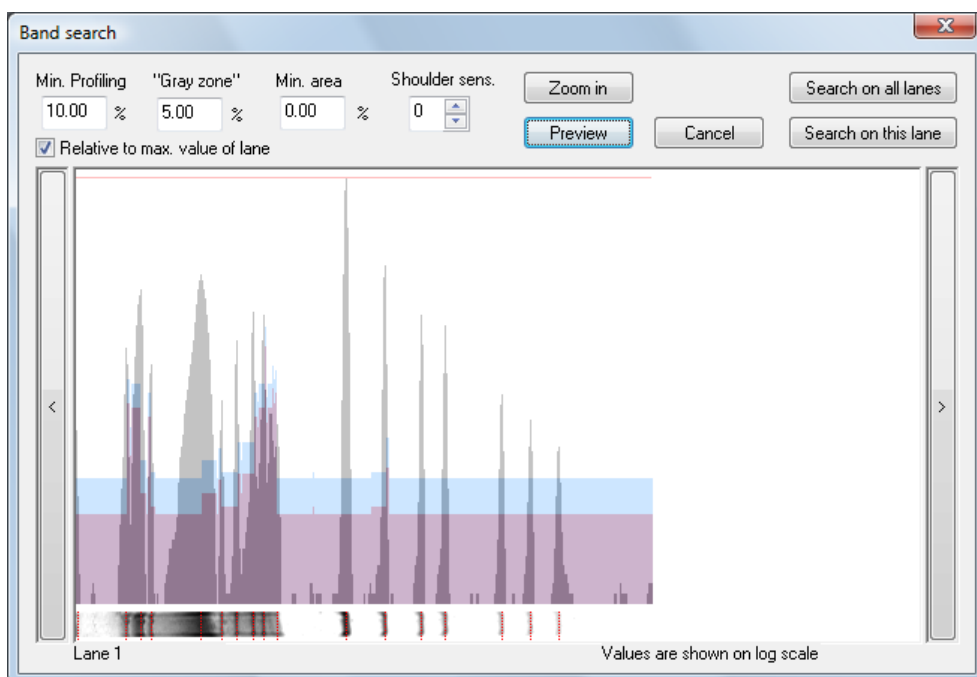



Figure 3-21. *Band search preview window.*

NOTES:

(1) The cursor automatically jumps to the closest peak; to avoid this, hold down the TAB key while clicking on a band.

(2) When there is evidence of a double band at a certain position, you can add a band over an existing one (3.2.6.6). Double bands (or multiplets) are indicated by outwards pointing arrows on the band marker:

 . Double uncertain bands are marked with a filled ellipse instead of an open ellipse. The clustering and identification functions using band based similarity coefficients (4.1.9) support the existence of double overlapping bands. For example, two patterns, having a single band and a double band, respectively, at the same position will be treated as having one matching and one unmatched band. Two patterns, each having a double band at the same position, will be treated as having two matching bands.

3.2.6.7 Hold the SHIFT key and drag the mouse pointer holding the left mouse button to select a group of bands.

3.2.6.8 Press the DEL key or **Bands > Delete selected band(s)** to delete all selected bands.

3.2.6.9 Select a band and **Bands > Mark band(s) as uncertain** (or press F5).

3.2.6.10 With **Bands > Mark band(s) as certain** (or press F6), the band is marked again as certain.

3.2.7 Advanced band search using size-dependent threshold

In many electrophoresis systems, staining intensity of the bands is dependent on the size of the molecules. In DNA patterns stained with ethidium bromide for example (e.g. Pulsed-Field Gel Electrophoresis, PFGE), larger DNA molecules can capture many more ethidium bromide molecules than small DNA molecules, resulting in large size bands to appear much stronger than small size bands.

In other electrophoresis systems, the definition of the bands (sharpness) might depend on the size, which can also result in apparent different height depending on the position on the pattern.

In such systems, a method that uses a single threshold parameter for finding bands on the patterns (i.e. the minimum profiling) might not work well: in case of PFGE for example, in the high molecular weight zone it might detect spots and irrelevant fragments whereas in the low molecular weight zone real bands might remain undetected.

In order to provide a more accurate band search for patterns with systematic dependence of intensity according to the position, InfoQuest FP provides a way to calculate a regression that reflects the average peak intensity for every position on the patterns in a given fingerprint type. The only requirement for this method is that a sufficient number of gels already needs to be processed, with the bands defined appropriately, before the regression can be calculated. The user can make a

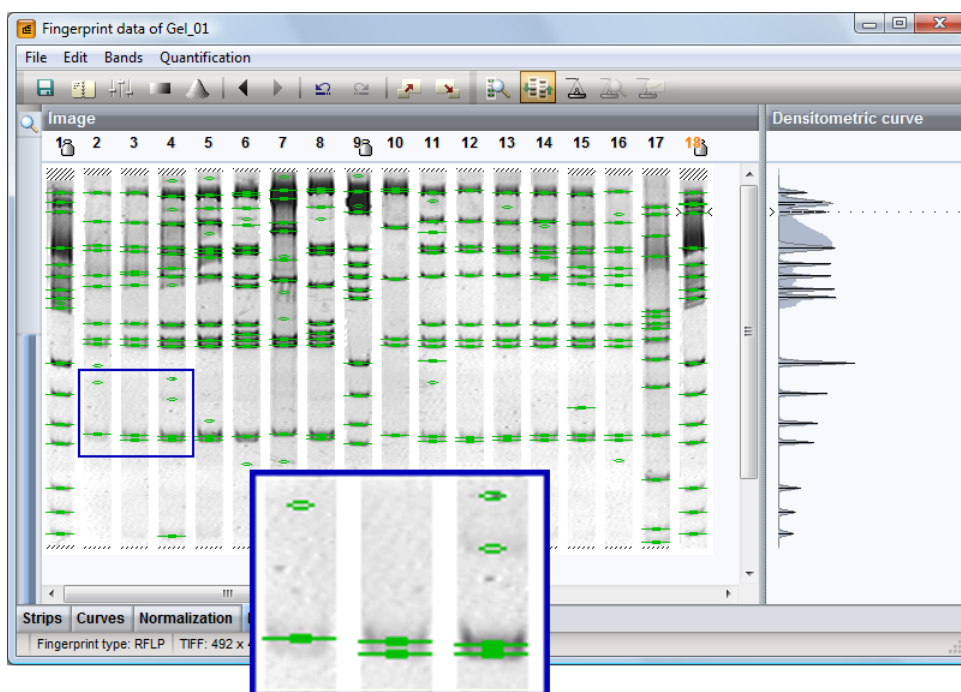



Figure 3-22. The *Fingerprint data editor* window. Step 4. Bands.

selection of entries from the database, and based upon that selection and the bands they contain in the fingerprint type, the regression is established.

To obtain the regression, we proceed as follows.

3.2.7.1 Open **DemoBase** in the *InfoQuest FP main* window.

3.2.7.2 Select **Edit > Search entries** or press F3 or .

This pops up the *Entry search* dialog box (see 2.2.8 for detailed explanation on search and select functions).

3.2.7.3 In the *Entry search* dialog box, check **RFLP1** and press **<Search>**. All entries having a pattern of **RFLP1** associated are now selected in the database, which is visible as a colored arrow left from the entry fields (see 2.2.8).

3.2.7.4 In the *Experiments* panel, double-click on **RFLP1** to open the *Fingerprint type* window.

3.2.7.5 In the *Fingerprint type* window, select **Settings > Create peak intensity profile**. This pops up the *Peak intensity profile* window, a plot of all intensities of the selected patterns in function of the position on the pattern (Figure 3-23).

3.2.7.6 Initially, the threshold factor is a flat line at 1.0. By pressing **<Calculate from peaks>**, a non-linear regression is automatically calculated from the scatter plot (Figure 3-23).

3.2.7.7 The regression line contains 5 nodes, of which the position can be changed independently by the user. To

change a node's position, click and hold the left mouse button and move the node to the desired position.

3.2.7.8 The regression can be reset to a flat line using the **<Reset>** button. To confirm and save the regression, press **<OK>**.

The regression can be edited anytime later by opening the *Peak intensity profile* window again (3.2.7.5). As a result of creating a peak intensity regression curve, the minimum profiling threshold (3.2.6.2) will be dependent on the curve. The value entered for the minimum profiling will correspond to the highest value on the intensity profile regression curve (the outermost left point in Figure 3-23). Therefore, after creating an intensity profile regression, you may have to increase the minimum profiling setting to find the bands optimally: noise and irrelevant peaks will be filtered out in the high intensity areas whereas faint bands will still be detected in the low intensity areas.

3.2.8 Quantification of bands

The right panel in the fourth step of the *Fingerprint data editor* window shows the densitometric curve of the selected pattern. For each band found, the program automatically calculates a best-fitting *Gaussian* curve, which makes more reliable quantification possible.

3.2.8.1 Select a band on a pattern.

3.2.8.2 Show rescaled curves with **Edit > Rescale curves**.

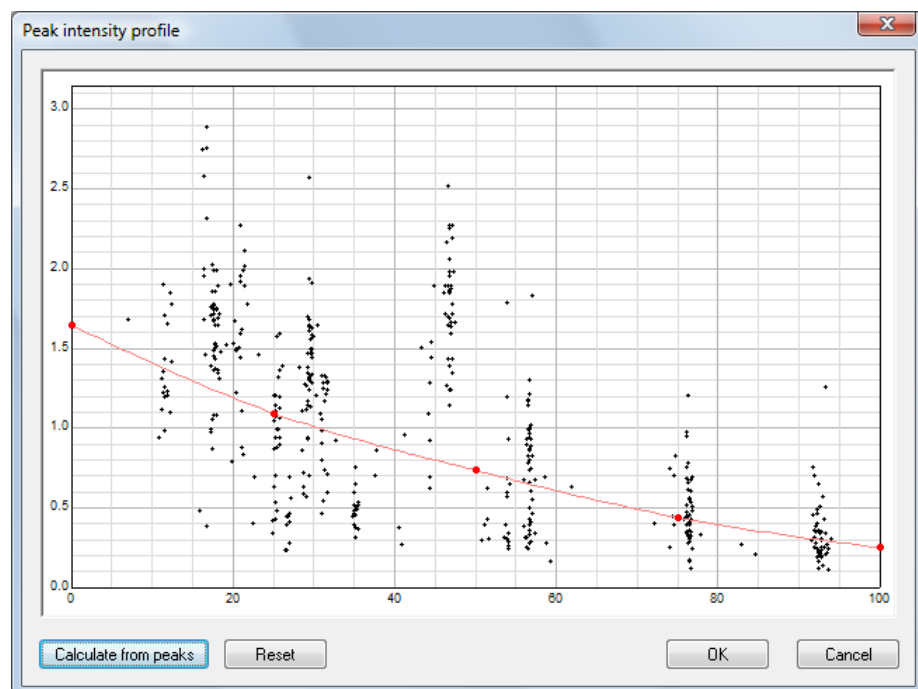



Figure 3-23. The *Peak intensity profile* window with peak intensity regression curve.

3.2.8.3 Zoom in on the band by pressing  repeatedly or use the zoom slider (see 1.6.7 for instructions on the use of zoom sliders). Figure 3-24 shows a strongly zoomed band with its densitometric representation and the Gaussian fit (red). The blue points are dragging nodes where you can change the position and the shape of the Gaussian fit for each band separately.

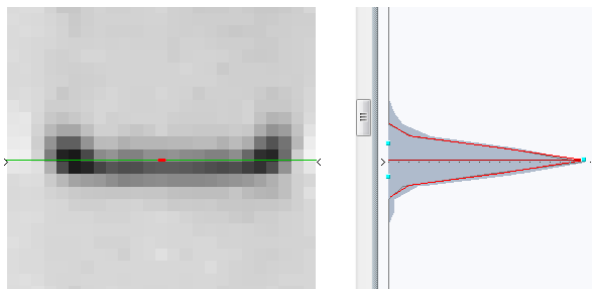



Figure 3-24. Zoomed band with its densitometric curve and best-fitting Gaussian approach.

3.2.8.4 Save the gel with *File > Save* (F2) or  .

3.2.8.5 It is possible to generate a text file or a printout of the complete band information of the gel, by selecting the command *File > Export report* or *File > Print report*, respectively.


The file lists all the bands defined for each pattern with their normalized relative positions, the metrics (e.g. molecular weight), the height, and relative one-dimensional surface, as calculated by Gaussian fit.

Once bands are defined, two-dimensional quantification is done as follows.


3.2.8.6 Bring the window in *Quantification mode* with



or *Quantification > Band quantification*. The

quantification button now shows as  and two additional band quantification buttons are shown.

3.2.8.7 To find the surfaces (contours) of the bands, use


Quantification > Search all surfaces or .

If you have added a band later, you can search the surface of that band alone with *Quantification > Search surface of band*.

When the contours are found, the program shows for each selected band its *volume* in the status bar: the sum of the densitometric values within the contour.

3.2.8.8 To change the contour of a band manually, first select the band and zoom in heavily (3.2.8.1 and 3.2.8.3).

3.2.8.9 Hold the CTRL key and drag the mouse (holding the left button) to correct the upper and lower contours.

3.2.8.10 For known reference bands, you can enter a concentration value by selecting the band and *Quantification > Assign value* (or from the floating menu that appears by right-clicking, or just double-click on the band). Known reference bands are marked with .

3.2.8.11 Once multiple reference bands are assigned their concentrations, a regression to determine each unknown band concentration is calculated by selecting *Quantification > Calculate concentrations*.

The *Band quantification* window (Figure 3-25) shows the real concentration in function of the band volumes, using cubic spline regression functions.

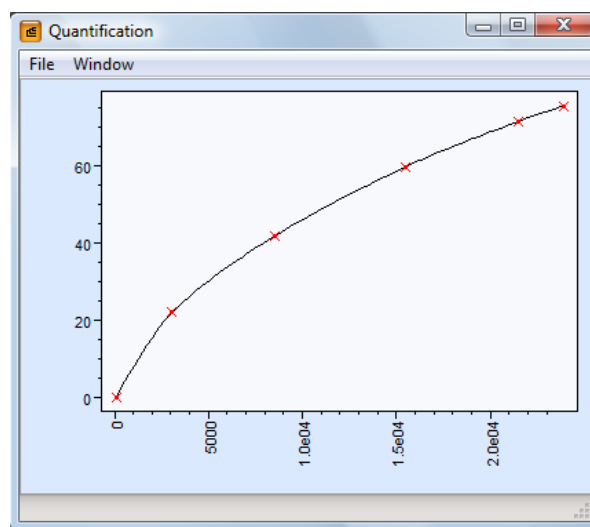



Figure 3-25. *Band quantification* window: concentration in function of known band volumes.

3.2.8.12 Save the gel with *File > Save* (F2) or  in order to store the quantification data.

3.2.8.13 It is possible to generate a text file or a printout of the complete two-dimensional band information of the gel, by selecting the command *File > Export report* or *File > Print report*, respectively.

The file lists all the bands defined for each pattern with their normalized relative positions, the absolute volume, and if regression is done, the relative volume as determined by the calibration bands.

We are now at a point that we can discuss the functioning of the *reference system*. We will explain how to calculate molecular weights for the fingerprint type and how to link a *standard pattern* to the fingerprint type.

3.2.8.14 Exit the *Fingerprint data editor* window: *File > Exit*.

The program asks “*Settings have been changed. Do you want to use the current settings as new defaults?*”. This question is asked when changes have been made to the fingerprint type-related settings, for example the gelstrip thickness, the rolling disk size, etc. If you answer <Yes>, the settings used for this gel will be saved in the fingerprint type’s settings, and all new gels will be processed using the same settings.

3.2.8.15 Answer <Yes> to save the changes made into the fingerprint type settings.

*NOTE: Answering <Yes> to the above question has the same effect as the menu function **Edit > Save as default settings** in the Fingerprint data editor window. Conversely, the current default settings can be copied to the current gel with **Edit > Load default settings**.*

3.2.9 Editing the fingerprint type settings

To show that the reference system is now defined for our gel type RFLP, we will open the *Fingerprint type* window.

3.2.9.1 In the *InfoQuest FP* main window, select **RFLP** in the *Experiments* panel (see Figure 1-15). Double-click on **RFLP**, or select *Experiments > Edit experiment type* in the main menu. This opens the *Fingerprint type* window (Figure 3-26).

3.2.9.2 The *Fingerprint type* window allows you to change all settings which we have defined when

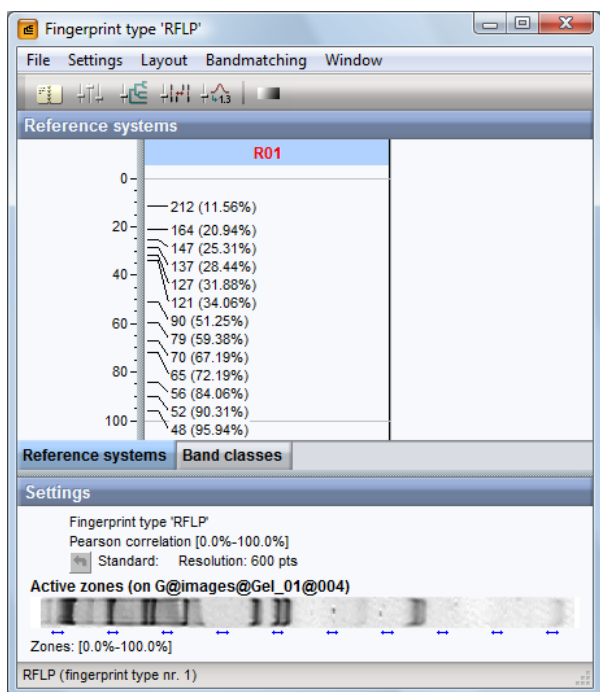





Figure 3-26. The *Fingerprint type* window; Standard is not yet defined.

creating the fingerprint type, and when processing the first gel with *Settings > General settings* or .

One setting which we have not discussed during the normalization of the example gel is the *Normalization* tab. This tab shows the *Resolution of normalized tracks* as only setting. In reality, the program always stores the real length of the raw patterns. For display purposes however, the program converts the tracks to the same length at real-time, so that the gel strips are properly aligned to each other. For comparison of patterns by means of the *Pearson* product-moment correlation, the densitometric curves also need to be of the same length. Thus, the resolution value only influences two features: the length of the patterns shown on the screen, and the length (resolution, number of points) of the densitometric curves to be compared by the *Pearson* product-moment correlation coefficient. By default, the program uses 600 as resolution, but when you normalize the first gel, the program automatically uses the *average track length* for that gel as the new resolution value. Whenever you save the gel, and the value differs more than 50% from the default value, InfoQuest FP will ask you to copy the resolution of the current gel to the default for the fingerprint type (see 3.2.3.13). Another option is *Bypass normalization*. You can use this option to have the program process the densitometric curves of the tracks *without any change*. This option is only useful to import patterns in InfoQuest FP that are already normalized, and for which you want the values of the densitometric curves to remain exactly the same after the normalization process.

The default brightness and contrast setting can be changed with *Layout > Brightness & contrast* or , and the quantification settings with *Settings > Comparative quantification* or . Further settings include the comparison settings, and the position tolerance settings, which will be discussed later.

The *Fingerprint type* window shows the defined reference positions in relation to the distance on the pattern (in percentage), and calls this reference system R01. Other reference systems (if created automatically) will be called R02, R03 etc. Currently, R01 is shown in red because it is the *active reference system*.

In this window, the panel for the **Standard** is still blank: the fingerprint type still misses a standard pattern. The standard pattern actually has no essential contribution to the normalization; it is only intended to show a normalized reference pattern next to the reference positions, in order to make visual assignment of bands to the reference positions easier. Another feature for which the standard is required is the automated normalization by pattern recognition. This algorithm requires a curve of a normalized reference pattern to be present in order to be able to align other reference patterns to it.

Now, link a standard to the fingerprint type as follows:

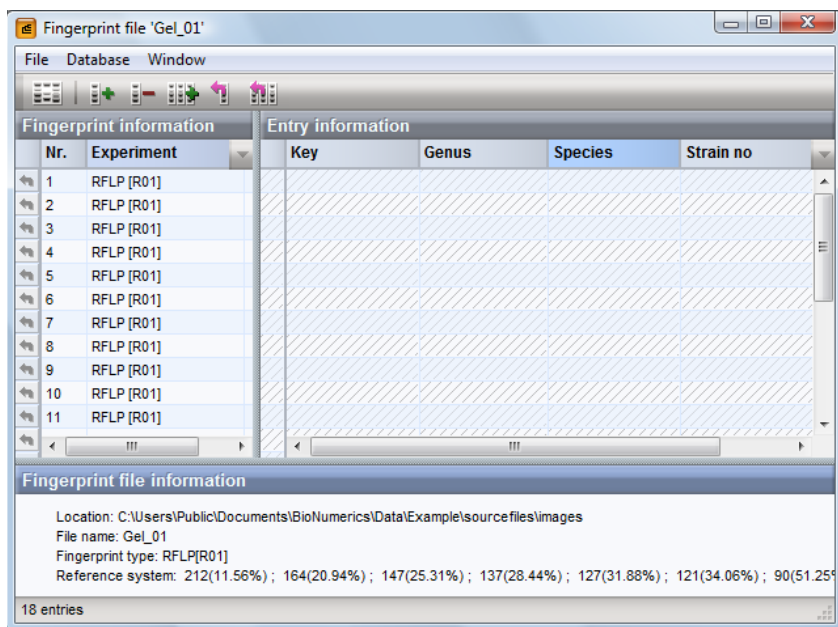





Figure 3-27. The Fingerprint entry file window.

3.2.9.3 Close the *Fingerprint type* window for now (**File > Exit**).

3.2.9.4 Select the gel file in the *Experiment files* panel () and choose **File > Open experiment file (entries)** from the main menu or press  in the toolbar of the panel.

This opens the *Fingerprint entry file* window, listing the lanes defined for the example gel (Figure 3-27).

These lanes are not linked to database entries yet. A *link arrow*  for each lane allows you to link a lane to a database entry, by clicking on the arrow and dragging it onto a database entry, and then releasing the mouse button. When the experiment is linked, its link arrow is purple: . The window also shows the *Fingerprint type* of the gel, the *reference system* according to which the gel is normalized, and the *reference positions* of this reference system.

3.2.9.5 In the *InfoQuest FP main* window, add a new database entry with **Database > Add new entries** (see 2.2.1.3 to 2.2.1.4).

3.2.9.6 Edit the new entry's information fields (see 2.2.3.1 to 2.2.3.2) and enter STANDARD as genus name.


3.2.9.7 Drag the link arrow of **lane 9** to the new database entry 'STANDARD': pattern 9 is now linked to this database entry.

3.2.9.8 In the *Fingerprint entry file* window, select the lane marked as STANDARD and choose **Database > Set lane as standard**. The program will ask a confirmation.

Alternatively, the standard can also be assigned using a drag-and-drop operation from the *Fingerprint type* window, as follows:

3.2.9.9 Close the *Fingerprint entry file* window with **File > Exit**.

3.2.9.10 In the *InfoQuest FP main* window, open the *Fingerprint type* window again for **RFLP** (3.2.9.1).

3.2.9.11 Link a reference lane (for example lane 9) to the fingerprint type by dragging the  button to the database entry STANDARD.

The standard pattern is now displayed in the *standard* panel next to the reference positions, and the database entry key of the standard is indicated next to the link arrow (Figure 3-28). From this point on, all further gels that are normalized will display the standard pattern left from the gel panel in the normalization step. This makes manual association of peaks easier and allows automated alignment using curve matching.

NOTE: The choice of a standard has no influence on the normalization process, since it is only used as a visual aid. One can change the standard pattern at any time later on, e.g. if another reference pattern appears to be more suitable for this purpose.

The *molecular sizes* of the bands are not calculated within a particular gel file, but for a whole reference system. This means that, once you have created a reference system and normalized one gel, you can define the molecular size regression for all further gels that will be normalized using the same reference system.

3.2.9.12 In the *Fingerprint type* window for **RFLP**, call **Settings > Edit reference system** (or double-click in the

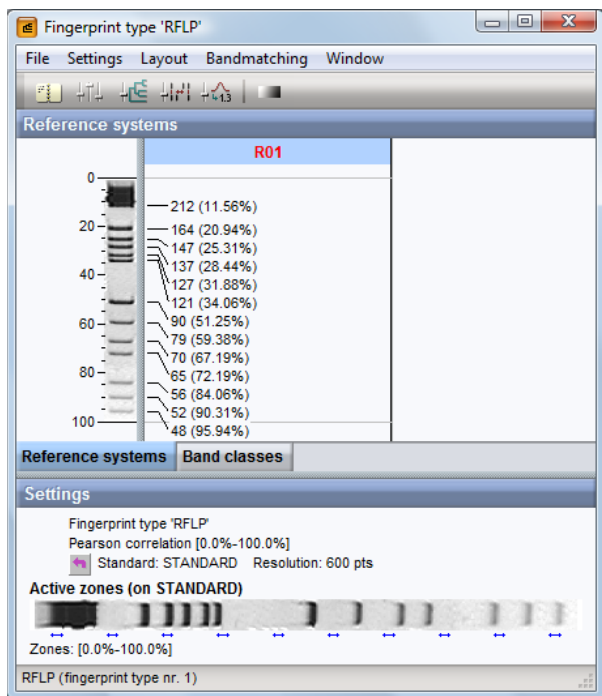


Figure 3-28. The *Fingerprint type* window; **Standard** is defined.

R01 panel). This pops up the *Reference system* window for fingerprint type **RFLP** (Figure 3-29).

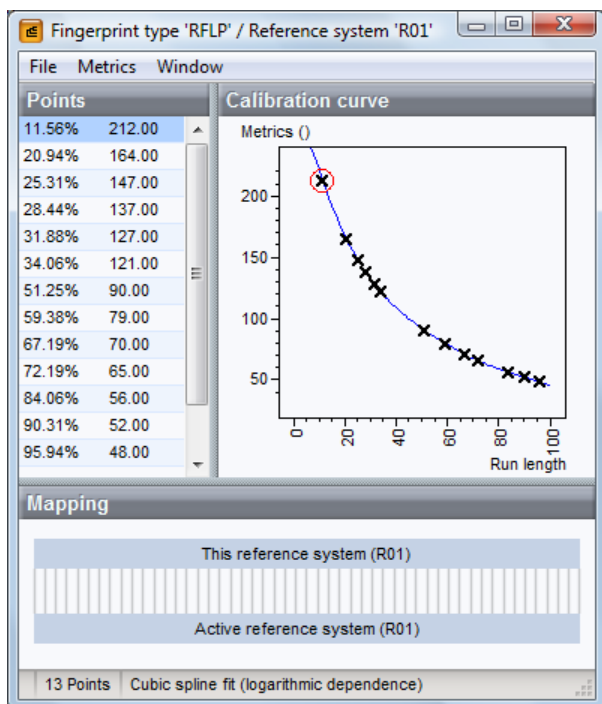


Figure 3-29. The *Reference system* window, showing molecular weight regression and remapping function to the active reference system (if different).

Initially, the regression cannot be calculated, since the program does not know where to take the marker points

from. The message “*Could not calculate calibration curve. Not enough markers*” is displayed.

3.2.9.13 You can add the markers manually (*Metrics > Add marker*), but if you have entered the molecular weights as names for the reference positions (see 3.2.5.4 and 3.2.5.5), the obvious solution is to copy these molecular weights: *Metrics > Copy markers from reference system*.

The result is a regression curve, shown in Figure 3-29. As regression function, you can choose between a first degree, third degree, cubic spline, and pole fit, and each of these functions can be combined with a logarithmic dependence.

3.2.9.14 For this example, choose *Metrics > Cubic spline fit with Logarithmic Dependence*.


3.2.9.15 Choose a unit with *Metric > Assign unit*, and enter **bp** (base pairs).

3.2.9.16 Close the *Reference system* window, and close the *Fingerprint type* window.

NOTE: The Band classes panel in the Fingerprint type window (displayed as a tab in Figure 3-28) will be discussed in .


3.2.10 Adding gel lanes to the database

In paragraph , we have seen how entries are added to the database. Once these entries are defined in the database, it is easy to link the experiments, which are gel lanes in this case, to the corresponding entries. We have done so with the STANDARD lane, explained in the previous paragraph. In summary, adding lanes to the database and linking experiments to them works as follows:


3.2.10.1 Select *Database > Add new entries* or  in the toolbar.


3.2.10.2 Enter the number of entries you want to create, e.g. 1, and press <OK>.

The database now lists one more entry with a unique key automatically assigned by the software.

3.2.10.3 Select the gel file in the *Experiment files* panel () and choose *File > Open experiment file (entries)* from the main menu or press  in the toolbar of the *Experiment files* panel.



This opens the *Fingerprint entry file* window, listing the lanes defined for the example gel (Figure 3-27).


These lanes are not linked to database entries yet. A *link arrow*  for each lane allows you to link a lane to a

database entry, by clicking on the arrow and dragging it onto a database entry, and then releasing the mouse button. When the experiment is linked, its link arrow is purple: .

3.2.10.4 Drag the link arrow of **lane 2** (lane 1 is a reference) to the new database entry: as soon as you pass over a database entry, the cursor shape changes into



3.2.10.5 Release the mouse button on the above created database entry; pattern 2 is now linked to this database entry, and its arrow in the *Fingerprint entry file* window has become purple  instead of gray .

The corresponding menu command is **Database > Link lane** (). The program asks you to enter the *key* of the database entry to which the experiment is to be linked.

NOTES:

(1) If you try to link a lane to an entry which already has a lane of the same experiment type linked to it, the program will ask whether you want to create a **duplicate key** for this entry. This feature is very useful in case you want to define experiments that are run in duplicate for one or more organisms. Rather than overwriting the first entry or disregarding duplicate entries, InfoQuest FP automatically considers them as duplicates and assigns an extension /#x to such duplicates. In case for a given entry a duplicate already exists (after import of another experiment), InfoQuest FP will automatically fill such existing duplicates that are still empty for the experiment type that is being imported. Database fields are automatically taken over from the "master" entry, i.e. the entry without extension. If the database fields from the "master" entry are changed, the /#x duplicates are automatically changed accordingly.


(2) If you enter an entry key which does not already exist, the program asks whether you want to create an entry with that key.

As soon as an experiment is linked to a database entry, the *Experiment presence* panel (see Figure 1-15) shows a colored dot for the experiment of this entry.

3.2.10.6 You can click on such a colored dot, which pops up the *Experiment card* for that experiment (see 3.8.2).


3.2.10.7 You can edit the information fields for this entry in several places: directly in the database (see 2.2.3.1 to 2.2.3.2), but also e.g. in the *Fingerprint entry file* window or in the *Comparison* window. In all cases, double-clicking on the entry calls the corresponding *Entry edit* window and clicking twice on the same information field enables direct editing.


If no database entries are defined for the current gel lanes, you can have the program create new entries and link the gel lanes automatically in a very simple way:

3.2.10.8 In the *Fingerprint entry file* window, select **Database > Add all lanes to database** (). All lanes that were not linked yet, will be added as new entries to the database, with the gel lanes linked.

*NOTE: In some cases, a gel can be composed of patterns belonging to different fingerprint types. For example, if you are running digests by three different restriction enzymes for the same set of organisms, for some remaining entries, you may want to run all three restriction enzyme digests on the same gel. In this case, you should process the gel according to one of the fingerprint types, and then, in the Fingerprint entry file window, select a lane that belongs to another fingerprint type and **Database > Change fingerprint type of lane**. A condition for this feature to work is that both fingerprint types are based upon the same reference system (the same set of reference markers, defined consistently using the same names). If the reference system for both fingerprint types is not the same, the software can still use the molecular weight calibration curves as a basis for conversion, if these are defined.*

If you do not wish to add all lanes to the database, you can select individual lanes, and use the menu command

Database > Add lane to database (.

You can unlink a gel lane from the database using **Database > Remove link** (). All entries from the gel are unlinked at once using **Database > Remove all links**.

3.2.11 Adding information to fingerprint files and fingerprint lanes

In InfoQuest FP, it is possible to assign information fields to fingerprint files in an easy way. This is useful to store information such as gel processing parameters, person who ran the gel, etc.

3.2.11.1 Right-click in the information fields header of the *Fingerprint files* panel.

3.2.11.2 From the floating menu, select **Add new information field** (see Figure 3-30).

3.2.11.3 Enter a name for the new information field (e.g. "Done by") and press <OK>.

The newly created information field is added in the information fields header. Clicking twice in an information field enables editing.

In addition to fingerprint file (i.e. gel-) specific information, it is possible to store information specific to individual fingerprint lanes. Recording lane-specific information could be useful e.g. to comment on PCR (RAPD, AFLP) or restriction digest (RFLP) efficiency of individual reactions. This feature is only accessible when working with a connected database (see section 2.3).

3.2.11.4 Double-click on the fingerprint file to open the *Fingerprint file* window.

3.2.11.5 Right-click in the information fields header of the *Fingerprint information* panel and select **Add fingerprint information field** from the floating menu (or select **File > Add fingerprint information field** from the menu).

3.2.11.6 Enter a name for the new fingerprint lane information field (e.g. "Comment").

The information field is added in the *Fingerprint information* panel. Clicking twice in the information field enables editing.

3.2.11.7 The lane information can also be visualized and edited from the *Experiment card* (Figure 3-84), by clicking the right mouse button inside the gelstrip window and selecting **Fingerprint information fields** in the floating menu that pops up.

3.2.11.8 Fingerprint lane fields can also be used in the *Advanced query tool* (see 2.2.9), using the search option **<Fingerprint field>**. This search option is only available if fingerprint lane information fields are defined. The first time after defining fingerprint fields, you will have to restart the program in order make the button **<Fingerprint field>** available in the advanced query tool.

3.2.12 Superimposed normalization based on internal reference patterns

This paragraph describes how to normalize patterns based upon "inline" reference patterns, i.e. reference patterns that are loaded in each lane, but that are revealed using a different color dye or hybridization probe. Examples within this category are (1) multichannel automated sequencer chromatograms, such as the **AFLP** experiment in the example database **DemoBase** and (2) RFLP gels that contain internal refer-

ence patterns which are visualized using a different color dye or hybridization probe.

In case (1), a special import program, **CrvConv** is required to convert the multichannel sample chromatogram files into the InfoQuest FP curve format. It can read chromatogram files from ABI, Beckman, and Amersham MegaBace. **CrvConv** splits the multichannel sample files into separate gel files for each available channel (color). Logically, the separate gels all contain the same lanes at the same position. One of the gels contains the internal reference patterns, whereas the other gel (or gels) contain the real data samples, to be normalized according to the reference patterns. The aim is to normalize the obtained reference gel, and to superimpose the normalization on the other gel(s). The only difference with TIFF files is that there are no two-dimensional gelstrips available for the sequencer patterns. InfoQuest FP creates reconstructed gelstrips instead. A non-reference gel (i.e. a real data gel) is normalized by first normalizing the reference gel (i.e. the gel containing the internal reference patterns), and then copying the normalization of the reference gel to the data gel. This can be done easily by simply linking the data gel(s) to the corresponding reference gel: each data gel is automatically updated when anything in the conversion and normalization of the reference gel is changed.

In case (2), the conversion from the sequencer sample files is not needed, but on the other hand, an initial alignment between the images of the reference gel and the data gel is needed, since both images are usually not scanned in exactly the same position. The further steps are the same as for the automated sequencer gels: A non-reference gel (i.e. a real data gel) is normalized by first normalizing the reference gel (i.e. the gel containing the internal reference patterns), and then copying the normalization of the reference gel to the data gel, or linking the data gel to the reference gel.

A. Multichannel sequencer gels

3.2.12.1 Create a new database (see 1.5.2).

A set of example files together composing one gel can be found on the CD-ROM in the **Sample and Tutorial data\AB sequencer trace files** directory. The same files are also available from the download page of the website (www.bio-rad.com/softwaredownloads). We are going to import these Applied BioSystems files in our database.

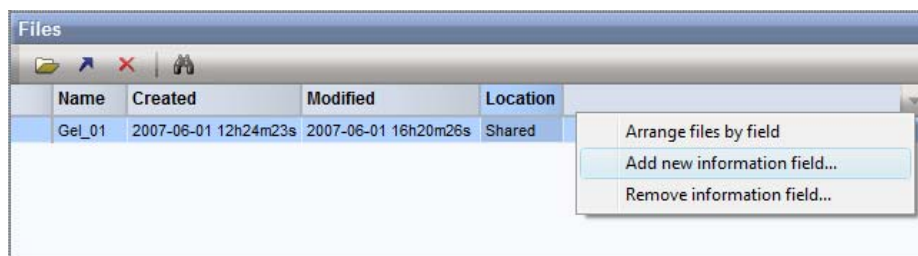


Figure 3-30. Adding information fields in the *Fingerprint files* panel.

In order to import chromatogram files from Applied BioSystems, Beckman, and Amersham MegaBace via the CrvConv program, the **Import** plugin needs to be installed.

3.2.12.2 Install the **Import** plugin (see paragraph 1.5.3 for more information).

3.2.12.3 Choose **File > Import > Import Fingerprint files from Automated Sequencers**.

A dialog box pops up, listing two different import options:

- If you are importing Applied BioSystems sequencer files into a connected database (both conditions should be met), you can select **Use automated import for AB files only**. With this option checked, InfoQuest FP imports Applied BioSystems chromatogram files without opening the CrvConv program.
- If you want to import other types of sequencer files and/or work in a local database, you should check **Open Curve Converter (all formats)**. This option opens the **CrvConv** program to import your chromatogram files in the InfoQuest FP software. Applied BioSystems, Beckman, and Amersham MegaBace chromatogram files are supported.

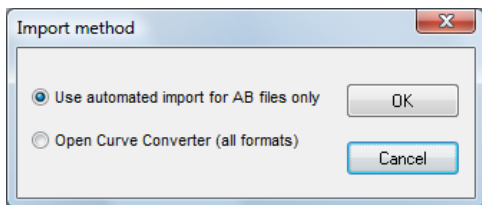


Figure 3-31. Two import methods for the import of multichannel chromatogram files.

More information about the automated import of AB files can be found in the separate Import plugin manual. A pdf version of this manual becomes available when you click on **<Manual>** in the *Plugin installation* toolbox (Figure 1-3).

3.2.12.4 Check **Open Curve Converter (all formats)** and press **<OK>**. The **CrvConv** window opens.

3.2.12.5 Open the sample chromatogram files in the **CrvConv** window with **File > Import curves from file**. A set of example files can be found on the CD-ROM in the **Sample and Tutorial data\AB sequencer trace files** directory. Alternatively, download the data from the website and browse to the unzipped folder. Select all files.

3.2.12.6 The program may produce a warning that the curve order is not specified, and that the default setting CTAG will be used. If you want to change the colors for the curves, you can use **View > Customize colors**.

3.2.12.7 If necessary, you can change the order of the lanes with **Edit > Move curve up** and **Edit > Move curve down**, or using CTRL+Up or CTRL+Down on the keyboard.

3.2.12.8 You can remove lanes if necessary with **Edit > Remove curve** (shortcut is DEL on the keyboard).

3.2.12.9 Select **File > Export curves** to save the curve gel files. Navigate to a directory on your hard drive, enter a name (e.g. **ABIGel01**) and press **<Save>**.

3.2.12.10 The program now asks "Do you want to reverse the curves?". If you know the top of the lanes is at the end of the curves, press **<Reverse>**, otherwise, press **<Don't reverse>**.

The program adds a number 01, 02, 03, 04 (the program supports up to 8 channels per lane) to each gel, depending on the color, and adds the extension .CRV to each gel.

3.2.12.11 In the *InfoQuest FP main* window, create a new fingerprint type (3.2.1), e.g. **ABI**, specifying **Densitometric curves** when the wizard asks "What kind of fingerprint type do you have?".

3.2.12.12 Specify **12-bit** OD range (4,096 gray levels) and leave the other settings unaltered.

When finished, the new fingerprint type **ABI** is listed in the *Experiments* panel. In a next step we are going to import the curve files.

3.2.12.13 Click the  button in the *Experiment files* panel or select **File > Add new experiment file** in the *InfoQuest FP main* window.

3.2.12.14 In the *Import fingerprint file* window, select **Curve files** as type of files and navigate to the path where the curve files are stored (see 3.2.12.9). Select a curve file (e.g. **ABIGel01_xx**) and press **<Open>**.


3.2.12.15 Repeat the previous step for the second curve file.

Two files are now listed in the *Files* panel. First we are going to process the reference file and then the data file(s).


3.2.12.16 Open the reference gel with **File > Open experiment file (data)**, and assign it to fingerprint type **ABI**.

The gel with the reference patterns is the gel containing 13 bands per lane. If you have opened the data gel and not the reference gel, close the window and repeat the previous step for the other gel.

It may be necessary to adjust the brightness and contrast (**Edit > Change brightness & contrast**), by enabling the **Dynamical preview** and slowly moving down the **Maximum value** until the darkest bands are (nearly) black.


3.2.12.17 Select **Lanes > Auto search lanes** or  to let the program automatically find the lanes.


You will notice that some setting options, applicable for TIFF files, are not available here: e.g. *Gelstrip thickness* and *Number of nodes*.


3.2.12.18 Move on to the next step with . This shows the densitometric curves.

Here again, *Averaging thickness* and *Number of nodes* do not apply. You may want to adjust the background subtraction and the filtering as described in 3.2.2.

3.2.12.19 Move on to the **Normalization** step .

3.2.12.20 Locate a suitable standard, place the gel in **Normalized view** , and define the reference bands (see 3.2.5.4). The example uses the reference mix from ABI, containing 13 bands with known molecular weight.

3.2.12.21 Select **References > Use all lanes as reference lanes** to mark all 16 lanes as **Reference lane**, and align the bands with **Normalization > Auto assign bands** or press .

3.2.12.22 Update the normalization with  and save the normalized reference gel.

3.2.12.23 Select the second gel, the gel containing the data, with **File > Open experiment file (data)**, and assign it to fingerprint type ABI.

3.2.12.24 Link this gel to the reference gel with the command **File > Link to reference gel** and enter the name of the reference gel in the dialog box that pops up.

NOTE: If the reference gel is selected in the Experiment files panel, the name of the reference gel is automatically shown in the dialog box.

The tracking info, curve settings, and alignment of the reference gel are now automatically superimposed to the data gel. You can run through the different steps till you reach the normalization step: the alignments as obtained in the reference gel are shown. If you wish, you can show the normalized view before you move to the last step, i.e. defining bands.

3.2.12.25 Whenever needed, you can pop up a reference gel to which a data gel is linked with **File > Open reference gel**.

3.2.12.26 If you have made changes to the reference gel without saving them, you can update the changes to the data gel with **File > Update linked information**. Once

you save the changes to the reference gel, the data gel(s) are updated automatically.

More information on the processing of the gels can be found in paragraph 3.2.6 and Section 4.1 - 4.2.

B. RFLP gel scans containing internal markers

The way of processing the gels is similar as described for the multichannel files, except that we start from two independent TIFF files here. An absolute condition is that the two TIFF files, containing the references and the data lanes respectively, have exactly the same resolution (dpi). If the images are shifted or rotated, they can be aligned to each other by applying two or more marker points to the gel. These marker points will be visible on the TIFF files, and the software allows such markers to be used to align the images.

3.2.12.27 Open the TIFF image of the reference gel and assign it to a fingerprint type.

If the reference gel and the data gel need to be aligned to each other, you should define marker points as follows:

3.2.12.28 In the first step (**1. Strips**), select **Lanes > Add marker point** and click on the first marker point of the gel.

3.2.12.29 Repeat the same action for the other marker points.

At least two marker points should be present before the program can copy the geometry from one gel to another.

If the TIFF images are already aligned (for example, when different fluorescent markers are used in the same gel, which are visualized at the same time), you should not add marker points.

3.2.12.30 Proceed with the full normalization of the reference gel as described in 3.2.2. Save the file.

3.2.12.31 Open the data gel and assign it to the same fingerprint type.

3.2.12.32 Link this gel to the reference gel with the command **File > Link to reference gel** and enter the name of the reference gel in the dialog box that pops up.

NOTE: If the reference gel is selected in the Experiment files panel, the name of the reference gel is automatically shown in the dialog box.


The tracking info, curve settings and alignment of the reference gel are now automatically superimposed on the data gel. In the second step (**2. Curves**), it is still possible to adjust the position of the track splines individually, or to add nodes and distort the curves where necessary. You can run through the different steps till you reach the normalization step: the alignments as obtained in the reference gel are shown. If you wish, you

can show the normalized view before you move to the last step, i.e. defining bands.

3.2.12.33 Whenever needed, you can pop up the reference gel to which a data gel is linked with **File > Open reference gel**.

3.2.12.34 If you have made changes to the reference gel without saving them, you can update the changes to the data gel with **File > Update linked information**. Once you save the changes to the reference gel, the data gel(s) are updated automatically.

NOTE: It is also possible to copy the geometry and normalization from one gel to another without linking them. In the reference gel, go back to the first step (1.

*Strips) with  and select **Lanes > Copy geometry**. In the data gel, use **Lanes > Paste geometry** to copy the gelstrip definition from the reference gel. The normalization from the reference gel is copied with **References > Copy normalization** and **References > Paste normalization** in the normalization step. This approach may offer additional flexibility in special cases, but is not generally recommended.*

3.2.13 Import of molecular size tables as fingerprint type

InfoQuest FP allows the input of band size and band position tables, and reconstruct fingerprints of these, based upon the size and the amplitude (area or height) of the peaks.

3.2.13.1 In the **Example** database, create a new fingerprint type **AB-Genescan**. Leave every setting as default except in the second step, where you should specify **Densitometric curves** and **12-bit (4096 values)**.

We will now create a new reference system to allow the import of an AB Genescan table, part of which is shown in Figure 3-32. The whole file (5 patterns) can be found in the **Sample and Tutorial data\Sample text files for import** directory on the installation CD-ROM as **Genescan.txt**. This text file is also available from the download page of the website (www.bio-rad.com/softwaredownloads). There are two possible approaches to create a new reference system:

- Enter positions on the gel (running distances) and the corresponding band sizes. Based upon the positions and the corresponding sizes, the program is able to establish a regression curve, upon which all imported bands can be mapped. This option is particularly suitable when you know the exact positions of the size markers in a gel system, and you want to reproduce the real regression exactly.
- Allow the program to create its own regression curve between a defined maximum and minimum

molecular weight, so that it can map the imported bands on this synthetic regression curve. This method is useful if you want to import band tables of which you know nothing else than the sizes.

We will focus on the example Genescan file **Genescan.txt** to apply both methods. The file format contains a column with the sample number, a comma and then the band number (Figure 3-32), next is a column with the running time, next is the size in base pairs, then the height, the volume, and the running time again.

Option 1: Composing a regression curve by entering positions and sizes.

The running distance needed by InfoQuest FP is reciprocal to the running time given in the Genescan file (second column). Therefore, we will calculate the reciprocal value of the running time, keeping in mind that this value should never exceed 100%. Thus in order to calculate a running distance of a band (RD), we look for the *lowest* running time (RT_{min}) in the file (highest running distance), divide this number by the actual running time (RT) of that band and multiply by 100 to have it in percent:

$$\%RD = RT_{\min}/RT \times 100$$

In the example, RT_{min} = 30. For the reference lane 17B, this yields the extra column under the running time column (Figure 3-32).

Based upon this running distance in percent and the band sizes, we can create a realistic regression curve according to the first approach described above.

Sample & band no.	Running time	Size in bp	Height	Volume	
17B, 1	33.00	60.47	228	929	330
17B, 2	34.60	67.53	201	815	346
17B, 3	43.30	106.02	113	855	433
17B, 4	52.90	146.14	381	1908	529
17B, 5	88.20	298.95	131	690	882
17B, 6	89.00	302.68	1425	7821	890
17B, 7	155.40	709.46	304	1800	1554
17B, 8	158.50	736.00	182	966	1585
17B, 9	165.10	796.02	121	713	1651



90.9
86.7
69.3
56.7
34.0
33.7
19.3
18.9
18.2

Figure 3-32. Lane in an AB Genescan table and conversion of running distances to InfoQuest FP positions.

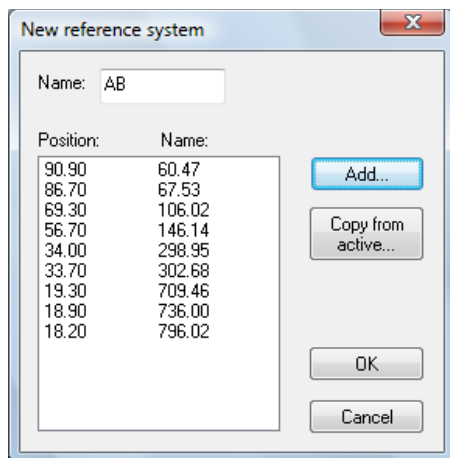


Figure 3-33. Defining a new reference system based upon known band positions and sizes.

3.2.13.2 In the *Fingerprint type* window, select **Settings > New reference system (positions)**.

The input box shown in Figure 3-33 allows all known reference bands to be entered.

3.2.13.3 Press the **<Add>** button and enter all running distances and sizes of lane 17B, as shown in Figure 3-33.

3.2.13.4 Enter a name for the reference system, e.g. **AB**.

3.2.13.5 When finished, press **<OK>**.

NOTE: Once a new reference system is defined, it is not possible to change it anymore! If you want to change a self-made reference system once it is saved, you will have to delete it and create it again.

3.2.13.6 Make the new reference system the *active reference system* by selecting it and **Settings > Set as active reference system** (not necessary if the reference system is the only one available).


3.2.13.7 Select **Settings > Edit reference system** or double-click to define the molecular weight regression.

3.2.13.8 In the *Reference system* window, copy the entered molecular weights with **Metrics > Copy markers from reference system**.

InfoQuest FP is now configured to import the Genescan tables.

3.2.13.9 Exit the *Reference system* window and the *Fingerprint type* window.

To import Applied BioSystems Genescan files, there are scripts available on the website of Bio-Rad. These scripts can be launched from the *InfoQuest FP main window*,

using the menu **Scripts > Browse Internet**, or 

The script to import Genescan data can be found under **Import tools** and is called **Import ABI Genescan tables**.

A description of how to use this script is available on the website.

3.2.13.10 When running the script, you can use the example **Genescan.txt** file in the **Sample and Tutorial data\Sample text files for import** directory on the CD-ROM.

Option 2: Importing band sizes by using a synthetic regression curve.

As an exercise, we will now import the same file using the second option described above, i.e. allowing the program to create its own regression curve.

3.2.13.11 In the *InfoQuest FP main window*, open the *Fingerprint type* window for **AB-Genescan**.

3.2.13.12 In the *AB-Genescan Fingerprint type* window, select **Settings > New reference system (curve)**.

The *New reference system* window (Figure 3-34) allows the size range to be specified as well as the type and strength of the regression.

3.2.13.13 Under **Metrics range of fingerprint**, enter 1000 as **Top** and 30 as **Bottom**.

3.2.13.14 Press the **<Add>** button to add the sizes for all reference bands available in the fingerprint type (see lane 17B, Figure 3-32).

The reference bands are shown as red dots on the regression curve. This makes the adjustment of the **Calibration curve** easier.

3.2.13.15 Optimize the **Calibration curve** and the strength (in percent) to obtain the best spread of the reference bands.

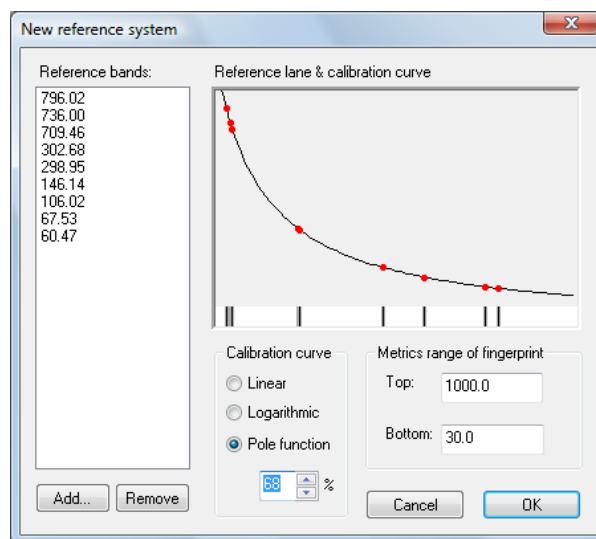


Figure 3-34. Defining a new reference system using a synthetic regression curve between user-defined size limits.

3.2.13.16 When finished, press <OK> to save the new reference system.

3.2.13.17 Make the new reference system the *active reference system* by selecting it and *Settings > Set as active reference system*.

Now you can import the same band table as described in 3.2.13.9 and further. When creating the database file, you should change the name **Genescan.txt** into another name, for example **Genescan2.txt**, because the program does not allow existing database files to be overwritten.

The two differently imported band size tables are an excellent example to illustrate the *remapping* functions in InfoQuest FP. Both gels have their bands on different positions because of the different logarithmic function that was used to reconstruct the gels.

3.2.13.18 Select an entry of the first imported file (should be **G@Example@Genescan@001** or similar if you used other names).

3.2.13.19 Select the corresponding entry of the second imported file (should be **G@Example@Genescan2@001** or similar if you used other names).

3.2.13.20 Create a comparison containing these entries and *Layout > Show image*. The patterns look the same except for very minor differences due to inevitable error caused by remapping.

3.2.14 Conversion of gel patterns from GelCompar versions 4.1 and 4.2

The installation CD-ROM contains a directory **GEXPORT**, in which the following two files are found: **BNexport.exe** and **BNexport.hlp**.

The program **BNexport.exe** and its help file **BNexport.hlp** should be copied to the home directory of GelCompar 4.1 or GelCompar 4.2.

The file **BNexport.hlp** is a Windows help file which explains step by step how to proceed to convert patterns from GelCompar to InfoQuest FP.

3.2.15 Dealing with multiple reference systems within the same fingerprint type

Under normal circumstances, a reference system is created once initially, and is never changed afterwards. In some cases however, it can be required that a second reference system is created. Some examples are:

(1) The gel used originally for defining the reference positions appears to be an aberrant one, so that repositioning the reference positions is required to allow most other gels to be normalized easily.

(2) One or more bands defined as reference positions are found to be unreliable or inappropriate and should be deleted or replaced with another band.

(3) The user switches to a new reference pattern for the fingerprint type.

(4) Gels of the same fingerprint type are imported from another database and need to be analyzed together with gels from the local database.

Case (1), shown in Figure 3-35, results in two reference systems with the same reference position names, but having different % distances on the gel. Gels processed under both reference systems are perfectly compatible and there is no loss of accuracy compared to gels analyzed under the same reference system.

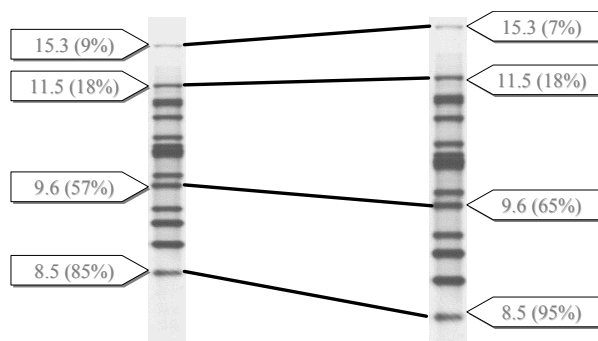


Figure 3-35. Example of different reference systems in the same fingerprint type for which remapping causes no loss of accuracy. See text for explanation.

The same situation can arise if gels are imported from another database, which have been processed under a different reference system [case (4)], but where the same marker pattern is used and the reference positions have been given the same name (even though the % distances are different).

Case (2) may result in a new reference system with more or less bands, or with bands having a different name (Figure 3-36). In either case, the new reference system will not be automatically compatible with the original, and compatibility can only be obtained by creating a molecular weight regression curve for both reference systems (see 3.2.9.12 to 3.2.9.16 on how to create a regression curve). Both reference systems can then be remapped onto each other, which inevitably causes some loss in accuracy. The degree of compatibility depends on the number of reference positions in both systems, the amount of overlap between regression curves, the predictability of the regression curve using one of the available methods, the spread of calibration points (reference positions), the definition of the reference bands, etc.

Case (3) obviously causes a situation where reference positions have different names, since one can assume

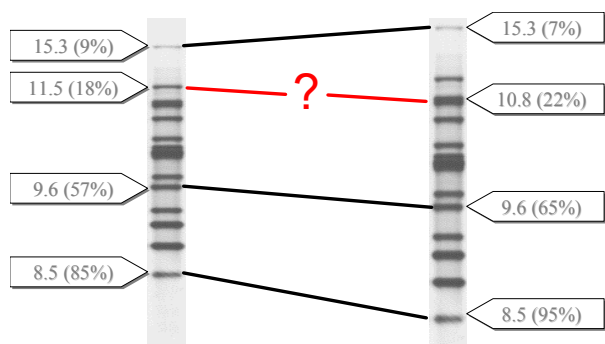


Figure 3-36. Example of different reference systems in the same fingerprint type for which remapping relies on molecular weight regression curves for both reference systems and as such, causes some loss of accuracy. See text for explanation.

that a new marker has different bands, and results in a situation where remapping is required.

When more than one reference system is present in a fingerprint type, one of the reference systems is specified as the “active” reference system. The active reference system is the one to which all new gels will be normalized. By default, the first created reference system is the active one. The name of the active refer-

ence system is shown in red in the *Fingerprint type* window.

3.2.15.1 To change the active reference system, open the *Fingerprint type* window, and select the reference system to become the active one. Choose *Settings > Set as active reference system*.


3.2.15.2 To remove a reference system that is not used anymore, select the reference system in the *Fingerprint type* window, and choose *Settings > Remove reference system*.

The program asks “Do you want to check if this reference system is in use?”. For large connected databases, this may take a long time. If you answer <No> to this question, the selected reference system is removed, regardless of whether it is used in gels or not. By opening and saving a gel that was processed under the removed reference system however, it will be restored. By answering <Yes>, the program checks the database for gels normalized with the reference system, and if any such gels are found, the reference system is not removed.

NOTE: To avoid any possible conflict situations, it is recommended to allow the program to scan the database for the presence of gels normalized with the reference system, and not to remove any reference systems that are in use.

3.3 Setting up character type experiments CH

3.3.1 Defining a new character type

3.3.1.1 Select *Experiments > Create new character type* from the main menu, or press the  button in the *Experiments* panel toolbar and select *New character type*.

3.3.1.2 The *New character type* wizard prompts you to enter a name for the new type. Enter a name, for example **Pheno**.

3.3.1.3 Press **<Next>** and check the kind of the character data files. Check *Numerical values* if the tests are not just positive or negative but can differ in intensity (choose *Numerical values* in this example).

3.3.1.4 For numerical values, enter the number of decimal digits you want to use. If you only want to use integer values, for example between 0 and 10, enter zero (this example).

3.3.1.5 After pressing **<Next>** again, the wizard asks if the character type has an open or closed character set.

In an *open character set*, the number of characters is not defined. For example, studying 10 bacterial strains by means of fatty acids can result in a total of 20 fatty acids found, but if some more strains are added, more fatty acids may become present in the list. In such cases, *Consider absent values as zero* should be checked, because if a fatty acid is not found in a strain it will not be listed in its fatty acid profile, and thus should be considered as zero.

In a *closed character set*, the same number of characters are present for all entries studied. This is the case with commercially available test kits. In such cases, *Consider absent values as zero* should not usually be checked.

3.3.1.6 Answer *No* to the *open character set* and leave the *absent values* check box unchecked.

If the character set is *closed*, i.e. when all the tests are predefined, the user is allowed to specify the *Layout* of the test panel. This layout involves a *Number of rows* and *Number of columns* to be specified, as well as the Maximum value for all the tests. By default, the number of rows and columns is set to zero, which means that the character set will be empty initially. In this case, you still can add all the tests one by one or by columns and rows, once the character type is defined. If you are defining a test panel based upon a microplate system (96 wells), you can now enter 8 as *Number of rows* and 12 as *Number of columns*. The program will automatically assign names to the tests: A1, A2, A3, ..., A12, B1, B2,


B3, ... These names can be changed into the real test or substrate names afterwards.

3.3.1.7 Press the **<Finish>** button to complete the setup of the new character type.

The new character type is now listed as a character type in the *Experiments* panel.

3.3.2 Editing a character type

The new character type exists by now, but the program still doesn't know which, and how many tests it contains. We will further define its tests.

3.3.2.1 Double-click on **Pheno** in the *Experiments* panel or select **Pheno** and *Experiments > Edit experiment type* or press  in the *Experiments* panel toolbar.

The *Character type* window appears (Figure 3-38), initially with an empty character list. Suppose that the phenotypic character kit exists of 10 tests; we will enter them one by one.

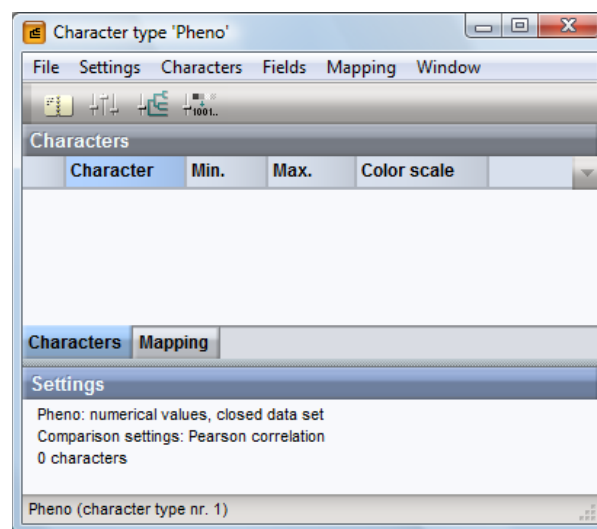


Figure 3-38. *Character type* window; no characters defined yet.

3.3.2.2 Select *Characters > Add new character*. Enter any name, e.g. **Glucosidase** and press **<OK>**. Enter the maximal value the character can reach (default 100) and press **<OK>**.

The character is now listed in the *Characters* panel (Figure 3-38).

3.3.2.3 Select *Characters > Rename character* if you want to give a character a different name. Select *Characters > Remove character* to delete the selected character from the list.

The default color scale for each character ranges from white (negative) to black (most positive). Its default intensity range is 0 to 100. If you want the character to cover another range, proceed as follows:

3.3.2.4 Select *Characters > Change character range*. Enter a minimum and maximum value and press **<OK>**.

3.3.2.5 If you want to change the range of all characters to the same range, select *Characters > Change all character ranges*.

NOTE: You can quickly access all menu commands that apply to a character by right-clicking on the character.

3.3.2.6 To change the color for a character, select *Characters > Change character color scale*. This pops up the *Edit color scale* dialog box (see Figure 3-39).

3.3.2.7 Select the start color (negative reaction) on the left hand side of the color scale: it is now marked with a black triangle (see arrow on Figure 3-39).

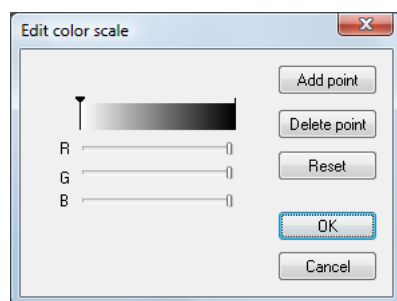


Figure 3-39. *Edit color scale* dialog box.

The three slide bars represent red, green, and blue, respectively. By moving the sliders to the right, the color becomes brighter.

3.3.2.8 Adjust the red, green and blue components individually until you have obtained the desired color for a negative reaction of this character.

3.3.2.9 Select the end color (positive reaction) on the right-hand side of the color scale and adjust the red, green and blue components individually until you have obtained the desired color for a positive reaction of this character.

3.3.2.10 If you want more transition colors, use the **<Add point>** button. A new color mark appears in the middle.

You can select this color, drag it to the left or to the right, and adjust it as described above.

3.3.2.11 Press **<Copy to character>** to copy the created color scale to the character.

3.3.2.12 Repeat action 3.3.2.2 to add more characters, and adjust the colors and the ranges individually as needed. When all characters have the same color range, you can use **<Copy to all characters>**.

A quick method to add a complete array of characters at a time, for example a microplate array, is *Characters > Add array of characters*. The program subsequently asks to enter the number of rows, the number of columns, and the maximum values for the tests. The program automatically assigns names to the tests: A1, A2, A3, ..., B1, B2, B3, ... These names can be changed into the real test or substrate names afterwards.

Each character is marked with a green \surd sign, which means that it is used in comparisons and identifications.

3.3.2.13 If you want a character to be disabled (not used) in comparisons, uncheck the *Characters > Use character for comparisons* menu item. This may be useful for a blank test which is often present in commercial identification kits. Unused characters are marked with a red cross.

3.3.2.14 In a character list, individual characters can be moved up or down. To achieve this, select the character and *Characters > Move character up* (CTRL+Up arrow key) or *Characters > Move character down* (CTRL+Down arrow key).

NOTE: In a local database, the software needs to be closed and opened again before the new arrangement of characters becomes effective in a Comparison window.

3.3.2.15 Using the menu *Settings > General settings* or



the character type settings which were entered in the setup wizard can be changed in the *Character type* tabs. The *Experiment card* tab lets you define some visual attributes of the experiment. These settings apply to the *Experiment card*, which is explained in Section 3.8.


With *Represent as Plate* and *Represent as List*, you can choose whether the individual tests are shown graphically on a panel, using colors, or as a list of characters with their name and intensities as a numerical value. In case of an *open character set*, only the list type can be chosen.

3.3.2.16 For the example here, choose *Plate*, and enter the number of columns in the test panel (if you entered 10 tests previously, enter 10 numbers of columns as well).

For test kits on microtiter plates, one would enter 96 tests and 12 columns.

In order to represent existing commercial kits as truthfully as possible, you can choose between three different circular cup types and elliptical cups. For blots and microarrays, you can choose between small blot, large microarray spots and small microarray spots.

3.3.2.17 Select the type of cells in the *Cell type* pull-down menu, e.g. elliptical, and press <OK>.

3.3.2.18 With the menu command *Settings > Binary conversion settings*, or , you can specify a binary cutoff value in percent.

Whenever converting the numbers to binary states, InfoQuest FP will consider all values above the cutoff value as positive and those below the cutoff value as negative. If you have entered 50% as cutoff value, you can choose the cutoff level to be 50% of the maximum value found in the experiment, or 50% of the average value from the experiment.

3.3.2.19 Enter 50% and *Of mean value*.

In InfoQuest FP, character values can be mapped to categorical names according to predefined criteria.

3.3.2.20 Select the *Mapping* tab in the *Character type* window (see Figure 3-40).

3.3.2.21 To add a criterion, select *Mapping > Add new mapping*.

3.3.2.22 In the dialog that pops up, give a name (e.g. Resistant) and enter a range (e.g. 0 and 1).

3.3.2.23 Repeat the previous steps if you want to add more criteria (e.g. Intermediate, Susceptible).

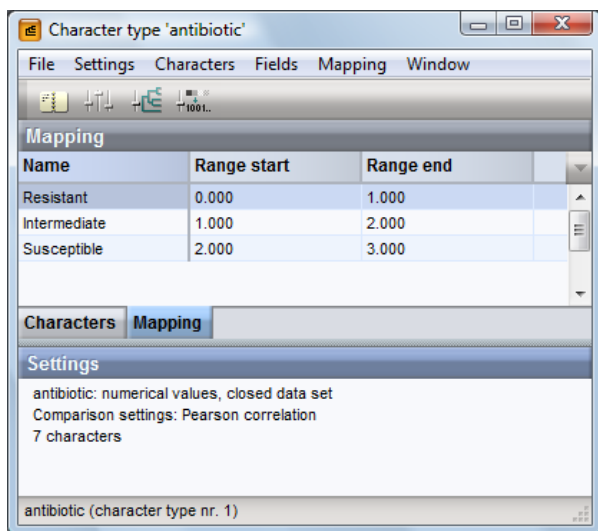


Figure 3-40. The *Mapping* tab within the *Character type* window, showing three predefined criteria.

Character	Value	Mapping
AMK	1.200	Intermediate
AMOX/CA	0.300	Resistant
AMPC	2.300	Susceptible
AMPVSU	0.900	Resistant
AZI	2.100	Susceptible
CCLO	0.600	Resistant
CFAZ	1.200	Intermediate

Figure 3-41. The *Mapping* column in the experiment card of a character type.

3.3.2.24 Close the *Character type* window.

If you open the experiment card of a character type, a *Mapping* column is listed (see Figure 3-41). If character values fall within the ranges of defined criteria for this character type, the name of the mapping is displayed in the *Mapping* column (see Figure 3-41). If the character value does not fall within the range of the defined criteria, or if no criteria are defined, a '<?>' is displayed.

The mapped names can be displayed in the *Comparison* window (see Figure 4-1), the *Pairwise comparison* window (see Figure 4-2) or in reports.

NOTE: The mapping of character values does not have any influence on the choice of coefficient or on the clustering of the character data since the clustering is based on the original values (see Section 4.3).

3.3.3 Input of character data

There are four possibilities for entering data of a character type:

1. Importing a character file or importing characters from text files or from an ODBC source using the **Import** plugin.
2. Defining a new character file in InfoQuest FP and entering the values manually.
3. Entering the data via the experiment card of the database entry (see 3.8.3).
4. Processing and quantification of images scanned as TIFF files (see 3.3.5).

• **Importing characters from text files or ODBC sources**


For the import of characters from text files or ODBC sources (databases or spreadsheets; e.g. Excel files, Access files, etc.), an **Import plugin** is available.

More information on the installation of the Import plugin can be found in paragraph . A full description on how to use this plugin can be found in the separate Import plugin manual. A pdf version of this manual becomes available when you click on **<Manual>** in the *Plugin installation* toolbox (Figure 1-13).

• **Defining a new character file**

A new character file is created as follows:


3.3.3.1 In database **Example**, select the new character type **Pheno**.

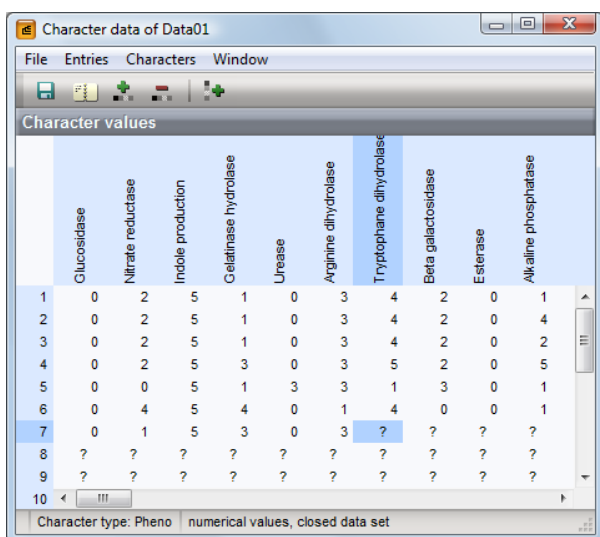
3.3.3.2 In the *InfoQuest FP main window*, select **File > Add new experiment file** or click on  in the *Files* panel.

NOTE: This feature is only accessible when working in a local database. Entering the data via the experiment card of the database entry works both on local and connected databases (see 3.8.3).

3.3.3.3 Enter a name, e.g. **Data01** and press **<OK>**.

3.3.3.4 Select **Data01** in the *Files* panel, and **File > Open experiment file (data)**.


This opens the *Character data file window* (Figure 3-42), which is empty initially. The test names, which we entered as an example, are shown in the column header. You can add characters here with **Characters > Add new character** or .



	Glucosidase	Nitrate reductase	Indole production	Cellulase hydrolase	Urease	Arginine dihydrolase	Tryptophane dihydrolase	Beta galactosidase	Esterase	Alkaline phosphatase
1	0	2	5	1	0	3	4	2	0	1
2	0	2	5	1	0	3	4	2	0	4
3	0	2	5	1	0	3	4	2	0	2
4	0	2	5	3	0	3	5	2	0	5
5	0	0	5	1	3	3	1	3	0	1
6	0	4	5	4	0	1	4	0	0	1
7	0	1	5	3	0	3	?	?	?	?
8	?	?	?	?	?	?	?	?	?	?
9	?	?	?	?	?	?	?	?	?	?
10	?	?	?	?	?	?	?	?	?	?

Figure 3-42. Character data file window.

Before you can enter data, you have to add new entries to the file. Suppose that we want to add character data for all entries of the database except the standard (17).

3.3.3.5 Select **Entries > Add new entries** or .


3.3.3.6 You are prompted to enter the number of entries; enter 17, and press **<OK>**.

Seventeen entries are now present, and all character values are initially represented by a question mark (Figure 3-42).

3.3.3.7 To enter values, you can either double-click on the question mark, or press the Enter key.

3.3.3.8 Enter values between 0 and 5 (the range of the characters), and press Enter again.



The next character of the same entry is automatically selected, so that you can directly enter the next value.


3.3.3.9 If you are entering large character files, we recommend to save now and then with **File > Save** () or the F2 shortcut.



3.3.3.10 **File > Exit** when you are finished entering the data.

3.3.3.11 In the *InfoQuest FP main window*, double-click on the file **Data01** or select **File > Open experiment file (entries)** with the file **Data01** selected.

The *Character entry file window* (cf. the *Fingerprint entry file window*, Figure 3-27) contains unlinked entries, which you can now link to the corresponding database entries.

A *link arrow*  for each entry allows you to link an entry to a database entry, by clicking on the arrow and dragging it onto a database entry, and then releasing the mouse button. When the experiment is linked, its link arrow is purple: .

3.3.3.12 Drag the link arrow of **entry 1** to the first database entry: as soon as you pass over a database entry, the cursor shape changes into .

3.3.3.13 Release the mouse button on the database entry; entry is now linked to this database entry, and its arrow in the *Character entry file window* has become purple  instead of gray .

NOTE: If you try to link an entry to a database entry which already has an entry of the same experiment type linked to it, the program will refuse the second link with the message: "The Experiment 'Pheno' of this database is already defined in XXX", where XXX is another lane

of the same character type, in the same or another experiment file.

As soon as an experiment is linked to a database entry, the *Experiment presence* panel (see Figure 1-15) shows a colored dot for the experiment of this entry.

You can edit the information fields for this entry in two places: directly in the database (see 2.2.3.1 to 2.2.3.2), or in the *Character entry file* window, by double-clicking on the entry.

NOTE: Experiment files added to the Experiment files panel can also be deleted by selecting the file and choosing File > Delete experiment file from the

main menu or clicking on  in the Files panel.

Deleted experiment files are struck through (red line) but are not actually deleted until you exit the program. So long, you can undo the deletion of the file by selecting File > Delete experiment file or by clicking

 again.

3.3.4 Character type settings restricted to connected databases

When a default connected database is defined for the current database (Section 2.3), it is possible to define additional information fields for the characters. In the *Character type* window, additional information fields can be added, renamed or removed with *Fields > Add new field*, *Fields > Rename field*, and *Field > Remove field*, respectively. Information can be entered for a given character by clicking twice on a field, or by clicking on a field and selecting *Fields > Set field content*.

By default, the Character field is displayed in a comparison. You can choose to display another field by clicking on the header of the desired field and selecting *Fields > Use as default field*. The column used as default field is highlighted in a pale green color (see Figure 3-43).

An important feature associated with a *connected database* (Section 2.3), is the possibility to define more than one quantification per character. Each quantification can be seen as a layer corresponding to one character table (see Figure 3-44). The standard analysis tools of InfoQuest FP (such as cluster analysis, identification, statistics) can only work with one layer at a time, so that one has to specify a default quantification to use. Using scripts, however, one can access the information of several layers simultaneously.

The possibility to create different layers in a character type is particularly useful if InfoQuest FP is used as a database platform in combination with GeneMaths XT. However, it can also be useful to store and retrieve different types of quantification, or to add error values or standard deviations to character measurements.

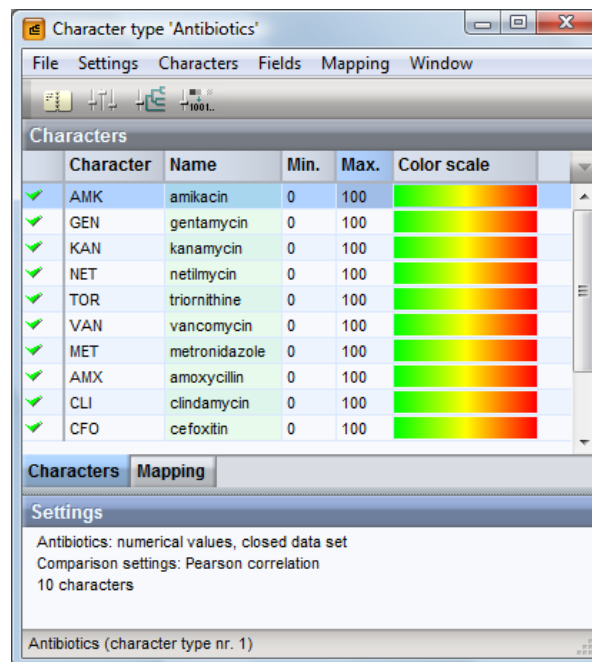


Figure 3-43. *Character type* window with the information field Name set as default field.

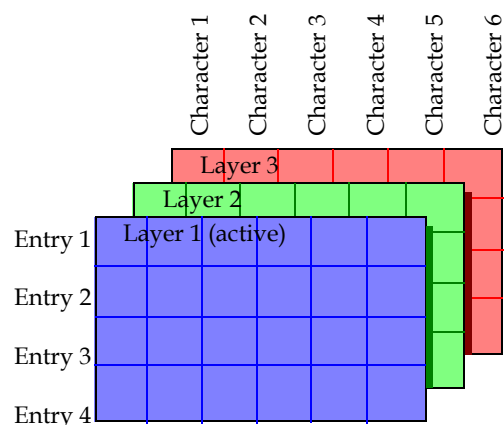


Figure 3-44. Schematic representation of different quantification layers in a character type.

3.3.4.1 To add or edit quantification layers to a character type, open the *Character type* window and select *Characters > Character quantifications*.

One default quantification layer, VALUE, is always present (Figure 3-45).

3.3.4.2 To add a new quantification layer, press the **<Add>** button.

3.3.4.3 Enter a name, for example “**Standard Deviation**”, and press **<OK>**.

The new layer appears in the list of quantification layers.

3.3.4.4 The selected layer at the time the dialog box is closed using the **<Exit>** button is the active quantifica-

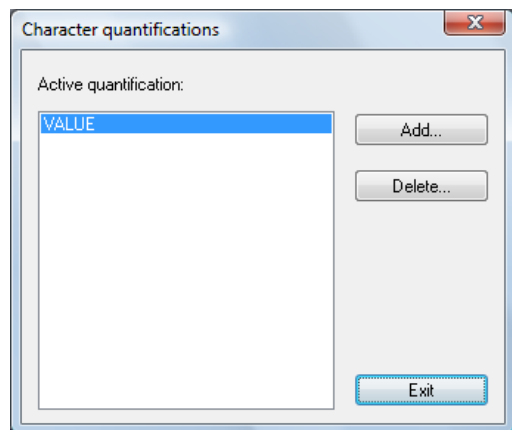


Figure 3-45. Character quantification layers editing dialog box.

tion layer used by the program. New data entered will also be stored in the currently active quantification layer.

3.3.5 Import of character data by quantification of images scanned as TIFF files

Similar as for gel images, InfoQuest FP can import *character type data* from TIFF images. This happens by quantification of the color intensity and/or color transitions on the TIFF file. Character data from phenotypic test panels often provide color transitions rather than changes in intensity. For example, many test panels have reactions that change from yellow to red, or blue - green - yellow, and hence, quantifying the colors by their intensity would provide no meaningful information. Rather, the program needs to be able to read the files as true RGB images and allow the possibility to define negative colors and positive colors, as well as transition colors. For example, in an acidification reaction with a bromophenol blue dye, non-reactive tests will be blue, whereas weak reactions will show the transition color green, and strongly positive reactions will show up yellow.

Using the same tool, InfoQuest FP also allows the import of micro-array and gene chip images scanned as TIFF files, offering for each gene or oligo a quantitative reaction value.

The character import tool is provided as a separate program, BNIMA.EXE, that can be started from within InfoQuest FP. BNIMA only works when the InfoQuest FP analyze program is running.

Example 1: import of microtiter plate image.

The first example we will use to illustrate the program is **Plate1.tif**. This TIFF file is available in the **Sample and Tutorial data\Microplate image** directory on the installation CD-ROM or from the download page of the

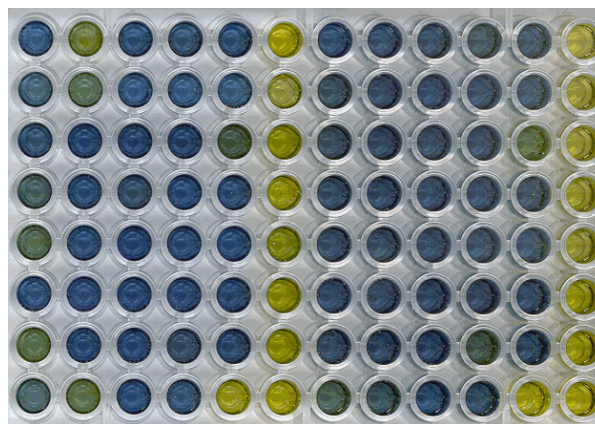


Figure 3-46. 96 wells microtiter plate with bromophenol blue as reaction indicator.

website (www.bio-rad.com). It is a photograph of a 96 wells microtiter plate with bromophenol blue as reaction indicator dye (see Figure 3-46).

3.3.5.1 Create a new *closed* character type as described in 3.3.1.1 to 3.3.1.6, and call it **Microplate**. Specify 8 rows and 12 columns under *Layout* (third step). Specify a color range from blue to yellow, over green (see 3.3.2.6).

3.3.5.2 Double-click on an entry to show its *Entry edit* window.

The experiment type **Microplate** shows an empty flask.

3.3.5.3 Click on the flask button. Since this experiment is not defined for the selected entry, the program asks "Do you want to create a new one?".


3.3.5.4 Answer **<Yes>** to create an *Experiment card* (see further, 3.8.1). An empty microplate image pops up.

3.3.5.5 Right-click on the empty microplate image and select **Edit image** from the floating menu.

This loads the BNIMA program.

3.3.5.6 Select **File > Load image** in BNIMA and load the file **Plate1.tif** from the **Sample and Tutorial data\Microplate image** directory on the installation CD-ROM or from the downloaded and unzipped folder from the website. The resulting window looks as in Figure 3-47.

3.3.5.7 First call the *Settings* dialog box with **Edit >**

Settings or .

The *Image* tab offers two choices for the *Image type*: **Densitometric** and **Color scale**.

In case the color reaction can be interpreted as a simple change in intensity (e.g. from light to dark), one should select **Densitometric**. The *Densitometric values* panel offers some additional tools to edit the TIFF file: **Inverted values** is to invert the densitometric values;

Background subtraction allows a two-dimensional subtraction of the background from the TIFF file, using the rolling ball principle. The **Ball size** can be entered in pixels. Spot removal allows all spots and irregularities below a certain size to be removed from the image, whereas larger structures are preserved. The background subtraction and spot removal changes are only seen when **Edit > Show value scale** is enabled in the *InfoQuest FP main window*.

In case the reaction causes a change from one color to another color, as in the above example, **Color scale** is the right option. An additional feature, **Hue only**, is particularly useful when the scanned images differ in brightness (illumination) or contrast. If the images do not contain black or white in their color range, it is better to enable this feature.

3.3.5.8 Select **Color scale** and **Hue only**.

3.3.5.9 Press **<OK>** to proceed with these settings. The other settings will be discussed later.

Like the process of normalization of gels, processing a character panel image exists of a number of steps: (1) Grid definition; (2) Cell layout; and (3) Quantification.

In **Step 1: Grid definition** we will create a grid that defines the wells of the microplate.

3.3.5.10 Select **Grid > Add new** and enter 8 as **Number of rows** and 12 as **Number of columns**.

3.3.5.11 Press **<OK>** and the grid appears.

At each edge of the grid, there is a dragging node (green square). The upper left dragging node is to *move* the grid as a whole; the lower right node is to *resize* the grid, and the upper right and lower left nodes are to *distort* the grid in case the image is not perfectly rectangular or not scanned horizontally.

3.3.5.12 Drag the nodes until the grid matches with all 96 wells.

NOTE: Using the SHIFT key, one can distort the grid locally if needed. The size of the local distortion area is indicated by a circle. It is even possible to reduce or enlarge the size of the distortion area as follows: hold the SHIFT key and left-click on any cell-marking cross of the grid. The area defining circle appears. Hold the left mouse button down and release the SHIFT key: the circle is still visible. While holding the left mouse button down, press the PgDn or PgUp key to reduce or enlarge the area of distortion. The size of the circle will

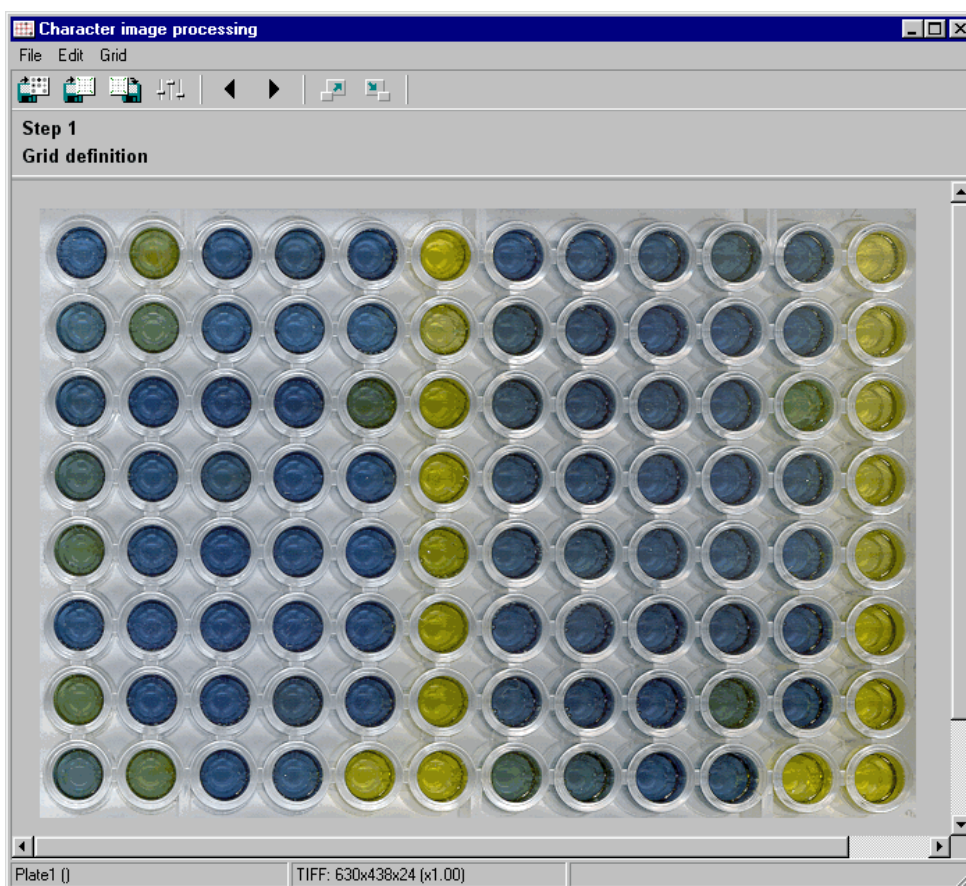



Figure 3-47. The BNIMA program with a microplate image loaded.

decrease or increase. Using a very small circle, it is possible to correct the grid in any individual cell.

3.3.5.13 In case you want to remove the grid and define it again, select one of the cells of the grid (the grid becomes red) and **Grid > Delete**.

In case the image consists of two or more subsets of cells (e.g. some more complex test panels or micro-arrays), it is possible to define more than one grid using the **Grid > Add new** command.

3.3.5.14 Move to the next step using **Edit > Next step** or the  button.

In this step, the layout of the cells is defined: the shape and size of the quantification area within each cell. In this step, we also define which cells we want to use for quantification and which cells not. By default, all cells of the grid are used for quantification.

3.3.5.15 Click in the upper left corner of the image and while holding the left mouse button down, select the left half of the test panel.

All the selected cells are marked in red.

3.3.5.16 Select **Cells > Delete selected**. The left half of the panel will not be used for quantification, and hence, cannot be used in the resulting character set.

Cross marks of unused cells are smaller than of used cells.

3.3.5.17 Select the non-used cells again with **Cells > Add selected**.

Before the program can do the quantification, it needs to know what the averaging area of the cells is. This is done using a *mask* which the user defines. One can define the same mask for all cells, or assign particular masks to individual (groups of) cells. In the case of a microplate it is obvious that all cells should have the same mask.

3.3.5.18 Select all cells as in 3.3.5.15.


3.3.5.19 Add a circular mask to all selected cells with **Cells > Add disk to mask**.

A dialog box prompts to enter a **Radius** for the disk in pixels, the **X offset** (horizontal shift from the cell marking cross) and the **Y offset** (vertical shift from the cell marking cross). For the offsets, a negative value can be entered.

3.3.5.20 Enter 8 as radius, and leave the offsets zero. Press **<OK>** to confirm.

The masks appear on all used cells of the grid as semi-transparent red disks. In order to see the masks, it is important that they are in a color that is complementary

to the reaction colors of the cells. One can change the color of the masks as follows:

3.3.5.21 Select **Edit > Settings** or .


3.3.5.22 Under **Layout**, pull down the **Mask color** menu and select the appropriate color.


In the example microplate, the most appropriate color is red.


*NOTE: In case of very small cells, e.g. microplate images, you can select **Small cross marks**, so that they don't overlap most of the cells.*

It is possible to add more than one mask to the cells. In case the cells have a more complex layout, i.e. not just circular, one can add two or more disks with different offsets to approach the shape of the cells. By selecting individual cells or groups of cells, it is also possible to change the shape of the masks per cell or per group of cells.

*NOTE: Some more advanced features allow the mask of individual cells to be changed manually: With **Cells >***


Add pixels to mask or  the cursor changes into a pencil which you can use to add pixels to the masks manually. When doing so, it is recommended to zoom in on the cell using the **Edit > Zoom in** command or

. Similarly, it is possible to remove pixels from the mask with **Cells > Remove pixels from mask** or

. Clicking a second time on these buttons or selecting the menu item finishes the pixel editing mode.

If you selected the image type to be **Color scale**, and not **Densitometric** in the settings (see 3.3.5.7 to 3.3.5.8), you can now specify the negative color, the positive color, and any transition colors between negative and positive. For each cell, you can define a unique color scale, which can be necessary for some commercial test panels containing more than one reaction dye.

3.3.5.23 In the example case, there is only one reaction dye, so select all cells as in 3.3.5.15.

3.3.5.24 Select **Cells > Edit color scale** or . This brings up the **Color scale** editor as shown in Figure 3-48.

By default, the color scale exists of two colors: white as negative and black as positive. In the case of the example microtiter plate, this scale would obviously not work. Since the scale ranges from blueish (negative) over greenish to yellow (positive), we will add a new intermediate color.

3.3.5.25 Press the **Add color** button. One new color (gray) is defined in the middle of the scale.

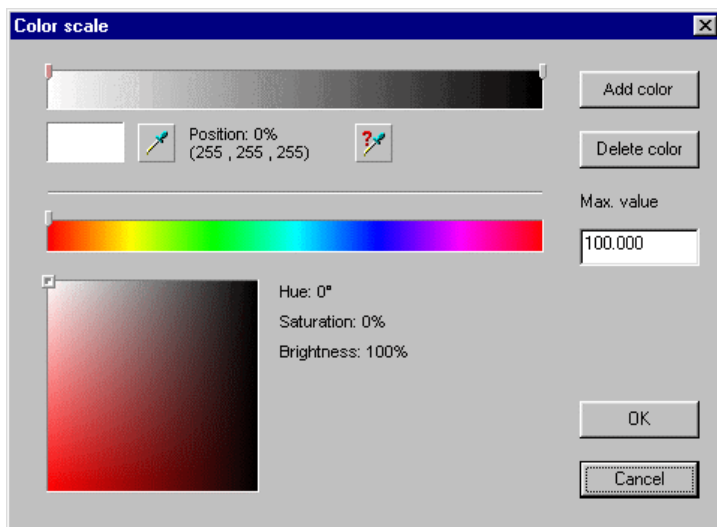


Figure 3-48. Color scale editor in the Cell layout step of the BNIMAGE program.

3.3.5.26 Select the color selector of the negative color (left).

3.3.5.27 Move the slider on the color scale of predefined colors to blue.

3.3.5.28 Move the slider in the Saturation/Brightness square to the lower left corner to obtain maximum brightness and saturation.


3.3.5.29 Repeat 3.3.5.26 to 3.3.5.28 for the intermediate color (middle), assigning green, and for the positive color (right), assigning yellow.

*NOTE: If you selected **Hue only** in the settings (see 3.3.5.7 to 3.3.5.8), changing saturation and brightness has no effect on the obtained color scale. If saturation or brightness transitions within the same color are to be registered, you should disable the **Hue only** feature in the settings.*


The upper color scale now should range from blue over green to yellow (Figure 3-49).



Figure 3-49. Appropriate color scale for the example microplate image.


NOTE: One can also pick up colors from the image in order to define the selected color in the upper color scale. To this end, click and hold the left mouse button on the left pipet button . The mouse pointer shape changes into a pipet which you can drag to the most negative cell, e.g. the blank control. The selected color in the color scale automatically changes into the color at the pipet's

*position. If **Hue only** is enabled, the closest hue color is selected.*

Once the color scale is defined you can interrogate the reaction of any cell using the right pipet button  as described above. This pipet does not affect the defined color scale, but only shows the position of the pointed cell graphically on the color scale, and the percentage reaction with error indication.

With the *Max. value* field you can enter the maximum value to which all characters will be rescaled.

3.3.5.30 Enter 100 as *Max. value* and press <OK> to confirm the color settings.

3.3.5.31 Move to the next step using *Edit > Next step* or the  button.

The next and last step involves quantification of the cells. First of all, the cells to be added to the character set need to be defined. In case one or more cells are intended only for calibration purposes, they can be excluded from the resulting character set, but used as calibration marker.


3.3.5.32 Select all cells by clicking in the upper left corner of the image and while holding the left mouse button down, select the complete test panel.

3.3.5.33 Select *Quantification > Add cells to character set*.

The cells are now numbered from 1 to 96.

3.3.5.34 If you click on a particular cell, its quantified value as rescaled according to 3.3.5.30 is given in the status bar as well as the value after calibration (see further).

Quantification is done by integrating the pixels within the defined mask. There are different options for integration:

3.3.5.35 Select *Edit > Settings* or  and choose the *Quantification* tab.

Cell integration methods include *Average*, *Median*, and *Sum*. In case the image contains spots that could influence the quantified values, the Median option will provide more reliable results than the arithmetic averages.


3.3.5.36 Select *Median* integration and press <OK>.

In order to illustrate the calibration feature, we will define one of the cells as negative control (minimum value), and another cell as positive control (maximum value).

3.3.5.37 Select cell A1 (negative control) and *Quantification > Define calibration point*. Enter 0 as value and press <OK>.

3.3.5.38 Select cell A12 (positive control) and *Quantification > Define calibration point*. Enter 100 as value and press <OK>.

Since only two calibration points are defined now, it is obvious that the program needs to calculate a linear regression through the defined points, in order to re-quantify the other cells according to the negative and positive controls:


3.3.5.39 Select *Edit > Settings* or  and choose the *Quantification* tab.

3.3.5.40 Under *Calibration*, enter 1 as *Polynomial degree*. This will result in a first degree regression.


3.3.5.41 Press <OK> to close the *Settings* dialog box.

3.3.5.42 Select *Quantification > View calibration curve*. This shows a linear regression between the two calibration points, zero and 100.


Finally, there is one more thing to do, i.e. to copy the character values in the microplate opened in InfoQuest FP.

3.3.5.43 Select *Quantification > Export to clipboard* or .

Before closing the BNIMA program, you can save the entire configuration defined for this microplate system. If you load a next microplate, you can reload the grid and all other settings such as color scale, disabled cells, quantification parameters etc.

3.3.5.44 Select *File > Save configuration as* or .

3.3.5.45 Enter a name e.g. microplate, and press <OK> to save the configuration.

For next microplates you can reload the configuration using *File > Load configuration* or .

3.3.5.46 Close the BNIMA program.

3.3.5.47 Right-click on the *Experiment card* (see also 3.8.1), and select *Paste from clipboard* from the floating menu.

The microplate now is filled with data and looks like in Figure 3-50.

3.3.5.48 Click the upper left triangular button to close the experiment card.

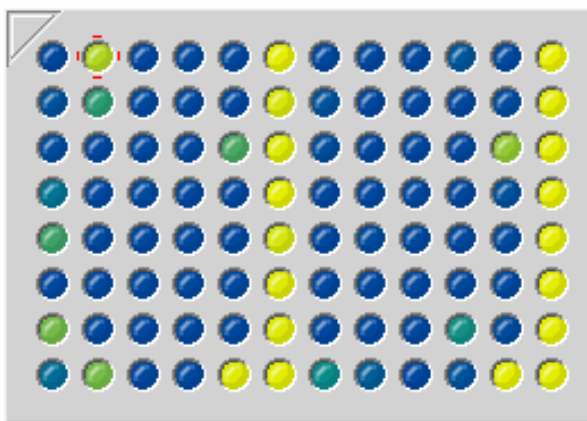


Figure 3-50. Example microplate experiment card after import of character values using BNIMA.

Example 2: import of gene-array scanning.

The second example we will use to illustrate the BNIMA program is **Array.tif**, a fragment of a gene array image which can be found in the **Sample and Tutorial data\Array image** directory on the installation CD-ROM. The array image was generated by chemiluminescent detection of digoxigenin-labeled cDNA¹. Each gene is characterized by two spots (horizontally next to each other), which can be considered as a control measure. For this example, we have used a fragment representing two blocks of 14 x 7 genes (the complete array is composed of six blocks of 14x7 genes, totalling 588 characters). The left and right half are separated by one blank column, and the two bottom rows contain calibration and reference spots (see Figure 3-51).

3.3.5.49 Create a new *closed* character type as described in 3.3.1.1 to 3.3.1.6, and call it **Gene array**. Specify 14

1. Courtesy S.D. Vernon, M.S. Mangalathu, and E.R. Unger (J. Histochemistry & Cytochemistry 1999; 47:337-342).

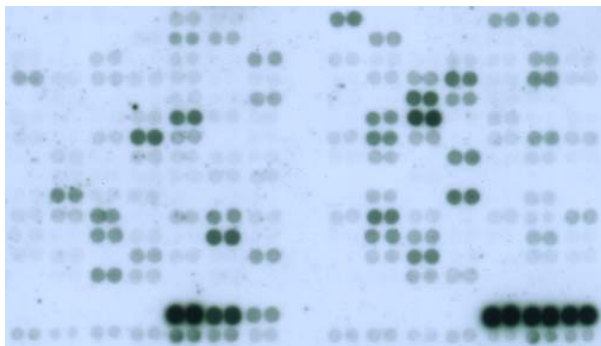


Figure 3-51. Fragment of gene array scanned as TIFF image (file Array.tif).

rows and 14 columns under *Layout* (third step), Figure 3-52.

3.3.5.50 When the **Gene array** experiment type is created, double-click on it in the *Experiments* panel.

3.3.5.51 In the appearing *Character type* window, select *Settings > General settings*, and click the *Experiment card* tab.

3.3.5.52 Under *Cell type*, select *Small blot*, which makes it possible to show large data sets in the experiment cards (see 3.8.1).

3.3.5.53 Click **<OK>** and close the *Character type* window.

3.3.5.54 Double-click on an entry to show its *Entry edit* window.

The experiment type **Gene array** shows an empty flask.


3.3.5.55 Click on the flask button. Since this experiment is not defined for the selected entry, the program asks "Do you want to create a new one?".

3.3.5.56 Answer **<Yes>** to create an *Experiment card* (see further, 3.8.1). An empty 14 by 14 array image pops up.

3.3.5.57 Right-click on the empty array image and select *Edit image* from the floating menu.

This loads the BNIMA program.

3.3.5.58 Select *File > Load image* in BNIMA and load the file **Array.tif** from the **Sample and Tutorial data\Array image** directory on the installation CD-ROM or from the downloaded and unzipped folder from the website. The resulting window looks as in Figure 3-52.

3.3.5.59 First call the *Settings* dialog box with *Edit > Settings* or .

The *Image* tab offers two choices for the *Image type*: *Densitometric* and *Color scale*.

Unlike the first microplate image, the color reaction of this gene array can be interpreted as a simple change in intensity (e.g. from light to dark), hence one should select *Densitometric*.

3.3.5.60 Select *Densitometric* under *Image type*.

The *Densitometric values* panel offers some additional tools to edit the TIFF file: *Inverted values* is to invert the densitometric values; *Background subtraction* allows a two-dimensional subtraction of the background from

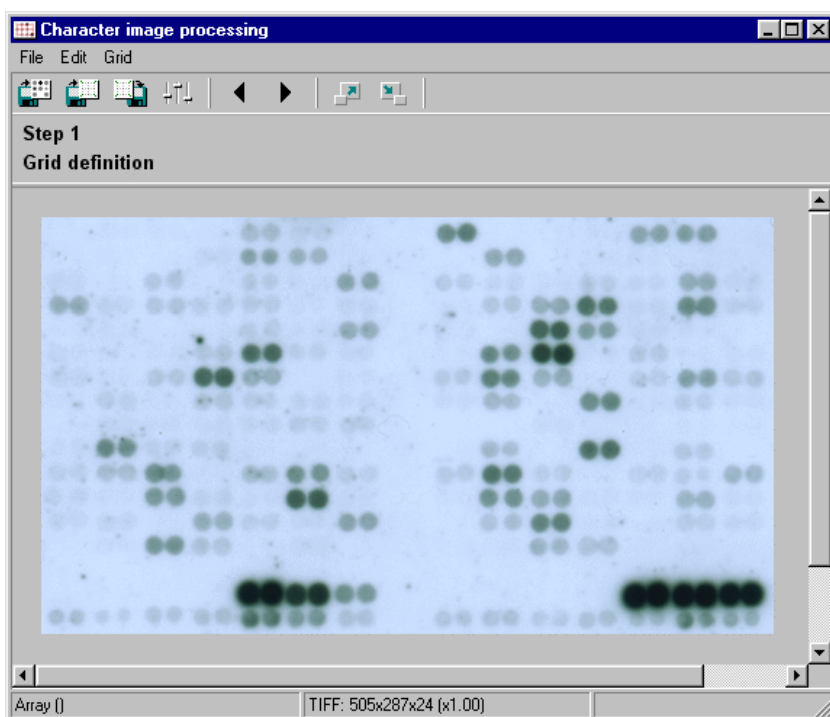


Figure 3-52. The BNIMA program with a gene array image (fragment) loaded.

the TIFF file, using the rolling ball principle. The *Ball size* can be entered in pixels. Background subtraction is only necessary if the illumination of the image is not uniform, which is not the case in the example image. Spot removal allows all spots and irregularities below a certain size to be removed from the image, whereas larger structures are preserved.

3.3.5.61 Leave *Background subtraction* disabled, and enable *Spot removal*, with a maximal *Spot size* of 3 pixels.

3.3.5.62 Press <OK> to quit the *Settings* dialog box.

The background subtraction and spot removal changes are only seen when *Edit > Show value scale* is enabled in the *InfoQuest FP main window*.

3.3.5.63 Check *Edit > Show value scale*. The image now looks “cleaned up”: spots are removed and the image is shown in grayscale rather than as 24 bit true color image.

In **Step 1: Grid definition** we will create a grid that defines the cells of the array.

3.3.5.64 Select *Grid > Add new* and enter 17 as *Number of rows* and 15 as *Number of columns*.

Choosing 17 and 15 rather than 14 by 14 is to allow the calibration spots to be included, and to take account of the blank column.

3.3.5.65 Press <OK> and the grid appears.

3.3.5.66 Move the upper left dragging node until the grid crosses match the middle of each double spot in the upper left area of the array (see Figure 3-53).



Figure 3-53. Correct alignment of grid on gene array spots.

3.3.5.67 Next, move the lower right dragging node until the grid crosses match the middle of each double spot in the lower right area of the array.

3.3.5.68 Then, move the lower left and upper right dragging nodes of the grid to distort the rectangle so that the grid crosses in the lower left and upper right areas, respectively, match with the double spots.

The grid on the image should now look as in Figure 3-54.

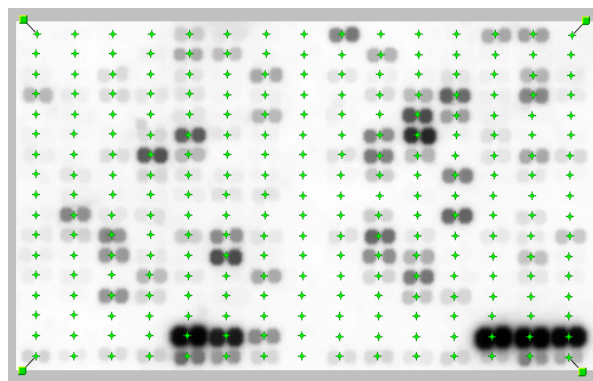



Figure 3-54. Correctly aligned grid on example gene array.

3.3.5.69 Move to the next step using *Edit > Next step* or the  button.

In this step, the layout of the cells is defined: the shape and size of the quantification area within each cell. In this step, we also define which cells we want to use for quantification and which cells not. By default, all cells of the grid are used for quantification.

3.3.5.70 Select the cells in the blank column of the image and *Cells > Delete selected*.

3.3.5.71 Similarly, select the three lowest rows and *Cells > Delete selected*.

Two cells of the second last row represent 0 and 100% hybridization respectively: the 4th and the 5th cell. We will include these cells for calibration, hence we have to include them again:

3.3.5.72 Select the 4th and 5th cell of the second last row and *Cells > Add selected*.

Before the program can do the quantification, it needs to know what the averaging area of the cells is. This is done using a *mask* which the user defines. In this case, it is clear that we will have to define two masks per cell, in order to cover the duplicate spots.

3.3.5.73 Select all cells as in 3.3.5.15.

3.3.5.74 Add a circular mask to all selected cells with *Cells > Add disk to mask*.

A dialog box prompts to enter a *Radius* for the disk in pixels, the *X offset* (horizontal shift from the cell marking cross) and the *Y offset* (vertical shift from the cell marking cross). For the offsets, a negative value can be entered.

3.3.5.75 Enter 6 as radius, and -6 as *X offset*. Press <OK> to confirm.

The masks appear on all used cells of the grid as semi-transparent red disks.

3.3.5.76 Add a second mask to all selected cells with *Cells > Add disk to mask*.

3.3.5.77 Enter 6 as radius, and 6 as *X offset*. Press **<OK>** to confirm.

After these steps, the *BNIMA* window should look like in Figure 3-55.

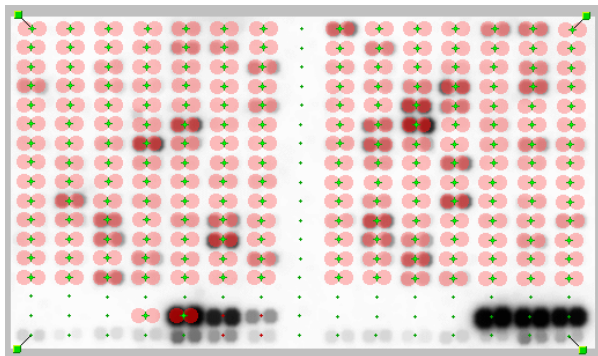




Figure 3-55. Array editing in BNIMA, with included and excluded cells, and masks defined.

3.3.5.78 If this is the case, move to the next step using *Edit > Next step* or the  button.

3.3.5.79 In the *Quantification* step, first call the *Settings* dialog box with *Edit > Settings* or .

3.3.5.80 Select the *Quantification* tab, specify a first degree polynomial fit and click **<OK>**.

3.3.5.81 Select the 4th cell in the second last row and *Quantification > Define calibration point*.

3.3.5.82 Enter 0 (zero).

3.3.5.83 Select the 5th cell in the second last row and *Quantification > Define calibration point*.


3.3.5.84 Enter 100.

All cells are now quantified between the zero and 100% hybridization control, and we now need to specify which cells to add to the character set. Since the calibra-


tion cells (second last row) are not part of the character set, these should not be included.

3.3.5.85 Select all but the three last rows and *Quantification > Add cells to character set*.

The cells to be used in the character set are now numbered 1 to 196.

3.3.5.86 Copy the quantified cells to the clipboard with *Quantification > Export to clipboard* or .

Before closing the *BNIMA* program, you can save the entire configuration defined for this gene array system:

3.3.5.87 Select *File > Save configuration as* or .

3.3.5.88 Enter a name e.g. "Gene array", and press **<OK>** to save the configuration.

3.3.5.89 Close the *BNIMA* program.

3.3.5.90 Right-click on the *Experiment card* (see also 3.8.1), and select *Paste from clipboard* from the floating menu.

The experiment card now is filled with data and looks like in Figure 3-56.




Figure 3-56. Example gene array experiment card after import of character values using BNIMA.

3.3.5.91 Click the upper left triangular button to close the experiment card.

3.4 Setting up sequence type experiments

3.4.1 Defining a new sequence type

3.4.1.1 Select *Experiments > Create new sequence type* from the *InfoQuest FP* main menu, or press  and *New sequence type*.

3.4.1.2 The *New sequence type* wizard prompts you to enter a name for the new type. Enter a name, for example **SSU-Ribo**.

3.4.1.3 Press **<Next>** and check the kind of the sequences: *Nucleic acid sequences* or *Amino acid sequences*. Select *Nucleic acid sequences*.

3.4.1.4 Press the **<Finish>** button to complete the setup of the new sequence type. It is now listed as a sequence type in the *Experiments* panel.

The new sequence type exists by now, and we can enter sequence data in several ways:

1. Importing sequences in GenBank, EMBL and Fasta formats using the Import plugin (see 3.4.2).
2. Assembling sequencer trace files into consensus sequences using *InfoQuest FP*' own Assembler program (see 3.4.3).
3. Defining a new sequence file in *InfoQuest FP*, and entering the bases manually, or pasting them from the clipboard.
4. Entering or pasting the sequences via the experiment card of the database entry (see 3.8.4).

In *InfoQuest FP*, sequences up to 200,000 bases can be imported and analyzed.

3.4.2 Importing sequences

•Importing sequences in GenBank, EMBL and Fasta formats

InfoQuest FP can import sequences in GenBank, EMBL and Fasta formats using the Import plugin.

As an example, we will import the **embl.txt** file that is provided in the **Sample and Tutorial data\Sample text files for import** directory on the installation CD-ROM. The same text file is also available via the download page of the website (www.bio-rad.com/software-downloads). In order to understand the import of the informa-

tion present in the text file, you may want to open the file **embl.txt** in Notepad (Wordpad) or another text editor. EMBL files (and GenBank files) contain for each sequence a header of which the information is characterized by tags. In EMBL, "DE" refers to the organism name, "AC" is the accession number, "KW" is the keyword, etc. Based on these tags we are going to import the information present in the text file.

First we are going to create two new information fields in our database.

3.4.2.1 Select *Database > Add new information field*, or right-click in the information fields toolbar of the *Database entries* panel. Enter the name "Gene". Repeat this step and name the second field "Name".

3.4.2.2 Install the Import plugin (see paragraph for more information).

3.4.2.3 Choose *File > Import > Import sequences*.

3.4.2.4 In the *Import* dialog box, select the **embl.txt** file and press **<Open>**.

3.4.2.5 Select **SSU-Ribo** in the first column, and specify EMBL as the file format. Press **<OK>**.

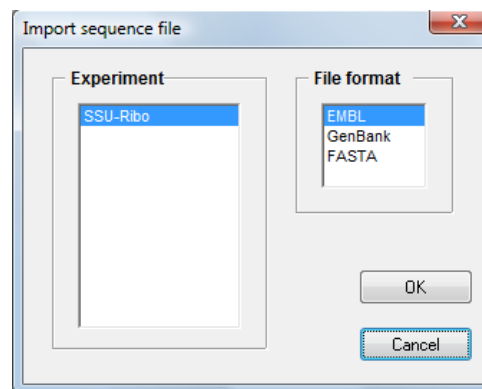


Figure 3-57. The *Import sequence file* dialog box.

3.4.2.6 In the next window, select *Key* from the drop-down menu next to the tag "AC", link *Gene* to the tag "KW" and *Name* to the tag "OS".

3.4.2.7 Press **<OK>**.

Twenty new database entries are created.

NOTE: Files to import should contain no more than 300 entries. If the file is larger, the number of imported

entries will be truncated after the first 300. If you want to import more than 300 sequences from single files, you should use the sequence import script(s) available on the website of Bio-Rad, as explained above.

• Defining a new sequence file

A new sequence file is created as follows:

3.4.2.8 In database **Example**, select the new sequence type **SSU-Ribo**.

3.4.2.9 Right-click in the *Files* panel, and choose **Add new experiment file** from the floating menu.


NOTE: This feature is only accessible when working in a local database. Entering or pasting the sequences via the experiment card of the database entry works both on local and connected databases (see 3.8.4)

3.4.2.10 Enter a name, e.g. **Seq01** and press **<OK>**.

3.4.2.11 Select **Seq01** in the *Files* panel, and **File > Open experiment file (data)**.

This opens the *Sequence data file* window, which is empty initially.

Before you can enter sequences, you have to add new entries to the file. Suppose that we want to add sequence data for three more entries of the database.


3.4.2.12 Select **Entries > Add new entries** or  .

3.4.2.13 You are prompted to enter the number of entries; enter 3 and press **<OK>**.

Three entries are now present, and all sequences are initially represented by a blank line.

With **Sequence > Paste from clipboard**, the contents of the clipboard is pasted into the selected sequence.



3.4.2.14 Double-click or **Sequence > Edit** to edit the sequence or to enter the bases manually.


3.4.2.15 If you are doing a lot of editing work, we recommend to save now and then with **File > Save** () or the F2 shortcut.



3.4.2.16 **File > Exit** when you are finished editing the sequences.

3.4.2.17 In the *InfoQuest FP main* window, double-click on the file **Seq01** or click on the file and select **File > Open experiment file (entries)**.

The *Sequence entry file* window (cf. the *Fingerprint entry file* window, Figure 3-27) contains unlinked entries, which you can now link to the corresponding database entries.

A link arrow  for each entry allows you to link an entry to a database entry, by clicking on the arrow and dragging it onto a database entry, and then releasing the mouse button. When the experiment is linked, its link arrow is purple:  .


3.4.2.18 Drag the link arrow of **entry 1** to any database entry that does not have a sequence linked: as soon as you pass over a database entry, the cursor shape changes into  .

3.4.2.19 Release the mouse button on the database entry; entry is now linked to this database entry, and its arrow in the *Sequence entry file* window has become purple  instead of gray  .

NOTE: if you try to link an entry to a database entry which already has an entry of the same experiment type linked to it, the program will refuse the second link with the message: "The Experiment 'SSU-Ribo' of this database is already defined in XXX", where XXX is another lane of the same sequence type, in the same or another experiment file.

As soon as an experiment is linked to a database entry, the *Experiment presence* panel (see Figure 1-15) shows a colored dot for the experiment of this entry.

You can edit the information fields for this entry in two places: directly in the database (see 2.2.3.1 to 2.2.3.2), or in the *Sequence entry file* window, by either double-clicking on the entry (opens the *Entry edit* card) or clicking twice on the entry (enables direct editing).

*NOTE: Experiment files added to the Experiments panel can also be deleted by selecting the file and choosing **File > Delete experiment file** from the main menu or clicking on  in the *Files* panel.*

*Deleted experiment files are struck through (red line) but are not actually deleted until you exit the program. So long, you can undo the deletion of the file by selecting **File > Delete experiment file** or clicking*

 again.

3.4.3 Input of sequences using the InfoQuest FP Assembler program

Assembler is a program to assemble contig sequences from partial sequences which result from sequencing experiments. The program accepts flat text files as well as binary chromatogram files from ABI, Beckman, and Amersham automated sequencers, including the SCF sequence trace format. In the latter cases, Assembler allows the user to verify base assignments by inspecting the chromatograms along with the partial sequences and


the consensus sequence. Assembler also investigates the quality and ambiguity of the curve profiles to assign a quality label to the partial sequences and trim off bad parts where necessary.

Contig sequences are saved into projects, which contain all the information about the partial sequences, the editing made by the user, the multiple alignment, and the editing done on the contig. A contig project and its full information can be opened at any time from the InfoQuest FP sequence entry to which it is associated. Assembler can handle thousands of sequences in one single contig project and is optimized for speed and editability in large projects. The program can be launched from InfoQuest FP but not as a separate program.

NOTE: The Batch sequence assembly plugin allows the Assembler program to run in batch mode, thereby assembling a large number of trace files into multiple contigs. See the separate Batch sequence assembly plugin for more information.

3.4.3.1 Double-click on an entry in the database which does not have a sequence assigned.

The *Entry edit* window of the entry appears.

3.4.3.2 Click on the  button next to the sequence type (e.g. SSU-ribo or another name you entered).

The program now asks "The experiment "SSU-ribo" is not defined for this entry. Do you want to create a new one?".

3.4.3.3 By answering <Yes>, the program will create a new empty sequence that is linked to this entry. The experiment card for the sequence type of this entry appears: a small empty window (see Figure 3-58 and 3.8.1).

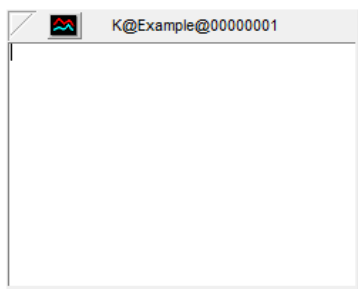




Figure 3-58. Empty sequence experiment card with button to launch Assembler.

3.4.3.4 It is now possible to simply paste a sequence in this window (see 3.8.4); however, pressing the  button will launch the Assembler program to assemble a contig sequence from a series of partial sequencing experiments for this entry.

A. The Assembler main window

The *Assembler main* window, initially empty, looks as in Figure 3-59.

3.4.3.5 Select *File > Import sequence files* or . Under *Files of type*, different formats can be imported, including flat text files. By default, the ABI file type is selected.

A set of partial 16S rDNA sequences from *Xanthomonas* strain ICMP 9121¹ run on an ABI 370 machine are provided in the **Sample and Tutorial data\16S rRNA sequencer trace files** directory on the installation CD-ROM or can be downloaded from the website (www.bio-rad.com/softwaredownloads). Being run on an old sequencer type, these trace files have a rather poor resolution, and hence, are suitable to illustrate the automatic sequence trimming and quality features.

3.4.3.6 Select all sequences **11 ICMP 9121 ... 18 ICMP 9121** in the **Sample and Tutorial data\16S rRNA sequencer trace files** directory on the CD-ROM or in the downloaded and unzipped folder from the website and press <Open>.

The six partial sequences are now shown in the *Assembler main* window as in Figure 3-59. The window consists of two tabs: *Trimming* and *Assembly*. The first tab, *Trimming*, displays the original sequences and gives an indication of the quality.

NOTES:


(1) The colors of text, background, bases, and all other symbols may be changed by the user. The descriptions below are given using the default colors, which can be obtained by selecting **View > Display settings** and pressing <Default>.


(2) Sequences can also be imported directly from the clipboard using the function **File > Import sequence from clipboard**.

The *Traces overview* panel (top right) shows the sequences in a graphical representation. For each sequence, there is a quality assignment, based on the quality of the densitometric curves and the base assignment. Based on the quality, the program will automatically trim the bad parts from the sequences, which are underlined with a black bar. Unknown bases (ambiguous positions) are indicated with a dark red flag on top of the sequence.



The *Traces list* panel (top left) shows the corresponding file names in the upper line. In the bottom line, the original size in base pairs and the size after trimming are shown for the sequence.



1. Hauben L., L. Vauterin, J. Swings, and E.R.B. Moore. 1997. *Int. J. Syst. Evol. Microbiol.* 47: 328-335.



3.4.3.7 You can zoom in on the *Traces overview* panel with **View > Zoom in (overview)**, by pressing  in the *Traces overview* panel toolbar or by using the zoom slider for the *Traces overview* panel (see 1.6.7 for detailed instructions on the use of zoom sliders).



3.4.3.8 To zoom out on the *Traces overview* panel, use **View > Zoom out (overview)**, press  left from the overview panel or use the zoom slider for the *Traces overview* panel.

3.4.3.9 The *Raw trace* panel (bottom) displays the densitometric curve in four colors and the corresponding bases for the selected sequence.

3.4.3.10 You can horizontally zoom in on the curve with **View > Zoom in (trace)**, with  in the toolbar of the *Raw trace* panel or by using the  zoom slider (see 1.6.7) in the *Raw trace* panel.

3.4.3.11 To zoom out on the curve in the horizontal direction, use **View > Zoom out (trace)**,  in the toolbar of the *Raw trace* panel or use the  zoom slider in the *Raw trace* panel.

3.4.3.12 To enlarge the curve vertically, click the  button in the toolbar of the *Raw trace* panel or use the  zoom slider (see 1.6.7) in the *Raw trace* panel.

3.4.3.13 To shrink the curve vertically, click the  button in the toolbar of the *Raw trace* panel or drag the  zoom slider in the *Raw trace* panel.

3.4.3.14 A sequence can be selected from the *Traces overview* panel, or from the *Traces list* panel. The selected sequence is highlighted and its graphical overview is bordered by a colored rectangle.

3.4.3.15 A position can be selected on any sequence in the *Traces overview* panel by clicking it with the mouse. The selected position is indicated with a blue vertical line. The corresponding sequence chromatogram is shown in the *Raw trace* panel, with the selected position centralized and highlighted in blue.

3.4.3.16 Likewise, a base position can be selected on the curve in the *Raw trace* panel, which causes the selection to be updated in the upper panels as well.

The logical working flow for a contig assembly is

1. Cleaning trace files and quality assignment

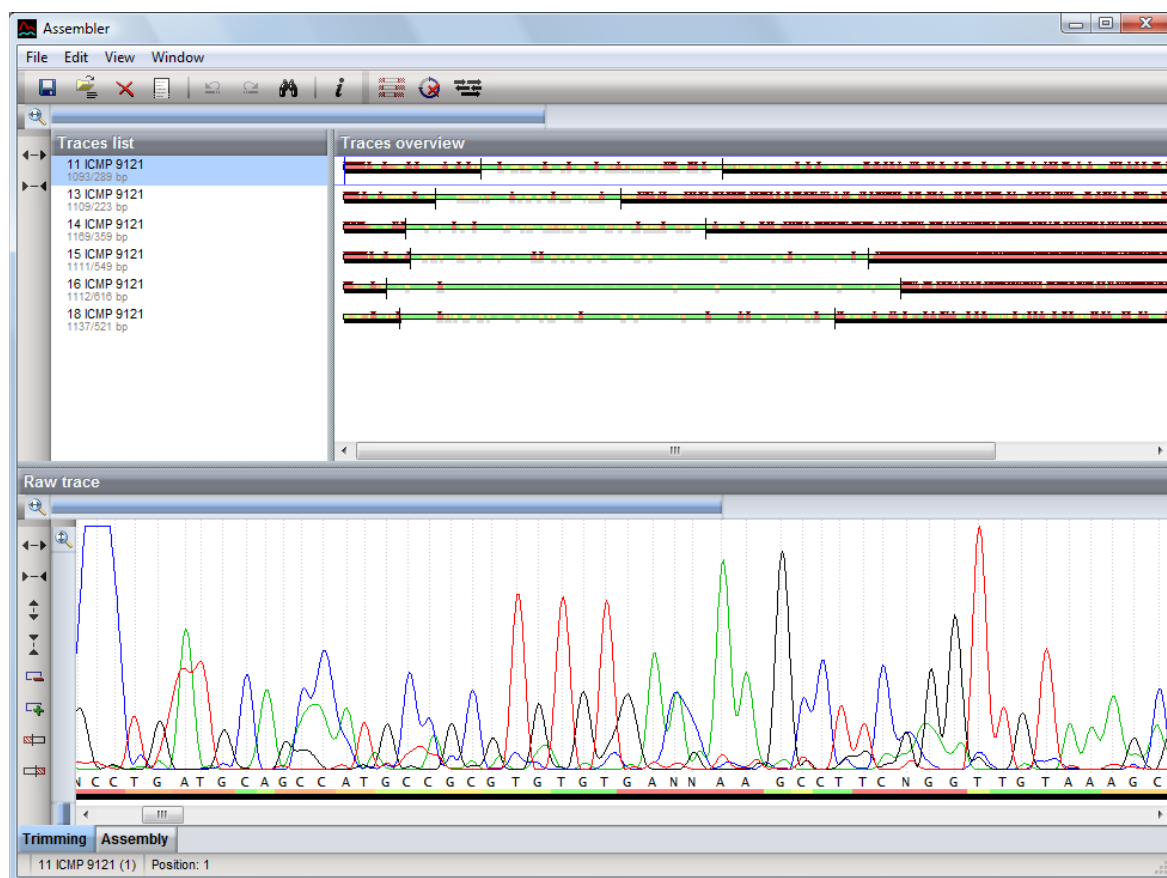


Figure 3-59. The *Assembler* main window.

2. Manual inspection of cleaning result
3. Removal of vector sequences (optional)
4. Assembling the contig (multiple alignment)
5. Manual inspection and correction of mismatches and unresolved positions
6. Trimming the consensus sequence according to known start and end signatures (e.g. primers) (optional)

The steps will be described in this order.


B. Cleaning chromatogram readouts

Before we actually align the sequences, we need to have the bad parts cut out, i.e. the outermost left and/or right parts from the curves with unreliable signal or no signal at all. This process, called cleaning of sequences, consists of two levels:

1. *Trimming* of the sequences, i.e. physically removing the unusable ends. This level of cleaning is based upon the percentage of unresolved positions at both ends of the sequence. Trimmed ends are neither used, nor shown in the *Assembly* view of the *Assembler main* window.
2. Inactivating doubtful parts of the sequence. This level of cleaning is based both on the quality of the densitometric curves and the proportion of unresolved positions. Inactivated parts are still shown, but do not actively contribute to obtain the consensus. However, they are aligned to the consensus. In case there is no consensus base at a position, the inactivated regions will not be considered by the program. The user can still compare the consensus position with the base in an inactivated sequence region. Inactive regions can still be set as active at anytime, whereas active regions can be set as inactive as well. In case an inactivated region is the only information available in a part of the consensus sequence, it will be used to fill in the consensus sequence. In case a position on an inactivated region conflicts with other sequences, it will be ignored.

3.4.3.17 Cleaning of the sequences happens automatically and is based on the *quality assignment* settings. The quality of the sequence is shown on the *Traces overview* panel in the *Trimming* view (Figure 3-59). A color scale ranges from green (acceptable quality) over yellow and orange to red (unacceptable quality). The trimmed ends are indicated by a black bar underlining the sequence. Inactivated zones are indicated by a gray bar. Unresolved positions ('N') are indicated with a small flag on top of the sequence.

3.4.3.18 The quality assignment can be changed by modifying the settings in the *Quality assignment* dialog

box (Figure 3-60). This dialog box can be popped up with *File > Quality assignment* or .

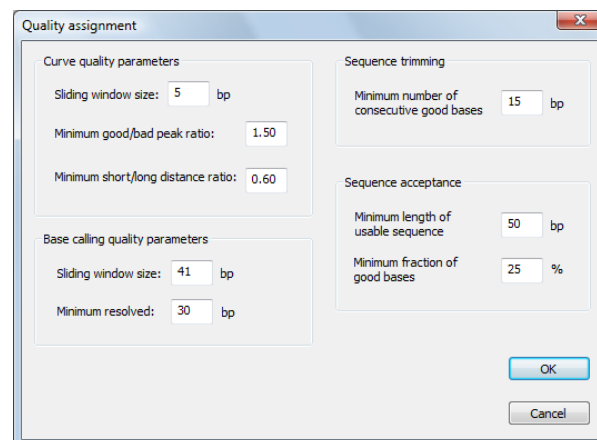


Figure 3-60. The *Quality assignment* dialog box.

3.4.3.19 The *Curve quality parameters* determine how the program will investigate the quality of signal derived from the curves. They include two ratios that are considered in a certain window, determined by the *Sliding window size*. The latter should be an odd number, including the position itself and a number of positions at either side.

The *Minimum good/bad peak ratio* is the ratio between the signal strength of the weakest peak resulting in a base and the strongest peak not resulting in a base within the sliding window. The higher this ratio is set, the more stringent the quality assignment becomes. A suitable starting value for most systems is 1.50.

The *Minimum short/long distance ratio* is the ratio of the shortest distance between two positions and the longest distance within the sliding window. A suitable starting value is 0.60; the larger it is set, the more stringent the quality assignment will be.

A typical value for the *Sliding window size* is 5 positions; increasing this value will result in a more stringent quality assignment.

3.4.3.20 Under *Base calling quality parameters*, you can specify a *Sliding window size*, and the number of resolved positions that should be found within the sliding window (*Minimum resolved*). Similar as under *Base quality assignment*, the *Sliding window size* should be an odd number. Suggested starting values are a *Sliding window size* of 41 of which minimum 30 resolved positions.


3.4.3.21 *Sequence trimming* is based upon the *Minimum number of consecutive good bases*, as defined by the *Curve quality parameters* and the *Base calling quality parameters*. A suggested value is 15; the larger the number, the heavier the sequence will be trimmed.


3.4.3.22 The *Sequence acceptance* parameters determine whether a sequence as a whole will be accepted to contribute to the consensus or not. The *Minimum length of usable sequence* determines the length of the non-trimmed part of the sequence. The *Minimum fraction of good bases* determines the ratio of good bases over the total number of bases in the usable part of the sequence. Suggested values are 50 bases of which minimum 25% good bases.


3.4.3.23 For the example sequences, which were generated on an old sequencer and have a rather poor quality, it is recommended to change the standard trimming settings slightly: under *Curve quality parameters: Sliding window size* 5; *Minimum good/bad peak ratio* 1.30; *Minimum short/long distance ratio* 0.60. Under *Base calling quality parameters: Sliding window size* 51; *Minimum resolved* 30. The other settings can remain unchanged, i.e. under *Sequence trimming*, 15 bp and under *Sequence acceptance* 50 bp and 25%, respectively.


3.4.3.24 Automatic cleanup (trimming and assignment of inactive zones) happens automatically after pressing the <OK> button in the *Quality assignment* dialog box. Any manual trimming and (in)activation done (see further) will be lost at this point.

After automatic quality assignment and trimming, the user can still manually correct the trimmed ends and inactive zones.

3.4.3.25 To mark the start of a sequence, click on the position to start (this can be done both on the overview and on the curve) and select *Edit > Mark start of sequence*. You can also use the  button in the toolbar of the *Raw trace* panel or the CTRL+Home shortcut key on the keyboard.

3.4.3.26 To mark the end of a sequence, click on the position to end (this can be done both on the overview and on the curve) and select *Edit > Mark end of sequence*. You can also use the  button in the toolbar of the *Raw trace* panel or the CTRL+End shortcut key on the keyboard.



3.4.3.27 To mark a zone as inactive, click on the start position of the zone, then hold down the SHIFT key while clicking on the end position of the zone (this can be done both in the overview and on the curve). Choose *Edit > Inactivate selected region* or press the  button in the toolbar of the *Raw trace* panel. A shortcut is the - (minus) key on the keyboard.

3.4.3.28 To mark a selected zone as active, choose *Edit > Activate selected region* or press the  button in the

toolbar of the *Raw trace* panel. A shortcut is the + (plus) key on the keyboard.


A sequence can be inactivated as a whole with *Edit > Inactivate selected sequence*. When inactivated, a sequence is marked with a red cross in the *Traces list* panel (upper left).

A sequence that was inactivated by the *Sequence acceptance* parameters (3.4.3.22) can be activated manually with *Edit > Activate selected sequence*.

3.4.3.29 A sequence can be removed from the project with *File > Remove selected sequence* or . Conversely, sequences can be added to a project at any time with *File > Import sequence files* or .

C. Removing vectors

If the sequences contain residues from vector sequences, these need to be removed before the sequences are assembled.

3.4.3.30 Vectors can be removed from the unaligned sequences with *File > Remove vectors* or . This pops up the *Remove vectors* dialog box (Figure 3-61).

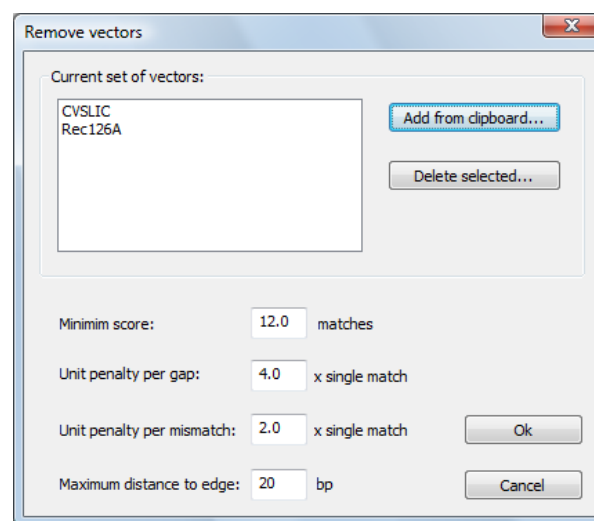


Figure 3-61. The *Remove vectors* dialog box.

3.4.3.31 Vector sequences to be removed can be added from the clipboard (by copying from another application). They can be pasted in the list by pressing <Add from clipboard>. This opens a new window, the *Import vectors from clipboard* editor (Figure 3-62). The sequence on the clipboard is automatically pasted into the editor,

which the user can still edit. An input field *Name* allows a name to be entered for the vector.

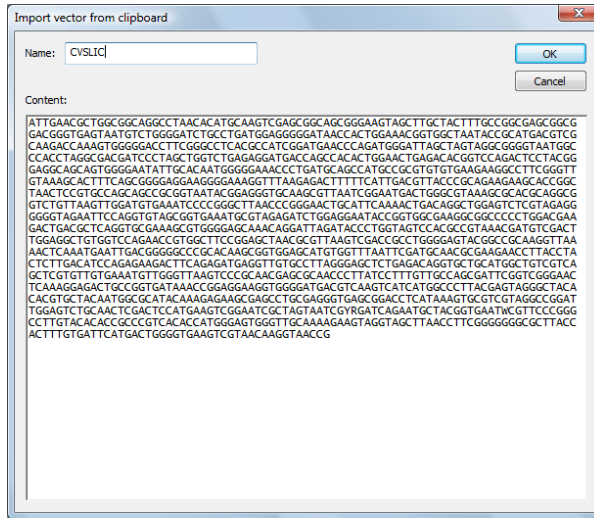


Figure 3-62. Import vectors from clipboard editor.

3.4.3.32 Vectors can be deleted from the list using the **<Delete selected>** button.

Vectors entered are automatically saved along with the project.

The *Remove vectors* dialog box (Figure 3-61) contains a number of alignment parameters:

3.4.3.33 **Minimum score:** the minimum number of matching bases the sequence and the vector should have in order for the vector sequence to be removed. This number is the result of the total number of matching bases minus the total penalty resulting from mismatches and gaps.

3.4.3.34 **Unit penalty per gap:** the penalty, as a factor of the match score, assigned to a gap in either the sequence or the vector after the alignment.

3.4.3.35 **Unit penalty per mismatch:** the penalty, as a factor of the match score, for a single mismatch between the vector and the sequence after the alignment.


3.4.3.36 **Maximum distance to edge:** the maximum number of unmatched bases at the end of the sequence. Normally a vector sequence will extend over the end of the trace sequence, so one will not expect unmatched bases at the end of the sequence. Therefore, this number should be set very low (e.g. 5 or less).

3.4.3.37 By pressing **<OK>** the vector sequences are automatically searched for and removed from the unaligned sequences. Removed vector sequences are indicated in blue in the overview panel.

NOTE: To undo vector removal, open the Remove vector dialog box, delete all vectors defined and press **<OK>**. Vector removal as well as undoing vector

removal can only be executed on unaligned sequences. If sequences are already aligned, you will first have to remove the consensus (see below).

D. Alignment to consensus

3.4.3.38 The sequences are assembled into a consensus with the menu command **File > Assemble sequences** or by pressing the  button. The *Calculate assembly* dialog box is displayed (Figure 3-63), allowing the various alignment parameters to be entered.

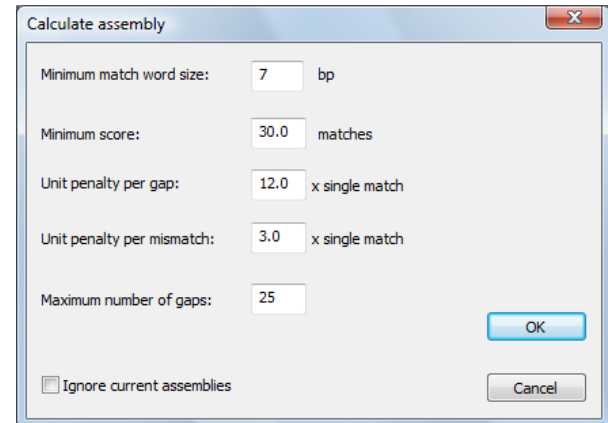


Figure 3-63. Calculate assembly dialog box with alignment parameters.

The **Minimum match word size** determines the number of bases that are taken together into one *word*. The algorithm creates a lookup table of groups of bases to accelerate the alignment, which increases the speed of the algorithm. In an alignment to a consensus sequence, no mismatches are expected, except due to bad base calling. In that case, it is justified to choose a high *word size* number. In the default setting of 7, only stretches of 7 identical bases or more will be considered as matches.

Minimum score: the minimum number of matching bases the two sequences should have before they will be aligned. This number is the result of the total number of matching bases minus the total penalty resulting from mismatches and gaps.

Unit penalty per gap: the penalty, as a factor of the match score, assigned to a gap introduced in one of the sequences after the alignment.

Unit penalty per mismatch: the penalty, as a factor of the match score, for a single mismatch between the two sequences after the alignment.

Maximum number of gaps relates to the alignment technique that is used, i.e. a fast algorithm based upon Needleman and Wunsch (1970)¹. The number of gaps the algorithm can create is proportional to the number of diagonals specified. The larger the number, the more



Figure 3-64. The *Assembler* main window, *Assembly* view (second tab).

accurate but the slower the calculations. The suggested default setting is 25 diagonals.


The check box *Ignore current assemblies* allows the algorithm to recalculate the consensus sequence(s) from individual trace sequences without taking into account any already calculated contigs.

3.4.3.39 Press <OK> to calculate the assembly or <Cancel> to exit the *Calculate assembly* dialog box without anything to happen.

E. Editing a consensus sequence

When the alignment is finished, the second view, i.e. the *Assembly* view, is shown (Figure 3-64). As compared to the first view (*Trimming* view, see Figure 3-59), a central *Alignment* panel now shows the consensus sequence (upper line) and the individual trace sequences that contribute to the displayed consensus.

The *Alignment overview* panel (top right) displays the aligned trace sequences. If the arrow points to the left, the program has invert-complemented the sequence to obtain the correct alignment.

3.4.3.40 You can have the names of the trace files displayed on top of the bars in the *Alignment overview* panel by pressing the  button, or by selecting *View > Show trace names*.

The *Alignment list* panel (top left) now displays the selected consensus with its length and the number of sequences that are part of it. If the program could not align all trace sequences to a single consensus, the panel lists the different consensus sequences with their lengths and number of trace sequences. One should click on a particular consensus sequence to select it for viewing and editing.

The *Traces* panel (bottom) has two tabs: the *Raw trace* and the *Aligned traces* tab. Depending on which view was last displayed when *Assembler* was closed, the *Raw trace* or *Aligned traces* view is shown.



The *Raw trace* view displays the chromatogram file for the selected trace sequence. Regardless of whether the



1. Needleman, S., and C. Wunsch. 1970. *J. Mol. Biol.* 48:443-453.



sequence is invert-complemented in the alignment, the chromatogram is always shown in original mode. This means that, when the sequence has been invert-complemented, a G on the original sequence, for example, will appear as a C on the consensus. Due to the fact that the direction of the curve can be opposite from the sequence and that the bases are not aligned, it is not possible to select bases on the raw curves directly.



The *Aligned traces* view has the following features:


- Curves have been stretched or shrunk to obtain equidistant spacing between the base positions
- Trace sequences are always shown as transformed and oriented in the consensus. If a sequence is invert-complemented, the complement of the bases is shown, and the colors of the curves are adjusted likewise.
- Multiple trace chromatograms can be shown together and are aligned to each other and to the consensus (see Figure 3-64).
- Arrows on the curves indicate the direction of the sequence: if the sequence has been inverted, the arrow points to the left (Figure 3-64).
- In the *Aligned traces* view, it is possible to select bases directly on the curves.

3.4.3.41 You can zoom in on the curves with **View > Zoom in (trace)**, with  in the toolbar of the *Traces* panel or by dragging the  zoom slider (*Aligned traces zoom horizontal*) in the *Traces* panel. See 1.6.7 for a description of zoom slider functions.



3.4.3.42 To zoom out on the curves, use **View > Zoom out (trace)**,  in the toolbar of the *Traces* panel or drag the  zoom slider (*Aligned traces zoom horizontal*) in the *Traces* panel.

3.4.3.43 To zoom in on the curves vertically, click the  button in the toolbar of the *Traces* panel, or drag the  zoom slider (*Aligned traces zoom vertical*) with the mouse. See 1.6.7 for a description of zoom slider functions.

3.4.3.44 To zoom out on the curves vertically, click the  button in the toolbar of the *Traces* panel, or drag the  zoom slider (*Aligned traces zoom vertical*) with the mouse.

3.4.3.45 With a second vertical  zoom slider (*Aligned traces stretch vertical*) in the *Traces* panel, the vertical space reserved for the curves can be determined. See 1.6.7 for a description of zoom slider functions.

3.4.3.46 A sequence can be moved up or down by selecting it and choosing **Edit > Move sequence up**

(PgUp or ) or **Edit > Move sequence down** (PgDown or ) , respectively.

3.4.3.47 Bases on the consensus sequence are assigned according to the *Consensus determination* parameters, which can be set with **Assembly > Consensus determination**. The dialog box (Figure 3-65) allows four parameters to be set:

- **Required bases to include position:** The percentage of sequences that need to have a base at a certain position in order for the position to be inserted in the consensus. For example using the default value 50, if the consensus is determined by three sequences at a certain position, it will not be accepted as a base if there is a gap in two of the three sequences (33.3%).
- **Required consensus for unique base calling:** The percentage of sequences that need to have the same base at a position in order for the base to be accepted as resolved.
- **Required consensus for 2-fold degeneracy:** The summed percentage of sequences having two different bases at a position in order to be denoted with IUPAC code for 2-fold degenerated positions (R, M or S for A/G, C/A or C/G, respectively). Only applicable for positions that do not fulfill the criterion for unique base calling.
- **Required consensus for 3-fold degeneracy:** The summed percentage of sequences having three different bases at a position in order to be denoted with IUPAC code for 3-fold degenerated positions (Y, K, W or V, for C/T/U, T/U/G, T/U/A or A/C/G). Only applicable for positions that do not fulfill the criteria for unique base calling and 2-fold degeneracy. Any position that does reach the required consensus for 3-fold degeneracy is denoted as "N".
- **Allow group editing of sequences** is a feature that allows bases to be changed directly on the consensus sequence. If this feature is enabled (default), corresponding positions on the trace files will be changed

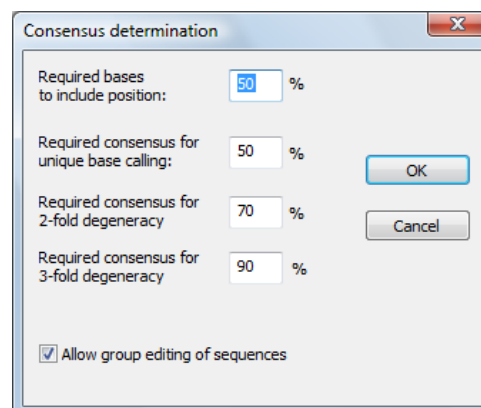


Figure 3-65. The *Consensus determination parameters* dialog box.

accordingly. If disabled, bases can only be changed on the individual trace sequences.

*NOTE: If you do not wish to use the parameters **Required consensus for 2-fold/3-fold degeneracy**, you can leave these parameters empty.*

Unresolved positions on the consensus are indicated in pink and extend over all sequences shown (*Alignment* panel, see Figure 3-64).

Problem positions on individual trace sequences, which have been solved under the current *Consensus determination* parameters (3.4.3.47) are indicated in orange. Such problem positions include mismatches as well as unresolved positions.


3.4.3.48 To change a base in a trace sequence, place the cursor on the base or on the position on the chromatogram and type the base, which can be A, G, C, or T, or any IUPAC code for denoting ambiguous positions.

3.4.3.49 To delete a base, select *Edit > Delete base* or press the DEL key.

3.4.3.50 To insert a position, select *Edit > Insert column* or press the INSERT key.

3.4.3.51 If consensus editing is enabled in the *Consensus determination* parameters (3.4.3.47), it is also possible to place the cursor on the consensus sequence and type a base, which causes the base to be changed on all sequences that have signal at the selected position.

3.4.3.52 As mentioned before, the Assembler program contains a multistep undo and redo function. In addition, the program also stores a history of editing actions done on each individual sequence. This information can be popped up by selecting the sequence (clicking on any position on the sequence, in the chromatogram or on the overview) and calling *Edit > Sequence information*

(CTRL+I) or pressing the  button. The *Sequence editing information* dialog box (Figure 3-66) lists all base corrections that are made to the sequence. The corrections recorded include base changes, deletions and insertions.

3.4.3.53 From the *Sequence editing information* dialog box, you can select a particular editing action in the list, and press *<Select on sequence>*. The position will be selected on the sequence. A correction made can be undone by pressing the *<Discard change>* button.

3.4.3.54 A range of bases can be selected on the curves or on the sequences in the central panel by clicking the first position of the range, then holding down the SHIFT key while clicking on the last position. A selected range is highlighted by a blue rectangle in the sequence view. Range selection by dragging the mouse is also possible in the sequence view.

3.4.3.55 If a selected selection of bases is flanked by a gap at one side, it is possible to shift the selection towards

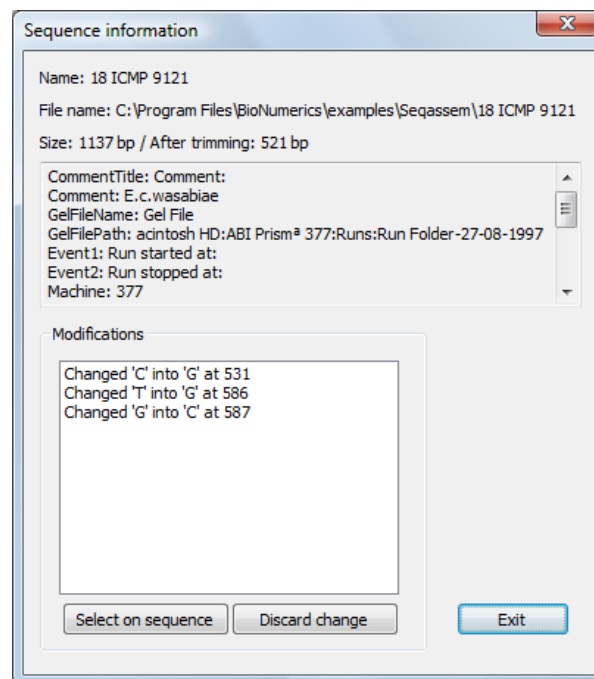




Figure 3-66. The *Sequence editing information* dialog box.

that gap, to correct misalignments. Shifting towards the left can be done with ALT+Left arrow key and shifting towards the right with ALT+Right arrow key. These commands can also be found in the menu (*Edit > Shift block left*, and *Edit > Shift block right*, respectively).


3.4.3.56 To check the consensus sequence for correctness, you can let the program jump to each next unresolved problem position using *View > Next unresolved problem* or  (or using the shortcut CTRL+Right arrow key on the keyboard).

3.4.3.57 To jump to the previous unresolved problem, use *View > Previous unresolved problem* or  (or using the shortcut CTRL+Left arrow key on the keyboard).

3.4.3.58 In case the program has incorrectly aligned a sequence to one or more other sequences, you can place the cursor on the misaligned sequence and select *Assembly > Break selected sequence apart*.


3.4.3.59 New sequences can be added at any time to the existing alignment project by switching to the first view and selecting *File > Import sequence files*, and subsequently selecting *File > Assemble sequences*. In the *Calculate assembly* dialog box (Figure 3-63), *Ignore current assemblies* should normally be unchecked, to preserve the assembly or assemblies already present.


Although Assembler automatically inverts and complements subsequences wherever necessary to obtain the consensus sequence, the program cannot know the correct orientation of the consensus sequence. Hence, it may be necessary to invert and complement the consensus sequence before entering it into the database.


3.4.3.60 Invert-complement the consensus sequence by selecting the consensus to invert and **Assembly > Invert direction** or .


NOTE: In case the program could not find one single consensus for all subsequences, two or more assemblies will exist. Therefore you will need to select the assembly to invert from the list in the Alignment list panel (upper left) before executing the invert-complement function.

The following editing actions are available to further clean up sequences (see also 3.4.3.25 to 3.4.3.28 in the *Trimming* view).

3.4.3.61 To mark the start of a sequence, click on the position to start and select **Edit > Mark start of sequence**. You can also use the  button in the toolbar of the *Alignment* panel. A shortcut is CTRL+Home on the keyboard.

3.4.3.62 To mark the end of a sequence, click on the position to end and select **Edit > Mark end of sequence**. You can also use the  button in the toolbar of the *Alignment* panel. A shortcut is CTRL+End on the keyboard.


3.4.3.63 To mark a zone as inactive, click on the start position of the zone, then hold down the SHIFT key while clicking on the end position of the zone (this can be done both on the sequence and on the curve, not on the overview). Choose **Edit > Inactivate selected region** or press the  button in the toolbar of the *Alignment* panel. A shortcut is the - (minus) key on the keyboard.

3.4.3.64 To mark a selected zone as active, choose **Edit > Activate selected region** or press the  button in the toolbar of the *Alignment* panel. A shortcut is the + (plus) key on the keyboard.


3.4.3.65 It is also possible to extend a sequence that has been trimmed off too far. To do so, select the outermost base on the sequence and **Edit > Extend sequence** (CTRL+X). An input box will ask you to enter the number of bases to extend.

3.4.3.66 A region on an individual sequence or on the consensus can be selected as explained in 3.4.3.54, and can be copied to the clipboard using **Edit > Copy**.

3.4.3.67 The entire sequence on which the cursor stands, or the entire consensus, can be selected with **Edit > Select all**.

3.4.3.68 A selected sequence can be removed from a contig with **Edit > Remove selected sequence** or by pressing the  button.

3.4.3.69 A consensus sequence and its associated alignment can be removed by selecting it in the *Alignment list*

panel and choosing **Assembly > Delete selected contig** or by pressing the  button.


3.4.3.70 All alignments and consensus sequences can be removed with **Assembly > Delete all contigs**.

The latter two options can be useful if you want to load stored templates (see further), remove vectors (3.4.3.30) or change the quality assignment parameters (3.4.3.18). Those actions cannot be performed if an alignment is present.

3.4.3.71 The overview panel of a contig project can be printed with **File > Print overview**.

F. Advanced alignment editing using the Dot plot window

Using a dot plot, regions of homology between two sequences are displayed graphically. To allow the dot plot to display the homology between very long sequences in an efficient way, three reduction factors will be applied: (1) bases are grouped together into *words* of a specific length, (2) a minimum number of bases should match before the match is displayed on the dot plot, and (3) the entire plot is reduced in size.

3.4.3.72 The *Dot plot* window is called with **Assembly > View dot plot** or by pressing the  button.

3.4.3.73 The *Dot plot parameters* dialog box that appears (Figure 3-68) prompts to enter the parameters for **Word size**, **Minimum score**, and **Reduction factor**. The values to enter depend strongly on the size of the project.

3.4.3.74 When pressing <OK> the *Dot plot* window appears (Figure 3-67). In this window, each consensus sequence is represented as one gray square. Repeats found within a consensus are shown within the gray squares; whereas repeats found between the consensus sequences are shown in the rectangles that form the intersections between the consensus sequences. The upper left part of the window displays the *direct repeats* (in green), whereas the lower left part of the window displays the *inverted repeats* (in blue).

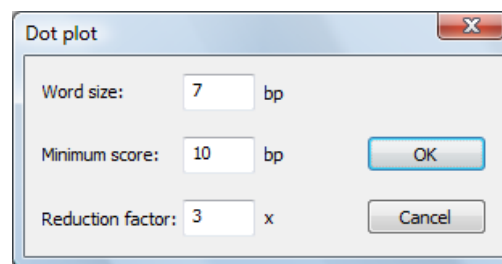


Figure 3-68. The *Dot plot parameters* dialog box.

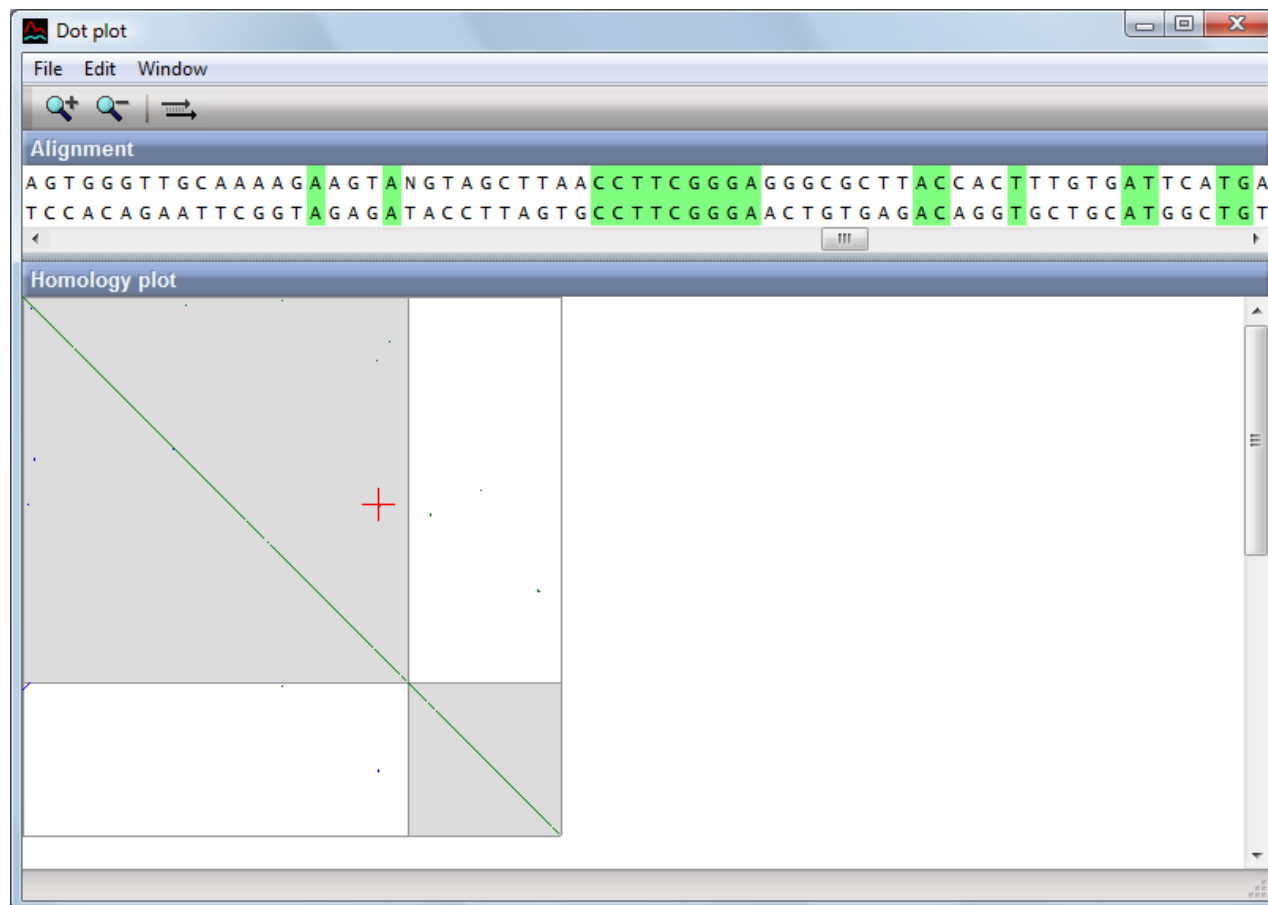





Figure 3-67. The *Dot plot* window.

3.4.3.75 You can zoom in or out on the plot using *Edit > Zoom in* and *Edit > Zoom out*, or by using the zoom buttons  and  in the toolbar.

A repeat (direct or inverted) can only be considered interesting in a contig project if it extends from a vertical side of a rectangle to a horizontal size: only then there is a complete overlapping end between consensus sequences, which can thus be merged.


3.4.3.76 Inside the dot plot window, you can click on a particular dot or stretch of dots. A red cursor appears, and the upper panel displays the matching region between the two sequences, matching bases on a green background.

3.4.3.77 To merge two sequences that have a terminal match, select *Edit > Merge contigs* or press . The consensus sequences are now merged in the *Assembler main* window and the *Dot plot* window is updated accordingly.

G. Approving and storing a contig project

A contig project can be marked as being approved or not. When working in a connected database, the user

can specify to display the status of the contig projects (approved or not) in the *Experiment presence* panel (see 2.3.3). For approved sequences, the colored dot indicating experiment presence is surrounded by a square of contrasting color, whereas for non-approved sequences the dot appears in a transparent square. The same squares are also indicated on the sequence experiment card (Figure 3-58). Sequences can be marked approved or non-approved with the *File > Approved* command.

3.4.3.78 When the aligned sequences are ready for importing in the sequence database, select *File > Save* (CTRL+S), or press the  button.

In case the program could not align the trace sequences to one single consensus, the different contig sequences will be saved into one sequence, separated by a vertical slash (/). They will be saved in the same order as they appear on the screen.

3.4.3.79 The order of the contigs can be changed by selecting a contig in the *Alignment list* panel and using *Assembly > Move contig up* or *Assembly > Move contig down*.

3.4.3.80 From a sequence in InfoQuest FP assembled using the Assembler tool, the project can be opened in

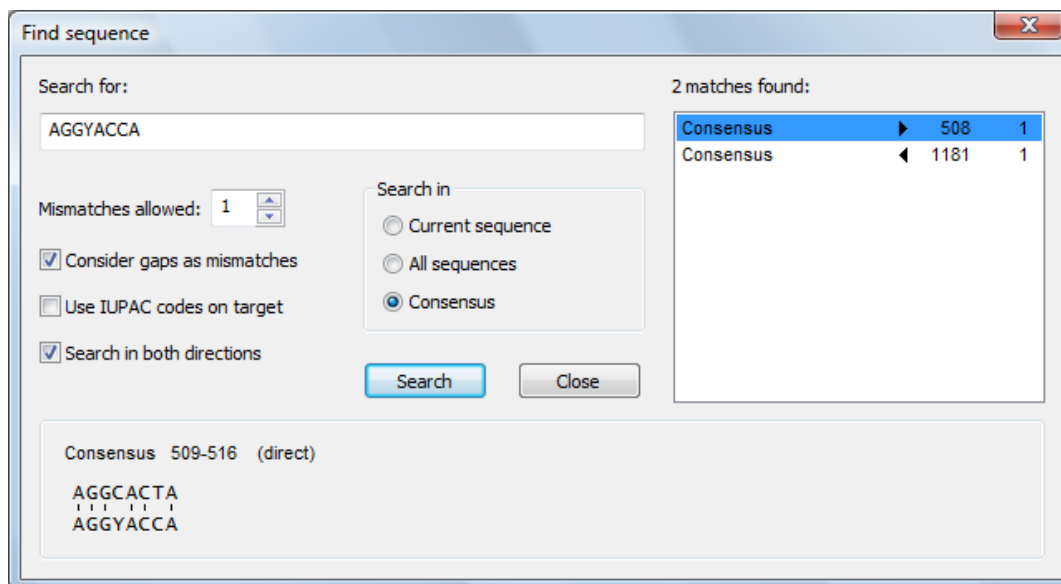




Figure 3-69. The *Find sequence* dialog box.

Assembler by pressing the  button in the small sequence edit box opened from the *Entry edit* window, or in the Kodon Sequence Editor if Kodon is linked to InfoQuest FP. Such projects can be changed at any time and are updated automatically in the InfoQuest FP database.

H. Finding subsequences

3.4.3.81 With *Edit > Find* or by clicking  in the toolbar (shortcut: CTRL+F) you can pop up a *Find sequence* tool in Assembler (Figure 3-69) to find subsequences. You can fill in a subsequence including unresolved positions according to the IUPAC code.

Under *Search in*, you can choose between *Current sequence* (the selected one), *All sequences*, and *Consensus*.

Using *Mismatches allowed*, it is possible to find subsequences that differ in a defined number of bases from the entered string.

The check box *Consider gaps as mismatches*, allows the search algorithm to introduce gaps in either the search sequence or the target sequence to match them. Gaps are considered in the same way as mismatches, and thus depend on the *Mismatches allowed* setting.

Use IUPAC codes on target allows the search sequence to be matched with uncertain positions denoted as IUPAC unresolved positions (e.g. "R", "Y", etc., including "N").


With *Search in both directions* enabled, the invert-complemented sequence will be searched through as well.

3.4.3.82 Press **<Search>** to execute the search command. The result set displays all the instances that were found (Figure 3-69), indicating with arrows if they have been found on the sequence as is, or after invert-complementing. The positions are also indicated.

3.4.3.83 If you click on an item in the result set, the matching subsequence is selected in the *Alignment* panel (central panel). The bottom panel of the *Find sequence* window displays the alignment of the search sequence and the target sequence, indicating mismatches and gaps introduced (if allowed).

I. Trimming the consensus

The purpose of this tool is to locate two fixed subsequences on the consensus to define the start and end position, respectively. One can choose to include or exclude the locator sequences in the final consensus. In many cases, but not always, these subsequences will correspond to primers used. For generality, the subsequences are called *trimming targets* in the program and in the description that follows.

3.4.3.84 Select *Assembly > Consensus trimming* or press the  button to open the *Consensus trimming* dialog box (Figure 3-70).

Under **Trimming targets**, you can fill in a *Start pattern* and an *End pattern*. For both the start and end patterns, you can specify *Mismatches allowed*, and fill in a *Target range* on the consensus. The latter is to restrict the search to certain regions on the consensus, e.g. to prevent incidental matches inside the targeted consensus sequence.

With *Search both directions*, the entered trimming targets will be searched for on the consensus as it appears as well as on its complementary strand. In case

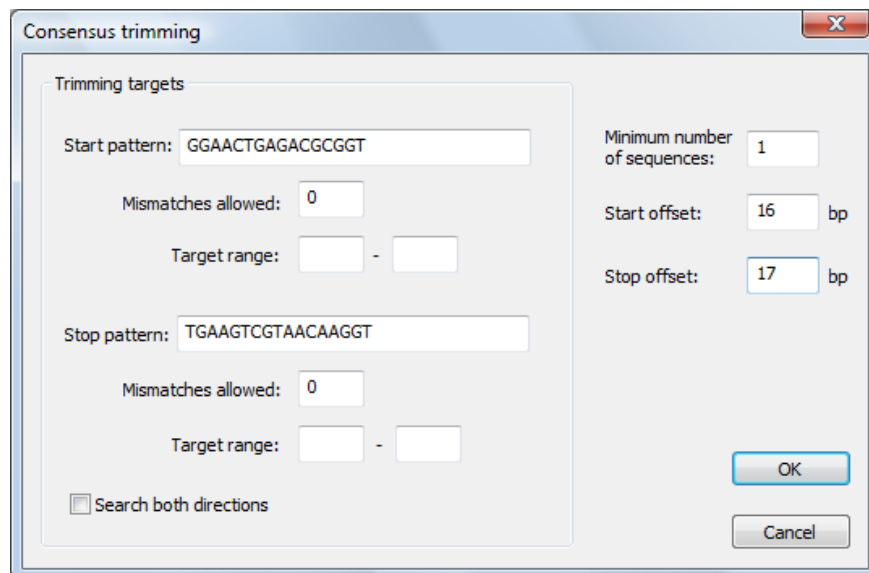


Figure 3-70. The *Consensus trimming* dialog box.

the trimming targets match the complementary strand of the consensus, it will be automatically invert-complemented.


Minimum number of sequences specifies a minimum number of trace sequences that should be contributing to the subsequence on the consensus that matches the trimming targets. For example, if 2 is entered, a trimming target will only be set if the matching region on the consensus is *fully* defined by at least 2 sequences.

With *Start offset* and *End offset*, one can specify that the consensus is trimmed at a certain offset from the start and end trimming target positions, respectively. If no offset is specified (zero), the trimming targets are **included** in the trimmed consensus.

3.4.3.85 When the trimming targets have been set by pressing the **<OK>** button in the *Consensus trimming* dialog box, the *Alignment overview* panel shows black hatched lines at the positions of the trimming targets. Likewise, the consensus sequence in the alignment panel is grayed where it is trimmed off.

J. Storing and using assembly templates

The Assembler program automatically stores all user defined settings from the last saved project in a template called *DefaultSettings* (see Figure 3-71). These settings include the display settings, the quality assignment parameters, the vectors to remove and their parameters, the alignment parameters, the consensus determination parameters, the consensus trimming targets and their parameters. When opening a new project, these settings are automatically applied to the new project. In addition to the *DefaultSettings* template, other templates can be stored.

3.4.3.86 Select **File > Templates** or . This will open the *Templates* dialog box (Figure 3-71) which allows the current template to be saved with **<Save current>**, or a selected template from the list (left) to be loaded with **<Load template>**. A selected template can be deleted

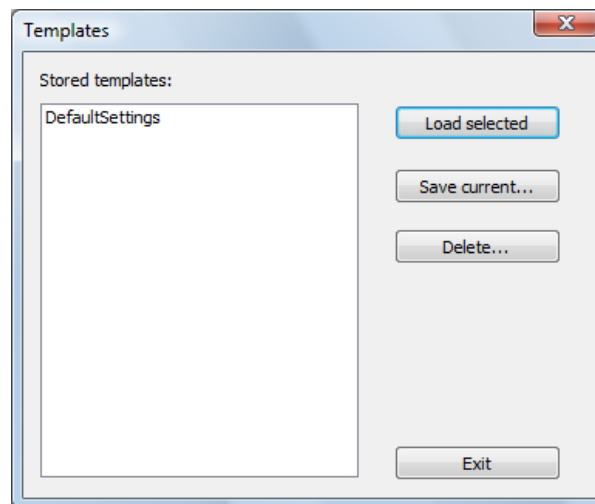



Figure 3-71. The *Templates* dialog box.

with **<Delete selected>**.

NOTE: A template can only be loaded if no alignment is present. To load a template, you will need to remove the assemblies first, which can be done with **Assembly > Delete all contigs** (see 3.4.3.70).


The sequence experiment card is filled with the assembled sequence as soon as the project is saved.

3.4.3.87 From a sequence in InfoQuest FP assembled using Assembler, the project can be opened in Assem-

bler by pressing the  button in the sequence experiment card (3.4.3.4). The base selected in the experiment card will automatically be selected in the Assembler editor. Such projects can be changed at any time and are updated automatically in the InfoQuest FP database.

3.4.3.88 From within a InfoQuest FP comparison, you can double-click on a base of a sequence, which pops up

the sequence experiment card with that base selected.

Pressing the  button in the sequence experiment card in turn launches Assembler with the same base selected.

3.4.3.89 When finished, exit the window with *File > Exit*.

3.5 Setting up trend data type experiments

3.5.1 Introduction

Reactions to certain substrates or conditions are sometimes recorded as multiple readings in function of time, as *kinetic* readings. The kinetic reading of enzymatic or metabolic activity is thought to be both more informative and more reliable than measuring the degree of activity at one point in time. Examples are the kinetic analysis of metabolic and enzymatic activity, real-time PCR, or time-course experiments using microarrays. Although multiple readings per experiment are mostly done in function of time, they can also depend on another factor, e.g. measurements in function of different concentrations.

These different data types have in common that they measure a trend of one parameter in function of another. We therefore call them *trend type* data. Analysis is usually done by fitting a *curve* through the measurement points using a *fit model*. To use a fit model, we have to assume that the biological data that are being studied behaves according to a certain predictable pattern. A model is a function that fits the biological data as closely as possible. Specific parameters can be deduced from the model function, and therefore, comparing the samples is done using the parameters of the curves rather than the original measurement points. Bacterial growth or activity is usually analyzed using a *Logistic Growth* fit. A number of parameters can be calculated from the curve fit (Figure 3-72), including the times at 5% growth increase (T_{05}), 50% growth increase (T_{50}) and 95% growth increase (T_{95}), the maximum slope (S_{max}), the time at maximum slope ($T_{S_{max}}$), the initial value (MIN), the final value (MAX), the initial exponential growth rate (r), and the initial doubling time (T_{doubl}).

Depending on the data type, other fit models may be used, such as linear, logarithmic, exponential, hyperbolic, Gaussian, Gompertz, power function, etc., each resulting into specific parameters that describe the fit.

In InfoQuest FP, analysis and comparison of curve type data can be done on one or more parameters derived from the curve fit. For example, if one uses S_{max} and MAX, each curve is translated into two character values. Figure 3-73 illustrates in a schematic way how a hypothetical test panel (in the example containing 6 tests) is processed into a data matrix in InfoQuest FP. Each test results in 5 readings (1), through which a curve is fit, using an appropriate model (2). The *Logistic Growth*

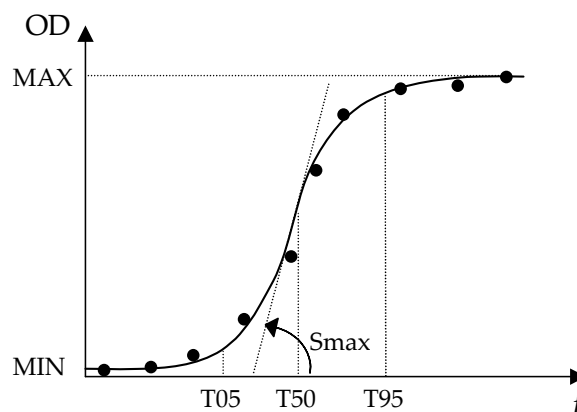



Figure 3-72. Trend curve that follows the logistic growth model and some derived parameters. T_{05} , T_{50} and T_{95} are the times at 5%, 50% and 95% growth increase, respectively. S_{max} is the maximum slope, MIN is the initial value, and MAX is the final value.

model is used in the example. For a given model, one or more characteristic parameters can be derived from the curves. In the example, the *maximum slope* S_{max} and the *final value* MAX are calculated (3). This leads to two data matrices, each containing one value per test and per organism or sample (4).

For taxonomy or typing purposes, one might be interested in combining the data from multiple parameters into one clustering or identification. In InfoQuest FP, it is possible to specify a comparison coefficient for each used parameter separately. The software then averages the respective similarity values into one similarity value per pair of entries compared. An important issue is that the parameters used can have different ranges, as is the case in the example in Figure 3-73. If a coefficient is chosen that has no inherent *scaling*, e.g. Euclidean distance, an appropriate range should be specified for each parameter, so that the weights of the different parameters are standardized when they are combined by averaging.

3.5.2 Defining a new trend data type

3.5.2.1 Select *Experiments* > *Create new trend data type*

from the main menu, or press the  button in the *Experiments* panel toolbar and select *New trend data type*.

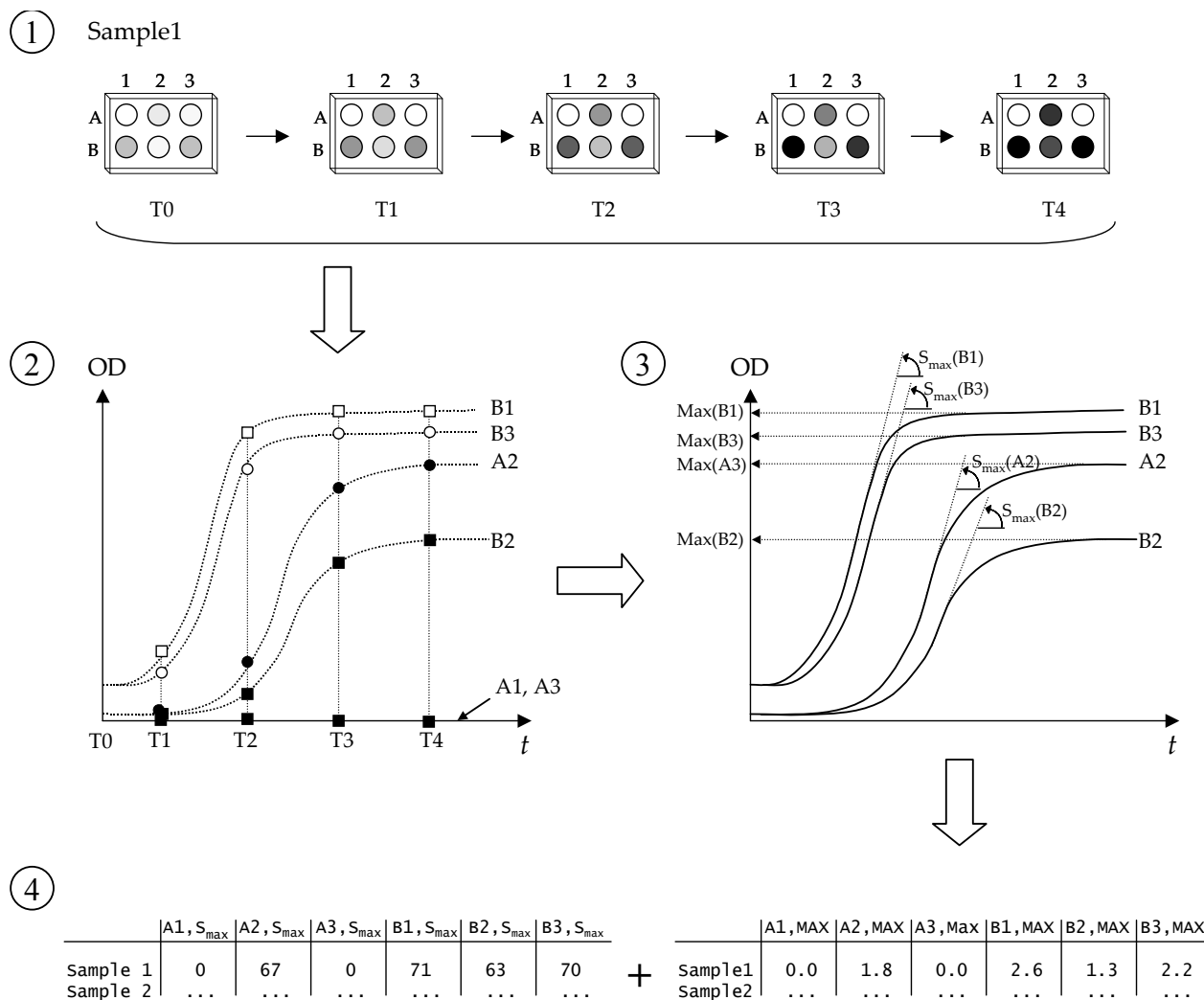


Figure 3-73. Example of the processing of kinetic readings of a phenotypic test panel. (1) Readings are done at different times T0...T4; (2) A curve model is fit through the values obtained for each well in the test panel (in the example, *Logistic Growth*); (3) One or more specific parameters are derived from the curves (in the example, the *final value MAX* and the *maximum slope S_{max}*); (4) A data matrix is constructed from a curve parameter obtained for each well, including all the samples analysed. In the example, two data matrices are generated as two parameters were chosen.

3.5.2.2 The *New trend data type wizard* prompts you to enter a name for the new data type. Enter a name and press the **<Finish>** button to complete the setup of the new trend data type. It is now listed in the *Experiments* panel.

Once the trend data type is created, one has to define a *trend curve* for each character that results into multiple readings. Figure 3-73 illustrates an example of an experiment consisting of 6 characters so that 6 trend curves will be calculated. In reality, a trend data experiment might as well exist of 96 characters, for example if a microtiter plate with 96 tests is read as kinetic data.

As an example we will create a trend data experiment consisting of 6 tests, as depicted in Figure 3-73.

3.5.2.3 Open the *Trend data type* window by double-clicking on the name in the *Experiments* panel. The window looks like in Figure 3-74, initially empty.

3.5.2.4 Select *TrendCurves > Add new trend curve*. Enter a name, for example "**Blank**", and press **<OK>**. The new trend curve appears in the *Curves* panel of the *Trend data type* window (Figure 3-74).

In the same way, you can enter more trend curves. Before any analysis can be done, we will have to define the parameters, derived from the appropriate model curves, which we want to use for the analysis and comparison.

3.5.2.5 Select *Parameters > Model parameters*. The *Trend curve parameters* dialog box that appears (Figure 3-75) lists the available models in the left panel.

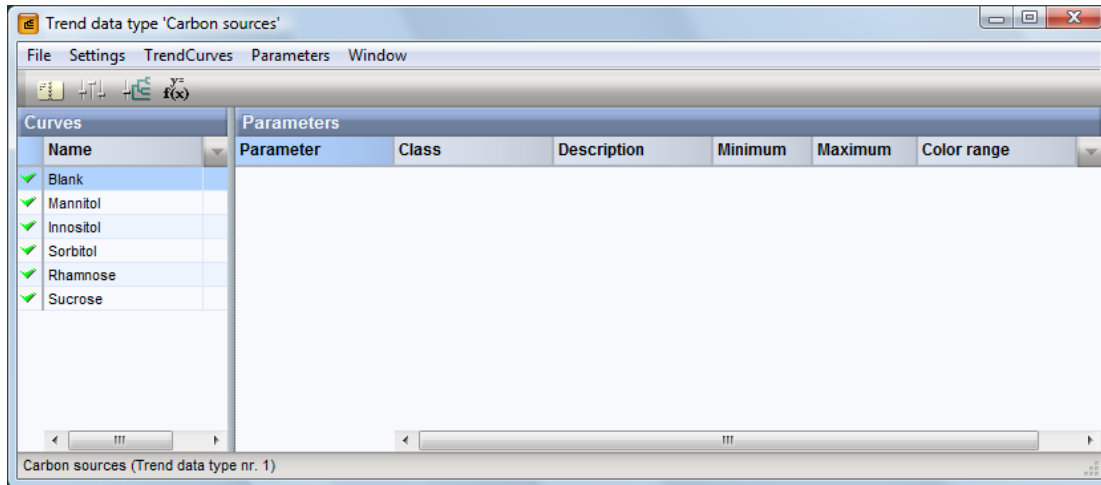


Figure 3-74. The *Trend data type* window, with 6 trend curves defined.

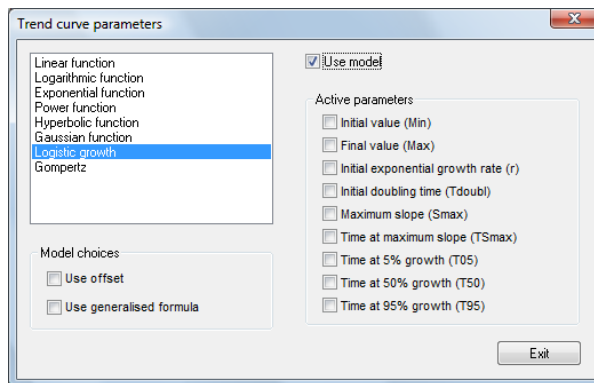


Figure 3-75. *Trend curve parameters* dialog box to select the models to use and the associated parameters to include.

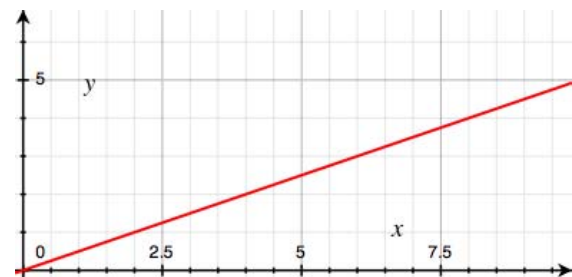
3.5.2.6 Select a model from the list, e.g. *Logistic growth*. A check box *Use model* allows the model to be included or not. If a model is selected, a number of parameters, listed under *Active parameters*, can be chosen to include in the analysis. Additionally, one or more choices (*Model choices*) can be specified for each model.

Below is a list of the models that are available and their parameters and model choices:

• **Linear function**

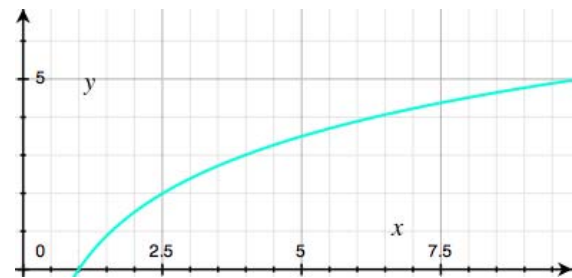
$$y = A + Bx$$

Available parameters are the *Intercept* A and the *Slope* B. The function can be forced to pass through zero, in which case the intercept A is always zero.



• **Logarithmic function**

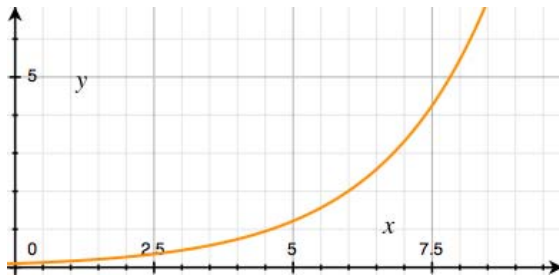
$$y = A + B \log x$$



Similar as for a linear function, the available parameters are the *Intercept* A and the *Slope* B. The function can be forced to pass through zero, in which case the intercept A is always zero.

• Exponential function

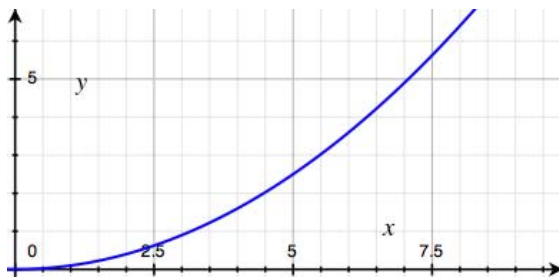
$$y = O + Ae^{Bx}$$



The function offers the *Amplitude* A and the *Exponential* r (exp) as parameters. If the model choice *Use offset* is checked, the *Offset* O is a parameter as well.

• Power function

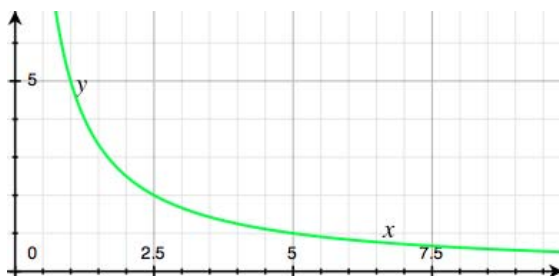
$$y = O + Ax^B$$



The function offers the *Amplitude* A and the *Power* B as parameters. If the model choice *Use offset* is checked, the *Offset* O (on the x-axis) is a parameter as well.

• Hyperbolic function

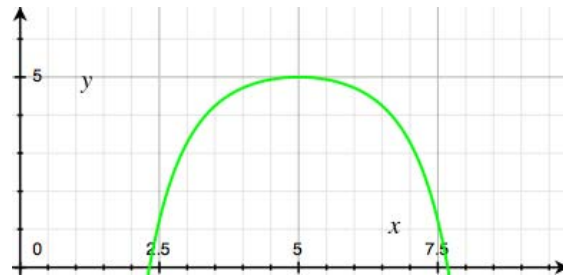
$$y = A + \frac{B}{x - C}$$



This model offers the *Offset* A and the *Amplitude* B as parameters. As a choice, an asymptote can be fitted with $C <> 0$ (*Fit asymptote*).

• Gaussian function

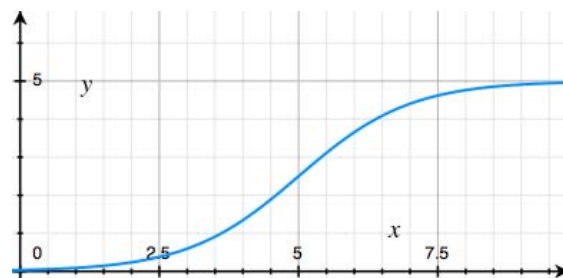
$$y = A + Ce^{-e^{B(x-M)}}$$



The Gaussian model offers the *Amplitude* A, the *Position of the center* M, the *Width of the Gaussian* S and the *Offset* O as parameters.

• Logistic growth

$$y = A + \frac{C}{[1 + e^{Q-B(x-M)}]^{1/Q}}$$



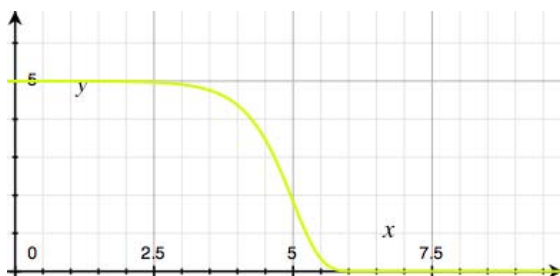
Following are the parameters for logistic growth:

- The *Initial value*, i.e. the minimum value derived from the curve
- The *Final value*, i.e. the maximum value derived from the curve
- The *Initial exponential growth rate* r
- The *Initial doubling time*, which is the time needed for y to double
- The *Maximum slope*: the maximum growth rate of y
- The *Time at maximum slope*, i.e. the x-value at maximum slope
- *Time at 5%, 50% and 95% growth* are the x values at 5%, 50% and 95% growth of the y value, respectively.

If the model choice *Use offset* is not checked, the A value becomes zero in all cases. If *Use generalized formula* is not checked, the value Q becomes zero in all cases.

• **Gompertz function.**

$$y = A + \frac{C}{e^{e^{B(x-M)}}}$$



Following are the parameters for Gompertz:

- The *Initial value*, i.e. the minimum value derived from the curve
- The *Final value*, i.e. the maximum value derived from the curve
- The *Maximum slope*: the maximum growth rate of y
- The *Time at maximum slope*, i.e. the x -value at maximum slope
- *Time at 5%, 50% and 95% growth* are the x values at 5%, 50% and 95% growth of the y value, respectively.

If the model choice *Use offset* is not checked, the A value becomes zero in all cases.

3.5.2.7 In the *Trend curve parameters* dialog box (Figure 3-75), check *Use model* for *Logistic growth*.

3.5.2.8 Check *Final value (Max)* and *Maximum slope (Smax)* as parameters to include.

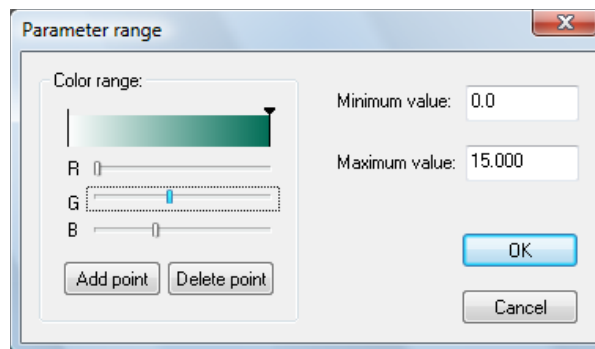


Figure 3-77. The *Parameter range* dialog box.

NOTE: Additional parameters can be picked up for analysis from the same or other models as well.

3.5.2.9 Press *<Exit>* to exit the parameter choice dialog box.

The *Curves* panel of the *Trend data type* window now contains the selected curve model parameters to be used for comparison (Figure 3-76).

3.5.2.10 For a selected parameter, the range can be specified with *Parameters > Change parameter range*.

The *Parameter range* dialog box (Figure 3-77) allows the *Color range* to be changed and a *Minimum value* and *Maximum value* to be specified.

3.5.2.11 Under *Color range*, the left and right ends of the color scale can be selected and a color can be assigned. R, G, and B stand for the red, green and blue component, respectively.

3.5.2.12 Using *<Add point>*, intermediate nodes can be added and assigned a color.

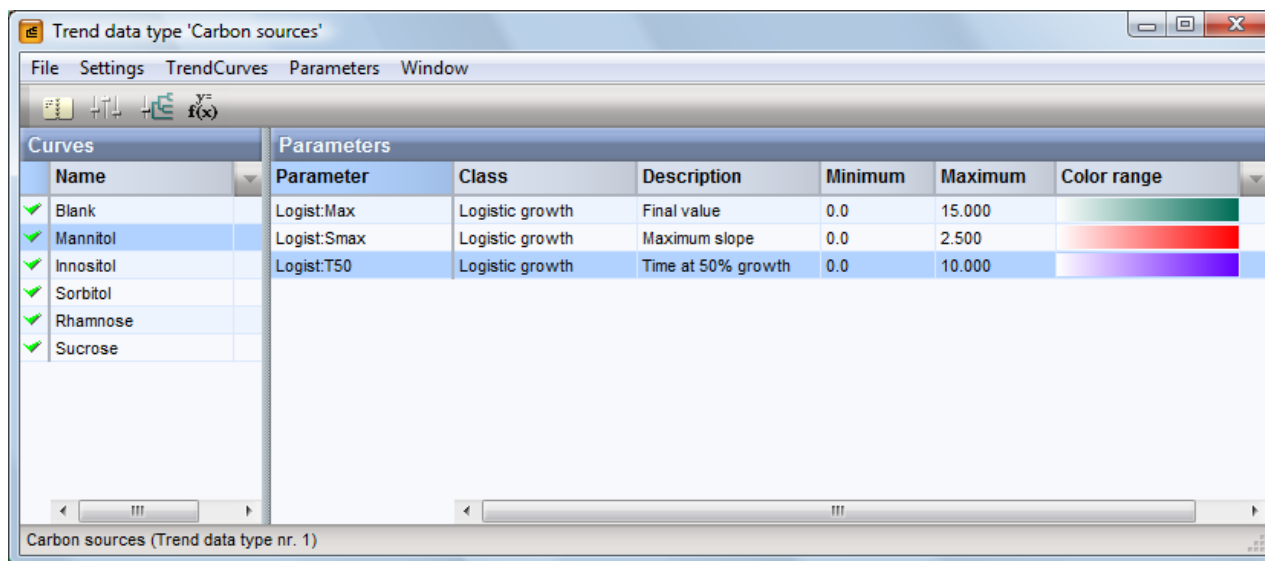


Figure 3-76. The *Trend data type* window for the example data set with 3 model parameters defined.


The color range specified will be used in the *Image* panel of the *Comparison* window (see 4.1.3).

The maximum and minimum values (range) of the parameter are important if the Euclidean distance coefficient is chosen, which has no inherent scaling. Using the ranges, the data values are normalized so that distance values from different parameters have equal weights.

3.5.3 Entering trend data in the database

Trend data type experiments usually generate quite massive data sets. Therefore, trend data cannot be entered using the keyboard. Since trend data can be stored in many different ways and formats, they can only be entered through scripts written in the InfoQuest FP script language (see the Script Manual).

One generic script for reading trend data from tabular formats is part of the Import plugin and can be installed as follows:

3.5.3.1 Select **File > Install/Remove plugins** in the *InfoQuest FP main* window, or click the  button.

3.5.3.2 Select **Import tools** in the *Plugins* window and press the **<Install>** button. If the plugin is installed correctly, the plugin is marked with a green check mark in the *Plugin installation* toolbox.

3.5.3.3 Close the *Plugin installation* toolbox.

A set of new menu items is now available under **File > Import**.

The trend data import script reads text files that contain a table of tab-delimited strings and values. The table should contain the trend data in the following format:

	Curve 1	Curve 2	...
X value 1	Y value	Y value	...
X value 2	Y value	Y value	...
...

The table can occur anywhere in the file; the user can select it in the script. Furthermore, certain rows and/or columns can be selected or unselected from the table. However, the first selected row (green) should be the header row describing the curve (character) names, and the first selected column (yellow) should be the X values. Each next column should contain the Y values for the curve named in the column header.

The script can create a new trend data type if required, and will automatically add all trend curves found in the data files.

Some 6 example data files (artificial data) are provided in the **Sample and Tutorial data\Trend data sample files** directory on the installation CD-ROM. The same files are also available from the download page of the website (www.bio-rad.com/softwaredownloads).

3.5.3.4 Launch the script after installation of the Plugin tools, by selecting **File > Import > Import trend data** in the *InfoQuest FP main* window.

3.5.3.5 Create a new trend data type by entering **Sugar metabolism** under **Create new**.

3.5.3.6 Press **<Create>** to create the new trend data type. It becomes automatically selected in the left drop-down list box.

3.5.3.7 Press **<Select files>** to select the sample trend data files (**BSU1072.txt**, etc.) on the CD-ROM or from the downloaded and unzipped folder from the website.

3.5.3.8 Press **<Proceed>**. The first file is now parsed as follows:

- A key is suggested from the file name. It can be changed if desired.
- The full file content is displayed in table format, and the data table located inside the file is selected in a multi-column list box.

In this example it is not necessary to change anything to the selection, but it is possible to unselect or select rows or ranges of rows with the CTRL and SHIFT keys, respectively. The first selected row should contain the curve names.

3.5.3.9 Press **<Next>** to proceed to the next step.

In this step, the columns that will be used can be selected. By default, all columns are selected. The first column should always contain the X-values, in the examples a measurement time.

3.5.3.10 Press **<Next>** again to proceed to the third and last step, which gives an overview of the parsed table. The column headers should contain the curve names, and the first column should contain the X values, i.e. the time in hours.


3.5.3.11 Press **<Save data>** to write the first experiment into the new trend data type. Trend curves with the names Mannitol, Inositol, etc. are automatically generated.

The same wizard will be repeated for each file selected. If all files are imported, the entries will show up in the database with a colored dot in the *Experiment presence* panel for the trend data type **Sugar metabolism**.

3.5.4 Displaying trend data

For visualization and comparison purposes, a default *curve fit model* will have to be chosen for a particular trend data type.

3.5.4.1 Open a *Trend data type* window, for example **Sugar metabolism** containing the data entered in Section 3.5.3.

3.5.4.2 A default curve model can be chosen for the trend data type by selecting **Settings > Default trend curve model** or by clicking on  in the toolbar.

The dialog box that pops up (Figure 3-78) lists the available models and regressions (left) and their additional parameters (right).

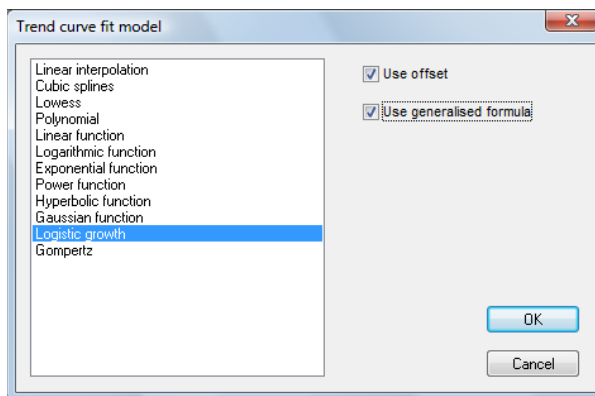



Figure 3-78. Dialog box to select the default trend curve model. For a number of models, additional parameters can be specified.

3.5.4.3 Choose **Logistic growth** and check both **Use offset** and **Use generalized formula** (see 3.5.2.6 for explanation). Press **<OK>** to set the parameters.

NOTE: This choice determines the fit model used for visualization of trend curves, and also the fit model used for curve comparisons, in case the fit model is used instead of the raw data values (see 4.5.1).

3.5.4.4 With **Settings > General settings** or via  from the toolbar, additional visualization settings can be specified: **Include zero in X axis** and **Include zero in Y axis**. If these settings are checked (enabled), the zero on the X axis and the Y axis, respectively, will always be shown on the plot, irrespective of the ranges of the components.

3.5.4.5 If you click on the colored dot in the *Experiment presence* panel that represents the trend data type for a particular entry, the curves for the selected entry are displayed in a *Trend curve* box (Figure 3-79).

The name of the entry to which the curves belong (i.e. the key) is written in the status bar. The box can be resized in the bottom right corner. The box can be

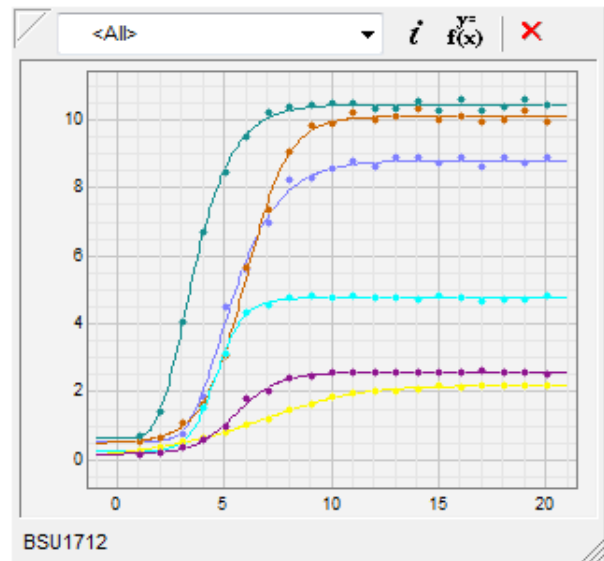



Figure 3-79. Trend curves box, showing curves for one entry.


moved by clicking and holding down the left mouse button anywhere inside the borders.


Upper left in the *Trend curves* box is a pull-down list box where you can choose which curves to display. The default setting is **<All>**.

3.5.4.6 You can select any particular curve by clicking inside the list box and selecting one of the curves present in the data type.

3.5.4.7 Using the  button, you can toggle between the *curve view* as depicted in Figure 3-79 and the *info view*, which contains detailed information about:

- The fit model chosen for visualization (see 3.5.4.2): the standard deviation and the parameters derived from the formulas (see 3.2.1) are indicated.
- The curve parameters selected for comparison (see 3.2.1).

3.5.4.8 With the  button, you can choose another regression or curve fit for the present set of curves. This choice only applies to the currently open *Trend curves* box, and will not influence the default curve fit as explained in 3.5.4.2.

3.5.4.9 Using the button , it is possible to remove the curves for the selected entry from the database.

3.5.4.10 The *Trend curves* box can be closed by clicking in the upper left corner (triangular button).

To compare curves between different entries, trend curves can be displayed for multiple entries at a time in the same window. This is achieved as follows.

3.5.4.11 Select a number of entries in the database for which trend curves are present, by holding the CTRL key and left-clicking (see 2.2.7 for manual selection functions). Selected entries are marked with a colored arrow.

3.5.4.12 Open the *Trend data type* window (Figure 3-74) and select *File > Create trend data window*.

The resulting *Trend data* window (Figure 3-80) displays the curves for all the selected entries in a single plot.

3.5.4.13 Inside the window, a legend shows the colors and the names of the corresponding curves. This legend is a box that can be moved inside the *Trend data* window.

3.5.4.14 Similar as in the *Trend curves* box (Figure 3-79), a pull-down list box allows either all curves or one particular curve to be displayed.

3.5.4.15 With one curve per entry displayed, all curves have the same color and the legend box is redundant. Using the *View* menu, however, one can change the type of labeling into *Label by entry*.

The curves now have different colors according to the entries, as indicated in the legend box.

3.5.4.16 The entries can also be queried interactively from this window:

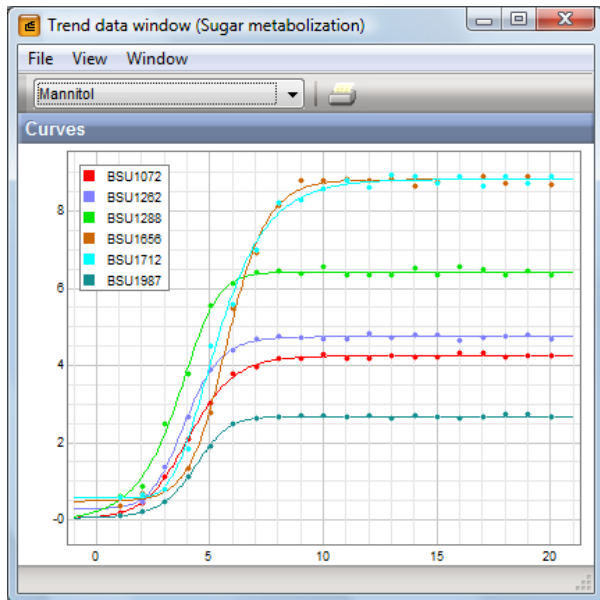


Figure 3-80. The *Trend data* window, displaying trend curves for multiple entries.


- By double-clicking on a dot (data point) of a curve, the *Entry edit* window of the entry to which the curve belongs will pop up.

- By holding down the CTRL key and clicking on a dot, the entry will be selected or unselected in the database.

3.5.4.17 The selection status of the entries can also be seen from the *Trend data* window by choosing *View > Label by selection*. Selected entries are shown in blue, unselected entries in black.

3.5.4.18 With *View > Use colors*, one can toggle between the color view and a black-and-white view, in which the data points of the different curves are represented by different symbols such as circles, squares, and triangles.

In the *Label by selection* view in black-and-white, selected entries are represented by a filled circle, whereas non-selected entries are represented by an open circle.

3.5.4.19 The image can be copied to the clipboard with *File > Copy to clipboard*, or printed directly with *File > Print* or .

3.5.5 Additional comparison parameters


In the *Trend data type* window (Figure 3-76), it is possible to define additional, non model-based comparison parameters.

3.5.5.1 With *Parameters > Statistics parameters*, the mean value and standard deviation of the X and Y component can be added as comparison parameters.

3.5.5.2 In addition, with *Parameters > Add value*, a Y value can be included that corresponds to a fixed X value. An input box prompts you to enter an X value.

3.5.5.3 Similarly, a slope can be calculated that corresponds to a fixed X value with *Parameters > Add slope*. An input box prompts you to enter an X value.


3.5.6 Comparison settings

The comparison settings for a trend data type can be accessed with *Settings > Comparison settings* or the  button, but also in the *Comparison* window. See Section 4.5 for a detailed explanation.

3.6 Setting up matrix type experiments

3.6.1 Defining a new matrix type

3.6.1.1 Select *Experiments* > *Create new matrix type*

from the main menu, or press the  button in the *Experiments* panel and select *New matrix type*.

3.6.1.2 Enter a name for the new type. Enter a name, for example **DNA-homol**.

3.6.1.3 Press the <OK> button to complete the setup of the new matrix type. The new matrix type is now listed in the *Experiments* panel.

Unlike other experiments, a matrix type does not provide an experiment for each entry. Instead, it contains *similarities between entries*. Hence, the “data file” which contains the experiment data, and the “entry file” which links the experiments to database entries are the same here. There are two ways to enter similarity values: by importing a matrix as a whole, and by entering the values from the keyboard.

To import a matrix, it must have the following format:

```
ENTRY KEY<tab>VALUE<eol>
```

```
ENTRY KEY<tab>VALUE<tab>VALUE<eol>
```

```
ENTRYKEY<tab>VALUE<tab>VALUE<tab>VALUE
<eol>
```

Etc.

<eol> means “end of line”, a simple return in MS-DOS text, which corresponds to ASCII character #13 followed by ASCII character #10.

Matrix files can be imported by selecting the matrix type in the *Experiments* panel, and *File* > *Import experiment file*.

The program compares the entry keys as provided in the import file with the entry keys in the database and assigns values to the corresponding keys. If entry keys are not found in the database, it will automatically create new database entries.

To enter similarity values manually, you first have to select the entries in the database for which you want to create a matrix.

3.6.1.4 Select some entries in the database by holding the CTRL key and left-clicking (see 2.2.7 for manual selection functions). Selected entries are marked with a colored arrow.

3.6.1.5 Double-click on the file **DNA homol** in the *Files* panel (not in the *Experiments* panel). This opens the *Matrix file* window (Figure 3-81).

The diagonals, i.e. the similarity values of the entries with themselves, are filled in already and cannot be changed.

3.6.1.6 To enter a value, press Enter or double-click on a field.

3.6.1.7 When finished, exit the window with *File* > *Exit*.

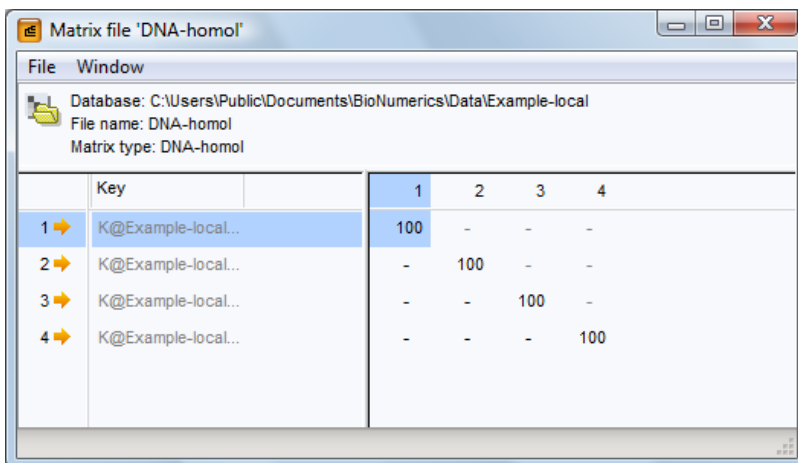


Figure 3-81. The *Matrix file* window to enter and edit similarity values.

NOTE: Importing matrix files according to the procedure described above and entering matrix data manually is only possible when working in a local

database. Import of matrix files in a connected database is done via a script (see 1.1.6). Contact Bio-Rad to obtain this script.

3.7 Setting up composite data sets


3.7.1 Introduction

A composite data set is a character table that contains all the characters of one or more experiment types. It is a “container” type of experiment type in InfoQuest FP, i.e. it holds the data coming from one or several other experiments, but it does not necessarily correspond to an actual physical experiment.

In addition to the obvious reason of creating a clustering based on multiple data sets, a composite data set also offers some additional interesting features compared to single character types. These include a function to discriminate groups based upon differential characters in the *Comparison* window (*Composite > Discriminative characters*) and a function to perform transversal clustering, i.e. based on the characters (*Composite > Calculate clustering of characters*). These functions will be discussed in . Performing bootstrap analysis (a cluster significance tool, see 4.1.13) on character type data is only possible via a composite data set. Lastly, composite data sets are used for creating band matching tables, a feature that will be discussed in .

3.7.2 Defining a new composite data set

We will now describe the setup of composite data sets in function of cluster analysis based upon multiple experiments. As an example, we will create a character table for all phenotypic tests defined in the **DemoBase** database.

3.7.2.1 In the *InfoQuest FP* main window, with the database **DemoBase** loaded, select *Experiments > Create new composite data set*, or press the  button in the *Experiments* panel toolbar and select *New composite data set*.

3.7.2.2 Enter a name, for example **All-pheno** and press **<OK>**.

The *Composite data set* window is shown for **All-Pheno** (see Figure 3-82). All experiment types defined for the database are listed, and when they are marked with a red cross, they are not selected in the composite data set.

3.7.2.3 Select **PhenoTest** from the experiment list and *Experiment > Use in composite data set*. Repeat this action for **FAME**.

When an experiment type is selected in the composite data set, it is marked with a green ✓ sign.

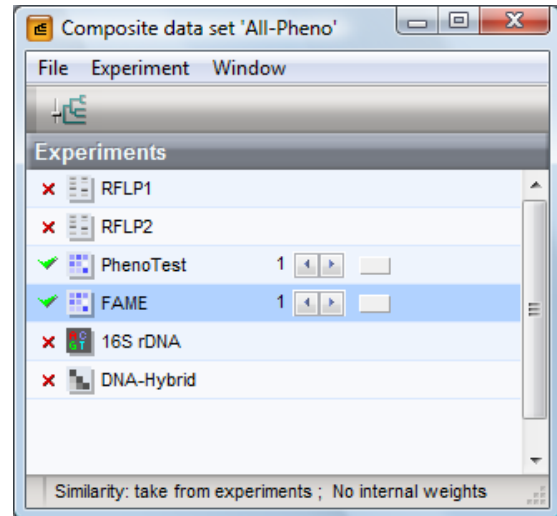


Figure 3-82. *Composite data set* window.

The scroll bar that appears in the **Weights** column allows the user to manually assign weights to each of the selected experiment types (see step 3 described in 4.7.2). If the individual matrices of the experiments are averaged to obtain a combined matrix, the similarity values will be multiplied by the weights the user has specified for each experiment.

In order to treat individual characters on an equal basis while averaging matrices, the program can automatically use weights proportional to the number of tests each experiment contains. This correction is achieved as follows:

3.7.2.4 Select *Experiment > Correct for internal weights*. The header now shows **Similarity: take from experiments; Correct for internal weights**.

NOTES:


(1) The correction for internal weights also applies to banding patterns: if technique **RFLP1** reveals 10 bands between entries A and B, whereas **RFLP2** only reveals 5 bands, the similarity value resulting from **RFLP1** will be twofold more important in averaging similarity between entries A and B.

(2) Both functions **Correct for internal weights** and the manual weight assignment can be combined. The program will then multiply the weights obtained after correction by the weights assigned by the user.

(3) In case step 4 described in 4.7.2 is chosen further in the analysis, i.e. the character sets are merged to a

combined character set to which a similarity coefficient is applied, the user defined weights also have their function: in this case, the program multiplies each character of a given experiment with the weight assigned to that experiment. This feature is useful in case the ranges of combined experiments are different; for example when one experiment has a character value range between 0 and 1 and another experiment has a range between 0 and 100, a quantitative coefficient such as the correlation coefficients, Gower, or Euclidean distance (for more information on these coefficients, see Section 4.6) would in practice only rely on the second experiment. Assigning a weight of $\times 100$ to the first

experiment makes them equally important for quantitative coefficients.

The comparison settings for the composite data set can be accessed with *Experiment > Comparison settings* or the  button, but also in the *Comparison* window. See Section 4.6 for a detailed explanation.


3.7.2.5 Close the *Composite data set* window with *File > Exit*. The new composite data set is shown in the *Experiments* panel of the *InfoQuest FP* main window.

3.8 Experiment display and edit functions

In , we have explained how you can edit the information fields for each database entry by double-clicking on the entry (2.2.3.1), which pops up the *Entry edit* window. It is possible to enter and view experiment data directly from the *Entry edit* window.

In order to explain the edit functions, we will use the **DemoBase** database:

3.8.0.1 Close the *InfoQuest FP main* window.

3.8.0.2 Back in the Startup screen, select **DemoBase** and click on  or just double-click **DemoBase** to start InfoQuest FP with this database loaded.

3.8.1 The experiment card

3.8.1.1 If we open the *Entry edit* window for any database entry (except a standard), the window lists all available experiment types for this entry, each of which contains two buttons (Figure 3-83).

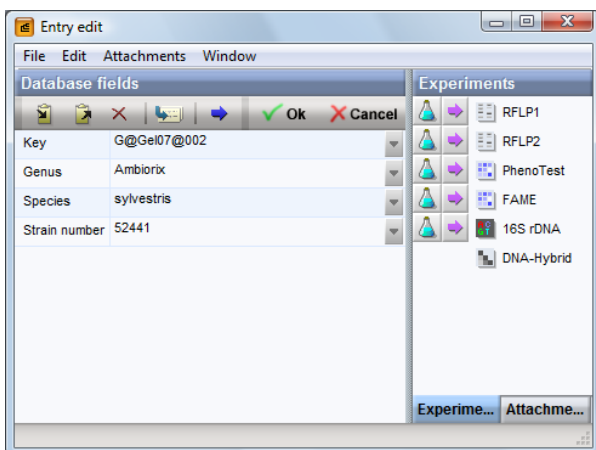



Figure 3-83. *Entry edit* window.

3.8.1.2 With the  button, you can display the *Experiment card* of an experiment (Figure 3-84).

An experiment card can also be opened from the *InfoQuest FP main* window, by clicking on the colored dot in the *Experiment presence* panel (see Figure 1-15).

3.8.1.3 You can move the experiment card by clicking and holding the left mouse button on the card, and then dragging it to its new position. For sequence experiment

cards and character experiment cards displayed as a list, move the window in the caption.

3.8.1.4 When you hover over the image card with the mouse, a small tag displays additional information. In case of a fingerprint, it shows the key of the entry, and the gel name and lane number. In case of a character type shown as a plate it shows the key of the entry, and the name and the value of the character being pointed to.

3.8.1.5 Close an experiment card by clicking in the small triangle-shaped button in the left upper corner.

You can open an experiment card for an entry, close its *Entry edit* window, and then show the corresponding experiment card for another entry, to arrange and compare them side by side. Only the screen size will be the limiting factor as to the number of experiment cards that can be shown together.

3.8.2 Gelstrips

Fingerprint type experiment cards (also called *gelstrips*) can be displayed in two modes, a raw mode, i.e. not normalized, and a normalized mode (see Figure 3-84). In the normalized mode, the band information is also shown. Band sizes are shown as molecular sizes (metrics) if the metrics regression curve is available for the reference system, or as relative distances from the top if no metrics regression curve is available.

3.8.2.1 To switch between the raw and normalized view, open the *Fingerprint type* window (Figure 3-26) and select *Layout > Show normalized gelcards*. If the feature is enabled, the menu item is flagged.

3.8.2.2 In case of a gelstrip, you can increase or decrease the size of the card using the keyboard, by pressing the **numerical + key** (increase) or the **numerical - key** (decrease).

3.8.2.3 You can right-click on the experiment card to pop up a floating menu, from which you can choose *Export normalized curve*, *Export normalized band positions*, and *Export normalized band metrics*. Selecting any of the above commands exports the corresponding information to the clipboard, from where it can be pasted as text, e.g. in Notepad.

3.8.2.4 In case multiple gelstrips are shown on the screen, it is possible to line them up by right-clicking on a gelstrip and choosing *Line up*. All gelstrips can be closed at once using *Close all* in the floating menu.

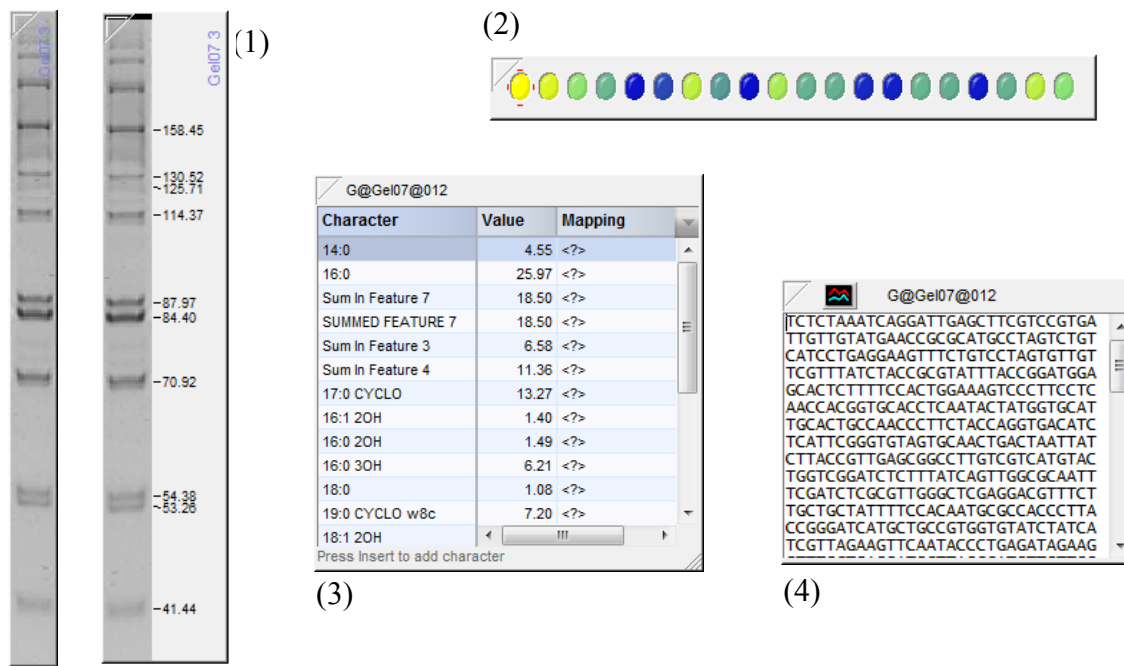


Figure 3-84. Experiment cards of fingerprint type (1) in raw mode (left) and normalized mode (right), character type with fixed number of characters (closed type) (2), character type with non-fixed number of characters (open type) (3), and sequence type (4).

3.8.2.5 In a connected database it is possible to show or edit the fingerprint lane information fields with *Fingerprint information fields* (see also 3.2.11 on fingerprint lane information fields and on connected databases).


3.8.3 Character experiment cards

3.8.3.1 In case of a character type, you can right-click on the experiment card to pop up a floating menu, from which you can call the character image import program BNIMA (*Edit image*) if applicable. You can copy the data set to the clipboard (*Copy data to clipboard*), or paste data from the clipboard into the experiment (*Paste data from clipboard*). *Export character values* creates a similar output, but provides the names of the characters in case of an *open character set* (see 3.3.1).


3.8.3.2 In a connected database (see Section 2.3), an additional option *Remove this experiment*, is available, making it possible to delete the character set from the database. This is an irreversible operation.

It is possible to enter or edit the character data directly on the experiment card. As an example, we will create a new character set for a database entry.

3.8.3.3 Open the *Entry edit* window. In this example, we choose the *Entry edit* window for a STANDARD, which has no character data available.

For experiments that are not available for the entry, an empty flask is shown: .

Depending on whether the character type is displayed as list or as plate (see 3.3.2), the input method is different.

3.8.3.4 Click the  button of an empty character type; in the example, we choose **PhenoTest**, which is a closed character type, displayed as plate.

A message displays “The experiment ‘PhenoTest’ is not defined for this entry. Do you wish to create a new one?”.

3.8.3.5 Answer **<Yes>** to this question. An empty experiment card appears.

3.8.3.6 In case of binary (plus or minus) data, you can enter the values using the **numerical + and - keys**. The cursor automatically jumps to the next test if you have entered a value.

3.8.3.7 You can move the cursor using the **Left and Right arrow keys**.


NOTE: If you use the + and - keys to enter non-binary data, the defined maximum for the character type is used if + is entered.

3.8.3.8 In case of non-binary values (real or integer values), each test can be varied continuously between the minimum and the maximum using the **PgUp key** (increase intensity) and the **PgDn key** (decrease intensity).

3.8.3.9 Press the close button of the experiment card. The program asks to save the changes made.

3.8.3.10 Open an *Entry edit* window of an entry which contains a **FAME** experiment (shown as a colored dot in the *Experiment presence* panel).

NOTE: You can also click on the colored dot to open the experiment card directly.

3.8.3.11 In the *Entry edit* window, press the  button of **FAME**. The entry cards for this experiment type are shown in list format (see 3.3.2). The card lists all fatty acids that are present in this entry, as percentages. Note that the window is resizable.

3.8.3.12 Click in the *Value* column next to a fatty acid to change/enter its value.

3.8.3.13 If you want to add a character to the list, press the Insert button. A dialog box shows all known characters for this character type which are not yet available in this entry, from which you can select one.

3.8.3.14 Press the **<Create new>** button to create a new character.


3.8.4 Sequence experiment cards

3.8.4.1 In case of a sequence type experiment card, you can click the right mouse button inside the card to call different edit functions (see 3.8.1).

3.8.4.2 Similar as for a character type experiment card, in a connected database (see Section 2.3), an additional option **Remove this experiment**, is available, making it possible to delete the character set from the database. This is an irreversible operation.

It is possible to enter the sequence data directly on the experiment card.


3.8.4.3 Open the *Entry edit* window. In this example, we will open the *Entry edit* window for a **STANDARD**, which has no sequence data available.


For experiments that are not available for the entry, an empty flask is shown: .

3.8.4.4 Press the  button of **16S rDNA**.

3.8.4.5 An empty sequence editor appears. You can enter bases or amino acids manually, or by pasting from the clipboard (SHIFT + INS).

NOTE: If you paste a nucleotide or an amino acid sequence in the experiment card, please use the correct IUPAC codes and amino acid abbreviations respectively. Avoid the use of improper letters, symbols and spaces.

3.8.4.6 When the sequence has been generated using the InfoQuest FP contig assembler program Assembler (in older versions GeneBuilder), pressing the  button will launch Assembler (GeneBuilder) with the contig project associated with this sequence.

3.8.4.7 When no contig project is available for this entry, pressing the  button will launch Assembler with a new project associated. See 3.4.3 for instructions how to work with Assembler.

3.8.4.8 By right-clicking in the experiment card, you can paste sequence data from the clipboard using **Paste from clipboard**. It is also possible to type bases directly from the keyboard. Right-clicking offers further editing tools **Undo** (ALT+Backspace), **Select all** (CTRL+A) and **Copy to clipboard** to copy the current selection.

4. COMPARISONS

4.1 General comparison functions CL

4.1.1 Definition


A *Comparison* in InfoQuest FP includes every function which allows to compare database entries. This involves the display of experiment images of selected entries, the calculation and display of cluster analyses, alignment of sequences, and calculation of principal component analysis (PCA) and multi-dimensional scaling (MDS) project.

Two different windows are available in InfoQuest FP for comparison of entries: The *Pairwise comparison* window offers a detailed comparison overview for all experiments available for two selected entries. Whenever more than two entries need to be compared, the *Comparison* window should be called.

The *Pairwise comparison* window and the *Comparison* window in InfoQuest FP present a comprehensive overview of all available experiments for a selection of entries and enables the user to show and compare any combination of images of experiments. A comparison is always created from a selection of database entries. These can be selected manually (see 2.2.7) or via the automatic search and selection functions (see 2.2.8 and 2.2.9).

4.1.2 The *Pairwise comparison* window

From within any window where you can select entries, you can display a detailed comparison between two entries. This pairwise comparison shows all the images of the experiment types as well as the similarities obtained using the specified coefficients.

4.1.2.1 In the Startup screen, double-click on **DemoBase** to open the database for analysis. Alternatively, you can select **DemoBase** from the list and then press the  button. In case **DemoBase** was already open, clear any previous selection with F4.

4.1.2.2 Select any two entries you want to compare.

4.1.2.3 In the *InfoQuest FP* main window, select *Comparison* > *Compare two entries* or use the **CTRL+2** (numerical 2) or **CTRL+ALT+C** shortcuts.

The **CTRL+2** or **CTRL+ALT+C** shortcuts work from within any window. The *Pairwise comparison* window appears (Figure 4-1).

The *Pairwise comparison* window consists of two dockable panels: the *Experiments* and the *Comparison* panel.

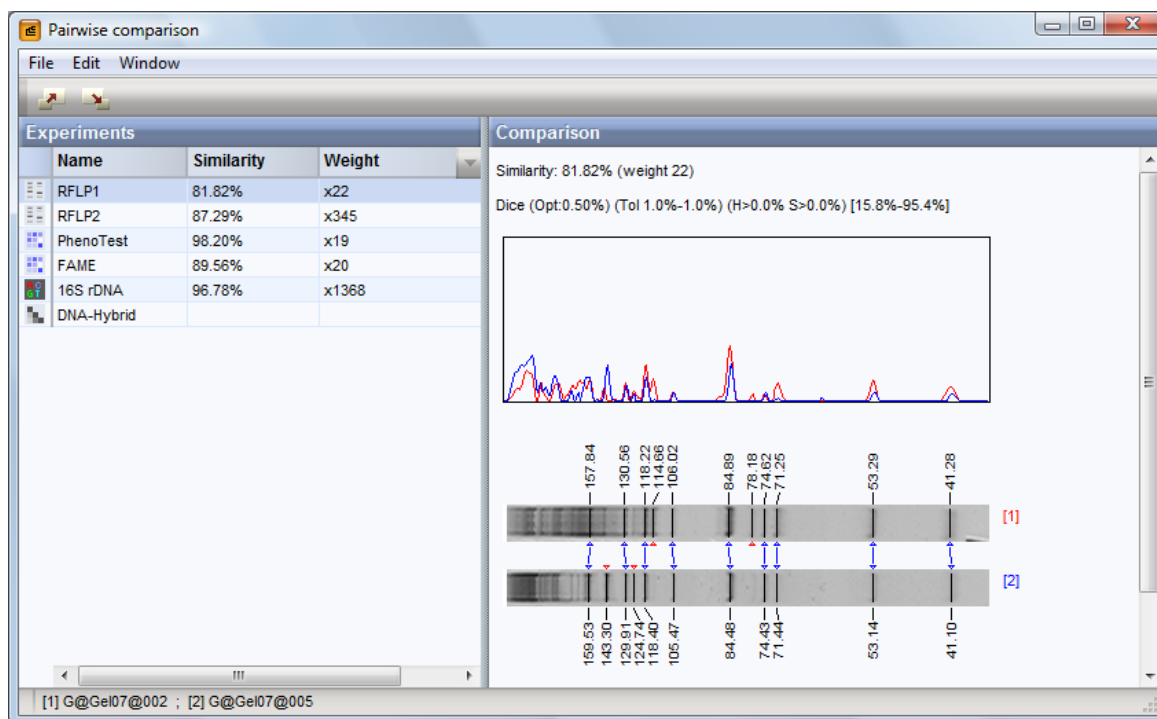




Figure 4-1. *Pairwise comparison* window.

For detailed information about the display of dockable panels, see 1.6.4.

The *Experiments* panel (left panel in default configuration) displays the names of all experiment types present in the database ('Name') and the type ('Type'). When an experiment type is present for both entries, the similarity value for this experiment type is shown in the information field 'Similarity'. 'Weight' displays the assigned weight to each experiment type. These information fields can be displayed or hidden by pressing the column properties button () and selecting the fields from the pull-down menu (see 1.6.6 for more information on grid panels).

4.1.2.4 Click on an experiment type in the *Experiments* panel to display the corresponding images in the *Comparison* panel (right panel in default configuration).

If you select a fingerprint type or a sequence type from the list, the *Comparison* panel lists the comparison settings used to calculate the similarity value.

NOTE: The comparison settings are defined in the Comparison settings dialog box. This dialog box can be accessed from each Experiment type window (via Settings > Comparison settings or ) and from the Comparison window (via Clustering > Calculate > Cluster analysis (similarity matrix)).

In case of fingerprint types, the detailed comparison of the band matching is shown if a band matching coefficient was chosen in the experiment settings (e.g. Dice coefficient in Figure 4-1).

In case of character types, all characters present in the character type are listed ('Character'), together with the character values and their corresponding colors. When a character mapping is defined, the categories are shown in the 'Mapping' column.

In case of sequences, the aligned sequences are shown.

In case of trend data types, the trend curves for both entries are shown.

4.1.3 The Comparison window

When more than two entries should be compared, this is achieved through the *Comparison* window. We will use the **DemoBase** database to explain this window.

4.1.3.1 In the Startup screen, double-click on **DemoBase** to open the database for analysis. Alternatively, you can select **DemoBase** from the list and then press the




button. In case **DemoBase** was already open, clear any previous selection with F4.

With all entries except the standards selected, we will create a new comparison. This selection can be done manually as described below or via the search and selection functions (see 2.2.8 and 2.2.9 for a description).

4.1.3.2 In the *Database entries* panel of the *InfoQuest FP main* window, click on the first database entry and, while holding the SHIFT key, click on the last entry to select all database entries. Alternatively, press CTRL+A on the keyboard to select all database entries at once.

4.1.3.3 Unselect the first entry marked as STANDARD by clicking it and selecting *Edit > Select/Unselect entry* or press the space bar on the keyboard. Repeat the same action for the second and third STANDARD.

4.1.3.4 Select *Comparison > Create new comparison* (ALT+C) or press the  button from the *Comparisons* panel toolbar. A *Comparison* window is created, with the selected database entries (Figure 4-2).

The *Comparison* window is divided in six main panels: the *Dendrogram* panel, which shows the dendrogram if calculated, the *Experiment data* panel, showing the images of the experiments, the *Information fields* panel, which shows the database fields in the same layout as in the database (see 2.2.5), the *Similarities* panel, which shows the similarity values, the *Experiments* panel, which shows the available experiment types and the *Groups* panel, which shows the groups if defined. Initially, the *Dendrogram* panel, the *Experiment data* panel, the *Similarities* panel and the *Groups* panel are empty.


4.1.3.5 You can drag the horizontal and vertical separator lines between the panels, in order to divide the space among the panels optimally.

4.1.3.6 All panels in the *Comparison* window are dockable and their position can therefore be changed according to your own preferences. For more information on the display of dockable panels, see 1.6.4.

NOTE: The Dendrogram, Experiment data, Information fields and Similarities panel behave as a group, i.e. these panels cannot be docked outside this group and they cannot be displayed in a window of their own (undocked).

The *Information fields* panel in the *Comparison* window is similar to the *Database entries* panel in the *InfoQuest FP main* window and contains the database information in tabular format (grid panel). For detailed information on the display options of grid panels, see 1.6.6.

4.1.3.7 In the *Information fields* panel, you can drag the separator lines between the information field columns to the left or to the right, in order to divide the space among the information fields optimally.

4.1.3.8 Clicking the column properties button () located on the right hand side in the information fields

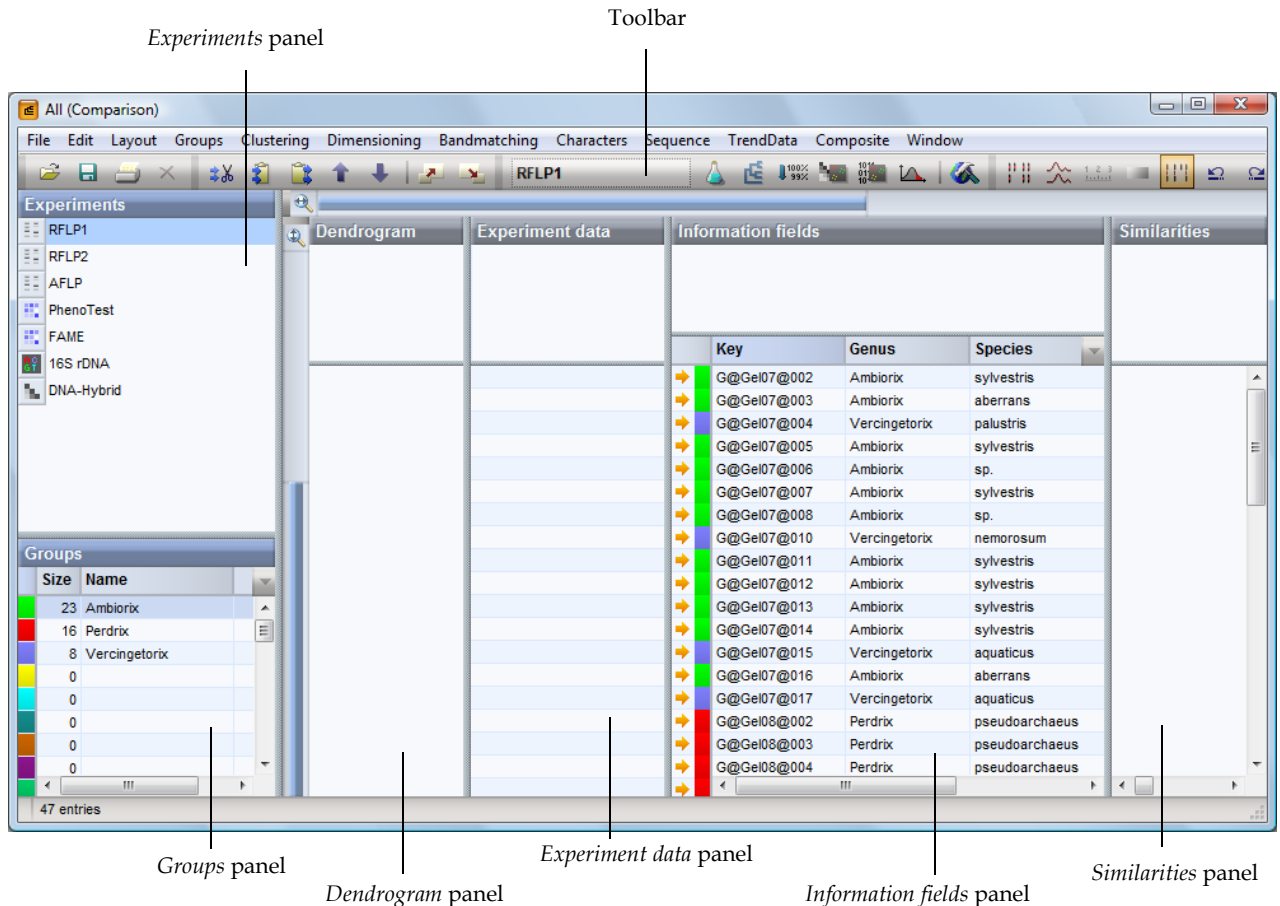










Figure 4-2. The Comparison window.



header in the *Information fields* panel gives access to functions allowing information fields to be displayed or hidden, frozen, or moved to the left or to the right (see 1.6.6 for details).

4.1.3.9 As explained in 2.2.5.6, it is possible to freeze one or more information fields in the *InfoQuest FP main window* using *Edit > Freeze left panel*, so that they always remain visible left from the scrollable area. The same fields will be frozen in the *Comparison window*. This feature can be combined with the possibility to change the order of information fields, which makes it possible to freeze any subset of fields.


From the *Experiments* panel, you can select one of the available experiment types, to show an image, calculate a dendrogram, or show a matrix. Each experiment type in the *Experiments* panel contains two objects: a button and the experiment type name, on the right hand side of the button. In case of a fingerprint type, the button is shown as ; character experiments as ; sequence types as ; matrix types as ; composite data sets as  and trend data types as . Clicking one of these buttons shows the image of the corresponding experiment type in the *Experiment data* panel.




NOTE: The experiments in the Experiments panel of the Comparison window are listed in the same order as they are listed in the Experiments panel of the InfoQuest FP main window. This feature also allows one to control the order in which experiment data are displayed in a composite data set (see 4.8.2).

4.1.3.10 Press the  button of **RFLP1**; the pattern images are shown for **RFLP1**. When the image of an experiment type is displayed, the button shows like .

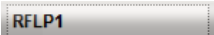
4.1.3.11 Press the  button of **RFLP2**; the pattern images of **RFLP2** are shown right from those of **RFLP1**. The button of **RFLP2** now shows like .

NOTE: To display more than one image at a time, we recommend to maximize the Comparison window, and to use maximal space for the Experiment data panel by minimizing the Dendrogram and Similarities panels (see 4.1.3.5).

4.1.3.12 If insufficient space is available to show both images at the same time, you can scroll through the *Experiment data* panel, or use the zoom functions .

and  (*Layout > Zoom in* and *Layout > Zoom out*). Shortcut keys for these actions are CTRL+PgUp and CTRL+PgDn, respectively. The zoom sliders indicated with  and  can be used to zoom selectively in the horizontal or vertical direction, respectively. See 1.6.7 for a detailed description of zoom slider functions.

4.1.3.13 In the caption of the *Experiment data* panel, you can drag the separator line between the images to the left or to the right, in order to reserve more or less horizontal space for a particular experiment image. The original aspect ratio (proportion height to width) of the image will not be maintained by this action.

4.1.3.14 To select an experiment type in the *Comparison* window, you can either click on the experiment type name in the *Experiments* panel, on the image itself or select the experiment type from the drop-down menu when pressing the *Active experiment* button  in the toolbar.

When an experiment type is selected, both the image caption in the *Experiment data* panel (if the image is shown) and its name in the *Experiments* panel are highlighted. All functions listed under *Clustering, Dimensioning, Bandmatching, Characters, Sequence, TrendData, and Composite* as well as some *Layout* functions, apply to the selected experiment type.


4.1.4 Adding and removing entries


Selections of entries made in the *Database entries* panel of the *InfoQuest FP main* window are also shown in the *Information fields* panel of the *Comparison* window and vice versa. The entries in a newly created comparison are all marked with a colored selection arrow, since they were all selected in the database. You can manually select and unselect entries in the *Information fields* panel (see Figure 4-2), using the CTRL and SHIFT keys as described in . Selections can be added or removed from an existing comparison.

4.1.4.1 First unselect all entries by pressing the F4 key.


4.1.4.2 Select some entries from the comparison (see 2.2.7).



4.1.4.3 With *Edit > Delete selection* (shortcut DEL on the keyboard), the selected entries are removed from the comparison. The program will ask for confirmation to remove the selection from the comparison. You will not be able to undo this operation.

4.1.4.4 With *Edit > Cut selection* or  (shortcut CTRL+X on the keyboard), the selected entries are removed from the comparison and are copied to the clipboard.

4.1.4.5 With *Edit > Paste selection* or  (shortcut CTRL+V on the keyboard), the same entries are placed back into the comparison. If no dendrogram is present, they are placed at the position of the selection bar. This tool can be used to rearrange entries in the *Comparison* window (see also 4.1.5).

Entries can be added to an existing comparison at any time. The entries first need to be copied to the clipboard from the *InfoQuest FP main* window or from another comparison.

4.1.4.6 To copy entries to the clipboard, select the entries (e.g. in the *InfoQuest FP main* window) first and use the *Edit > Copy selection* command or  (shortcut CTRL+C on the keyboard).

To cut entries from one comparison into another, use *Edit > Cut selection* or  (shortcut CTRL+X on the keyboard) in the one comparison and *Edit > Paste selection* or  (shortcut CTRL+V on the keyboard) in the other comparison.

New database entries can be added to an existing *dendrogram* (see 4.1.9 on how to calculate a dendrogram) in this way: select the new entries in the database, open an existing comparison with dendrogram, and paste the selection into the comparison. Both the similarity matrix and the dendrogram will be updated, which uses considerably less time than recalculating the whole cluster analysis.

NOTE: Entries can also be selected from the Dendrogram panel: hold the CTRL key and left-click on a branch node to select/unselect a cluster on the dendrogram at once (see also 4.1.11).

For adding and removing entries from a global sequence alignment, see 4.5.12.

4.1.5 Rearranging entries in a comparison

The cut and paste functions can be used to rearrange entries in the *Comparison* window (4.1.4.4 to 4.1.4.5). Some other convenient functions are available for rearranging entries in a comparison, as explained below.

4.1.5.1 Select *Edit > Arrange entries by database field* to sort the entries according to the highlighted database field.

When two or more entries have identical strings in a field used to rearrange the order, the existing order of the entries is preserved. As such it is possible to categorize entries according to fields that contain information of different hierarchical rank, for example *genus* and *species*. In this case, first arrange the entries based upon

the field with the lowest hierarchical rank, i.e. *species*, and then upon the higher rank, i.e. *genus*.

4.1.5.2 When a field contains numerical values, which you want to sort according to increasing number, use **Edit > Arrange entries by database field (numerical)**.

In case numbers are combined numerically and alphabetically, for example entry numbers [213, 126c, 126a, 126c], you can first arrange the entries alphabetically (**Edit > Arrange entries by database field**), and then numerically using **Edit > Arrange entries by database field (numerical)**. The result will be [126a, 126b, 126c, 213].

4.1.5.3 A group of selected entries (colored arrows) can be placed at the position of the cursor (the entry you last clicked on) with **Edit > Bring selected entries to cursor**.

4.1.5.4 A group of selected entries (colored arrows) can be moved to the top of the comparison with **Edit > Bring selected entries to top** (shortcut CTRL+T on the keyboard).


4.1.5.5 An individual entry can be moved up and down by left-clicking on it, and selecting **Edit > Move entry up**



or **Edit > Move entry down**



. Pressing the Arrow Up and Arrow Down keys on the keyboard while holding down the SHIFT key does the same.


4.1.5.6 When using the up/down buttons  and



, you can move an entry to the top or the bottom at once by holding the CTRL key.

4.1.6 Saving and loading comparisons

A comparison can be saved and all calculations done on the data it contains, will be stored along. This includes similarity matrices in all experiment types where they have been calculated, any dendrogram that has been calculated (see 4.1.9), band matchings and polymorphism analyses (see Section 4.3). In a connected database (see Section 2.3) it is even possible to share comparisons with other users who share the same database.

4.1.6.1 Select **File > Save as** or press  to save the comparison (shortcut CTRL+S on the keyboard). In a connected database, you can save the comparison either locally or in the connected database (see Figure 4-3).

4.1.6.2 Enter a name, e.g. **MyComp**.

4.1.6.3 Close the comparison with **File > Exit**. Comparison **MyComp** is now listed in the *Comparisons* panel of the *InfoQuest FP* main window.

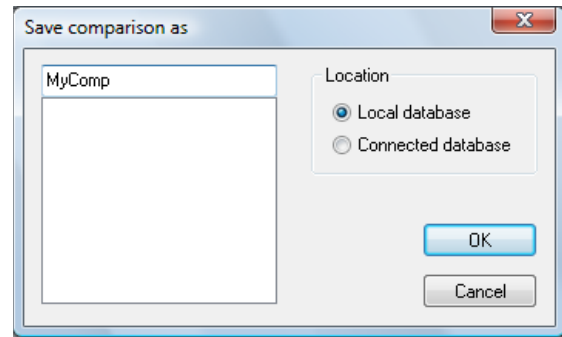



Figure 4-3. Save comparison dialog box in a connected database.

The default information field 'Location' in the *Comparisons* panel of the *InfoQuest FP* main window lists the location (i.e., Shared or Local) of the saved comparisons.

If a comparison is saved in the connected database (see Figure 4-3), the comparison will be visible by other users who are connected to the same database. Only the first user, however, will be able to save changes in the comparison; for subsequent users it will be read-only.

4.1.6.4 To open an existing comparison, select it from the list in the *Comparisons* panel of the *InfoQuest FP* main window and press the  button. Alternatively, just double-click on the comparison name.

4.1.7 Interaction between subsets and comparisons

Comparisons can be created from subsets and *vice versa* (for more information about subsets, see 2.2.10). To create a subset from a comparison there is a direct function available in the *Comparison* window.

4.1.7.1 First, make sure a subset is open. If you want to create a subset that contains only the members of the comparison, create a new subset (see 2.2.10 on how to create subsets).


4.1.7.2 In the *Comparison* window, select **File > Add entries to current subset**.

The current subset now contains the entries of the comparison, in addition to entries that may have been present in the subset before.

Conversely, a comparison can be created from a subset as follows:

4.1.7.3 Open a subset in the database (see 2.2.10).

4.1.7.4 Click on the first entry and subsequently, while holding the SHIFT key, on the last entry in the *Database entries* panel to select all entries from the subset. Alternatively, press CTRL+A on the keyboard.

4.1.7.5 Select *Comparison* > *Create new comparison* or press the  button from the *Comparisons* panel toolbar (shortcut ALT+C).

A new comparison is created, containing all entries from the subset.

4.1.8 Cluster analysis: introduction

In many cases, the user would want to perform a *cluster analysis* based on a certain experiment. The term *cluster analysis* is a collective noun for a variety of techniques that have the common feature to produce a hierarchical tree-like structure (*dendrogram*) from the set of sample data provided. The tree usually allows the samples to be classified based upon the *clusters* produced by the method. Apart from this common goal, the principles and algorithms used, as well as the purposes, may be very different (for more information about cluster analysis, see 4.10.1). Cluster analysis *sensu latu* has therefore been subdivided in different sections in this manual:


- Cluster analysis *sensu stricto* is based upon a matrix of similarities between database entries and a subsequent algorithm for calculating bifurcating hierarchical dendrograms representing the clusters of entries (Section 4.1 to Section 4.9).
- Phylogenetic cluster methods are methods which attempt to create trees that optimize a specific phylogenetic criterion. These methods start from the data set directly rather than from a similarity matrix (Section 4.9).
- Minimum spanning trees are trees calculated from a distance matrix, that possess the property of having a total branch length that is as small as possible (Section 4.11).


4.1.9 Calculating a dendrogram


4.1.9.1 Open the database **DemoBase**.

4.1.9.2 Select all entries except STANDARD and create a new comparison (see 4.1.3.2 to 4.1.3.4).

We will save this comparison (see 4.1.6) since it will be used throughout this manual.

4.1.9.3 Select *Edit* > *Save* or press the  button. The dialog box that appears prompts for a name for the comparison. Enter **All** for example.

4.1.9.4 Select an experiment type in the *Experiments* panel (e.g. **RFLP1**) and show the image by pressing the image button ( for **RFLP1**).

4.1.9.5 Select *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)*. You can also press the  button, in which case the following menu pops up (Figure 4-4).

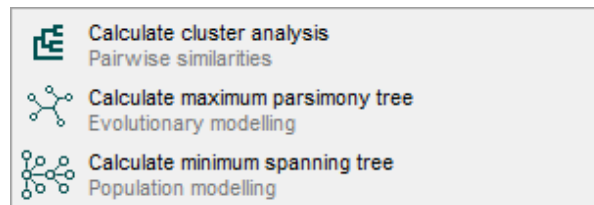


Figure 4-4. Cluster analysis menu popped up from the dendrogram button.

The first choice is the matrix-based cluster analysis discussed in this chapter, whereas the second and third choices are discussed in Section 4.9 and Section 4.11, respectively. Note that, in case of aligned sequence data, an extra option *Calculate maximum likelihood tree* becomes available, which is also discussed in Section 4.11.


A *Comparison settings* dialog box pops up. For each experiment type different settings are listed in this dialog box. More information about these settings can be found in the cluster analysis sections of each experiment type.

4.1.9.6 For this example, you can leave the default settings. Press <OK> in the *Comparison settings* dialog box to start the cluster analysis.

During the calculations, the program shows the progress in the *Comparison* window's caption (as a percentage), and there is a green progress bar in the bottom of the window that proceeds from left to right.

When finished, the dendrogram and the similarity matrix are shown (Figure 4-5). For more information about the different panels in the *Comparison* window, see 4.1.3.

The experiment type from which the dendrogram is generated, is shown in the header of the dendrogram panel. The parameters and settings of the cluster analysis are shown in the header of the matrix panel.

4.1.9.7 To save the comparison with the dendrogram, select *File* > *Save* or press the  button. Comparison **All** now contains a dendrogram for fingerprint type **RFLP1**.

4.1.10 Calculation priority settings

InfoQuest FP performs almost all its calculations in multi threaded mode. This means that you can further use InfoQuest FP or any other program while time-

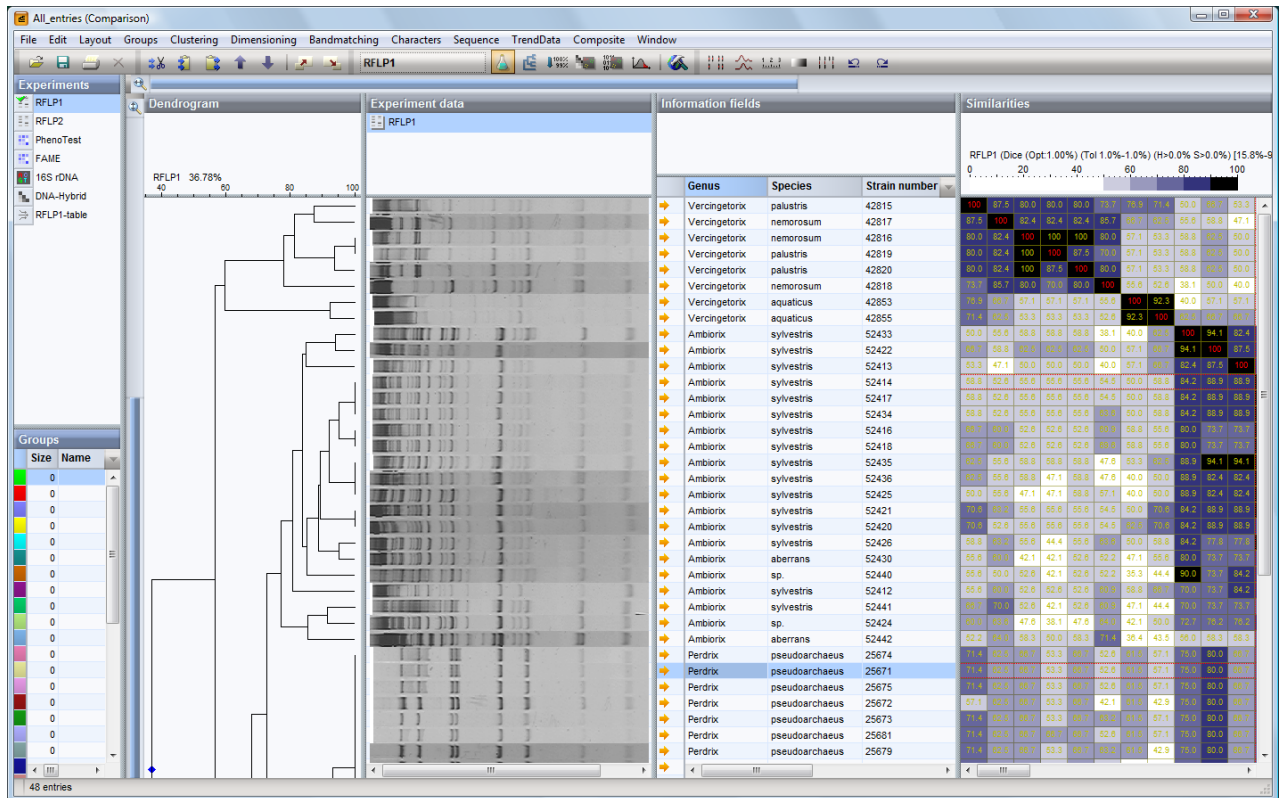


Figure 4-5. Comparison window with dendrogram, image, entry names and similarity matrix.

consuming calculations are going on (especially sequence alignments and phylogenetic clustering can take a long processing time). In order to speed up the calculations, or make multi tasking smoother, you may want to modify the priority settings for the calculations. The calculation priority settings are grouped with other preference settings in the *InfoQuest FP main window*.

4.1.10.1 In the *InfoQuest FP main window*, select **File > Preferences** and click on **Calculation priority settings** in the list on the left side of the *Preferences dialog box* (see Figure 4-6).

The dialog box offers the choice between five priority levels. If **Foreground** is chosen, it will not be possible to run other applications while the calculations are going on. **Idle time background** means that the computer will only process the InfoQuest FP calculations while it has nothing else to do.

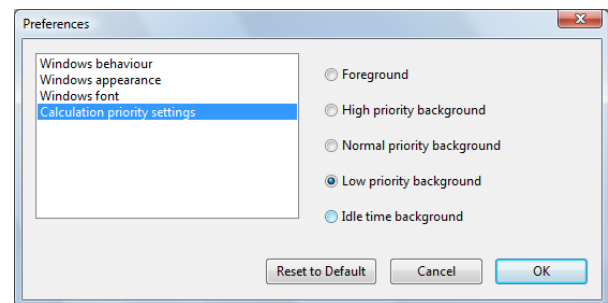



Figure 4-6. Calculation priority settings in the *Preferences dialog box*.

4.1.10.2 Select **Low priority background** and press **<OK>**.

While the program is calculating, you can abort the calculations at any time using the  button.

4.1.11 Dendrogram display functions

4.1.11.1 In the **DemoBase** database, open comparison **All** or any other comparison for which a dendrogram is calculated (see 4.1.9 on how to calculate a dendrogram).

4.1.11.2 Press **F4** to unselect any previous selection of database entries.

Entries can be selected from within the *Dendrogram* panel of the *Comparison* window:

4.1.11.3 To select an individual entry, hold the CTRL key and click on a dendrogram tip (where a branch ends in an individual entry). Alternatively, right-click on the dendrogram tip and choose *Select branch into list* from the floating menu. Repeat this action to unselect the entry.

4.1.11.4 To select a cluster on the dendrogram at once, hold the CTRL key and left-click on a branch node. Alternatively, right-click on a branch and choose *Select branch into list* from the floating menu. Repeat this action to unselect a branch.

When a dendrogram node or tip is clicked on, a diamond-shaped cursor appears on that position. The average similarity at the cursor's place is shown in the upper left corner of the *Dendrogram* panel. You can also move the cursor with the arrow keys.

In some cases, it may be necessary to select the root of a dendrogram, for example if you want to (un)select all the entries of the dendrogram. In case of large dendrograms, selecting the root may be difficult using the mouse.

4.1.11.5 With *Clustering > Select root*, the cursor is placed on the root of the dendrogram.

Two branches grouped at the same node can be swapped to improve the layout of a dendrogram or make its description easier:

4.1.11.6 Select the node where two branches originate and *Clustering > Swap branches*.

To simplify the representation of large and complex dendrograms, it is possible to simplify branches by abridging them as a triangle.

4.1.11.7 Select a cluster of closely related entries and *Clustering > Collapse/expand branch*.

4.1.11.8 With *Clustering > Show similarity values*, the average similarity of every branch is indicated on the dendrogram.

Another function, *Clustering > Reroot tree*, only applies to so-called *unrooted trees*, i.e. neighbor joining, parsimony and maximum likelihood trees. These clustering methods produce trees without any specification as to the position of the root or origin. Since users will want to display such trees in the familiar dendrogram representation, the tree is to be rooted artificially. "Rerooting" is usually done by adding one or more unrelated entries (so-called *outgroup*) to the clustering, and using the outgroup as root. The result is a *pseudo-rooted* tree.

To illustrate the rerooting of an unrooted tree, we will create a second dendrogram, based upon neighbor joining of the fingerprint type **RFLP2**.

4.1.11.9 In the *Experiments* panel, select **RFLP2**.

4.1.11.10 Select *Clustering > Calculate > Cluster analysis (similarity matrix)* and specify *Neighbor Joining* in the dialog box. A neighbor joining tree is calculated for **RFLP2**.

If you scroll through the tree, you will notice that two entries, i.e. *Perdrix* sp. strain numbers 53175 and 25693 protrude on a very long branch. These two entries are ideally suited as "outgroup".

4.1.11.11 Click somewhere in the middle of the branch. A secondary, X-shaped cursor appears.

4.1.11.12 Select *Clustering > Reroot tree*, and the new root connects the outgroup with the rest of the entries.

4.1.11.13 The software automatically limits the displayed similarity range to the depth of the dendrogram. If you want to change this range, select *Clustering > Set minimum similarity value*.

4.1.11.14 The similarity scale can be displayed in similarity (default for most clustering types) or in distance. To toggle between similarity and distance modes, select *Layout > Show distances*.

If a dendrogram is calculated for more than one experiment type in a comparison, you can toggle between the available dendrograms as follows:

4.1.11.15 In the *Experiments* panel, click on the experiment type for which you want to display the dendrogram (e.g. **RFLP1**) and select *Layout > Show dendrogram*. Alternatively, right-click on the experiment type and select *Show dendrogram* from the floating menu that appears.

4.1.12 Working with Groups

An important display function in the *Comparison* window is the creation of Groups. Groups basically are subsets of a comparison, that can be defined from clusters, from database fields, or just from any subdivision the user desires. Groups are normally displayed using rectangles of different colors next to the entries, each group having its own color. They can also be displayed using different symbols, or using alphanumeric codes. In the first place, Groups facilitate the comparison between a dendrogram or a dimensioning and a certain characteristic (database information). Groups also make the homology display between dendrograms obtained from different experiments easier. In addition, Groups are necessary in a number of derived statistical analysis functions, such as Partitioning, Group separation, Discriminant Analysis, and MANOVA (4.1.15). Finally, Groups form an easy link between dimensional representations such as PCA, SOM or graphs and scatter plots on the one hand, and database field information on the other hand. To make the distinction between groups as clusters on the one hand and groups as defined by the

Groups tool on the other hand, the latter Groups are always written with a capital.

First we will see how to define Groups based on the clusters in a dendrogram.

4.1.12.1 If a UPGMA dendrogram of **RFLP1** is already calculated for comparison **All**, show it as follows: right-click on **RFLP1** in the *Experiments* panel, and select **Show dendrogram**. Otherwise, calculate a dendrogram as described in 4.1.9.4 to 4.1.9.6.

This dendrogram reveals three major clusters: *Vercingetorix*, *Ambiorix*, and *Perdrix* (with some exceptions).

4.1.12.2 Make sure that no entries are selected by pressing F4.

4.1.12.3 Hold the CTRL key and click on the node that connects all entries belonging to the *Vercingetorix* cluster. The entries of this cluster are now selected (as indicated by the colored arrows in the *Information fields* panel).

4.1.12.4 In the menu, select **Groups > Assign selection to**. The menu lists 30 different colors and accompanying symbols, from which you can choose one (e.g. the first one, green).

The selected color is shown next to all selected entries in the *Information fields* panel.

4.1.12.5 The *Groups* panel displays the number of entries next to the group color (see Figure 4-7).

4.1.12.6 Click twice in the information field 'Name', to add descriptive information to the defined group (see Figure 4-7).

Size	Name
8	
0	
0	
0	
0	
0	
0	
0	
0	
0	

47 entries

Figure 4-7. The *Groups* panel.

4.1.12.7 Press F4 to clear the selection and click on the node connecting all *Ambiorix* entries while holding the CTRL key.

4.1.12.8 Select **Group > Assign selection to** and choose the second color (red).

4.1.12.9 Repeat actions 4.1.12.2, 4.1.12.3, and 4.1.12.4 for the third cluster mainly composed of *Perdrix*. Use for example the third color (purple).

4.1.12.10 You can repeat these actions for two outliers of *Perdrix*, using another color. The *Groups* panel is updated whenever a new group is created.

Whatever dendrogram you now display, you will be able to recover the groups of the **RFLP1** dendrogram at a glance.

4.1.12.11 Right-click on **RFLP2** in the *Experiments* panel, and select **Show dendrogram** or calculate a dendrogram based on **RFLP2** if not yet present. The *Perdrix* and *Ambiorix* strains are not well separated by this technique: the second and the third Group are mixed up.

The Group assignments are saved along with the cluster analysis.

An alternative method to define Groups is by selecting a database field and having the program automatically create Groups based upon the different names that exist in this database field. One should be aware, however, that any misspelled name or typographic error will result in a different group. The method works as follows:

4.1.12.12 Select a database field by clicking on the database field name, for example 'Genus'.

4.1.12.13 In the *Groups* menu, select **Create groups from database field** or right-click in the database field name 'Genus' and select **Create groups from database field**. The *Create groups from field* dialog box appears (see Figure 4-8).

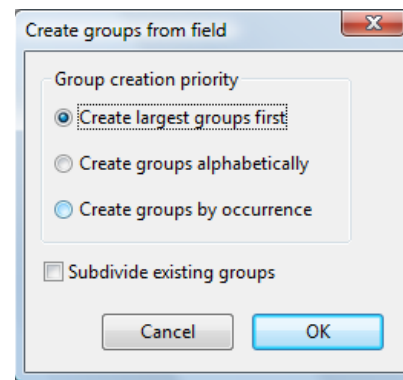


Figure 4-8. The *Create groups from field* dialog box.

The *Group creation priority* setting determines the order in which Groups are assigned. With **Create largest group first**, the group containing the largest number of entries will be Group 1, the second largest group will be Group 2, etc. With **Create groups alphabetically**, Groups will be created according to the alphabetical order of the information field. **Create groups by occurrence** assigns Groups in the order in which they are listed in the

comparison; the first occurrence of a Group member determines its Group number. When *Subdivide existing groups* is disabled (not checked), any previously defined Groups will first be removed and the program will assign the new Groups based upon the selected database field only. If you check *Subdivide existing groups*, the program will keep the Groups that are already defined, and split existing Groups into more Groups if differences in the selected database field are found.

4.1.12.14 Leave *Subdivide existing groups* disabled and press <OK>. The program creates three Groups according to the genus names.

NOTE: The maximum number of Groups that can be defined is 30. In case a database field contains more than 30 different names (text strings), the program will only assign Groups to the first 30. Depending on the selected option in the Create groups from field dialog box (see Figure 4-8), this is based on prevalence, alphabet or order of occurrence of the names.

The Groups panel displays the number of entries for each group and the genus name as Group name (Figure 4-9).

Size	Name
23	Ambiorix
16	Perdrix
8	Vercingetorix
0	
0	
0	
0	
0	
0	
0	

47 entries

Figure 4-9. The Groups panel with three defined groups.

4.1.12.15 Select the 'Species' database field and *Group > Create from database field* again.

4.1.12.16 Select *Subdivide existing groups* and press <OK>.

Every unique species name now is assigned to a different Group. In addition, if two different genus names would have the same species name, they would belong to a different Group too, since we kept the existing Groups based upon the genus database field (see Figure 4-10).

Since different colors are not equally distinguishable by different persons it may be useful to customize the Group colors in a user-defined scheme.

4.1.12.17 To define an own Group color scheme, select *Groups > Edit group colors*. This brings up the *Group colors* dialog box (Figure 4-11). For each color, three slide bars (red, green and blue, respectively) can be adjusted to produce any desired color.

Size	Name
16	Ambiorix,sylvestris
14	Perdrix,pseudoarchaeus
4	Ambiorix,aberrans
3	Ambiorix,sp.
3	Vercingetorix,nemorosum
3	Vercingetorix,palustris
2	Vercingetorix,aquaticus
2	Perdrix,sp.
0	

47 entries

Figure 4-10. The Groups panel after executing the *Subdivide existing groups* command.

4.1.12.18 A thus obtained color scheme can be saved by pressing the <Save as> button, and entering a name.

A user defined color scheme can be selected from the

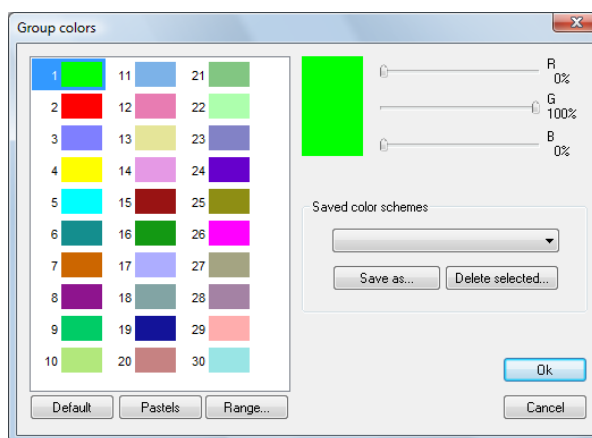


Figure 4-11. The *Group colors* dialog box.

drop-down list of *Saved color schemes*.

4.1.12.19 To delete a saved color scheme, first select it, and then press the <Delete selected> button.


4.1.12.20 To bring up the default color scheme, press <Default>. Another predefined scheme, using pastel colors, can be loaded by pressing <Pastels>.

4.1.12.21 It is also possible to generate a scheme of transition colors by pressing <Range>. The program will ask to enter the number of colors to include in the range. Enter a number between 2 and 30.

Besides using group colors, it is possible to assign a symbol to each group.


4.1.12.22 Uncheck *Groups > Show groups using colors* in the menu of the *Comparison* window.

The colors in the *Information fields* panel are replaced by symbols. To view the list of symbols in the *Groups* panel,

select the  button in the *Groups* panel and select *Sign* from the pull-down menu.

In many cases, the entry keys or particular information fields may be too long to be displayed in particular comparison types, e.g. phylogenetic trees, PCA plots, MANOVAs, rendered trees. In such cases, the entry keys can be replaced by a Group code. The program assigns a letter to each defined Group, and within a Group, each entry receives a number. The Group codes can be shown as follows:

4.1.12.23 In the *Comparison* window, select *Layout > Use group numbers as key*.

The keys in the *Information fields* panel are replaced by a letter followed by a number. The letter corresponds to the Group letter. To view the list of letters in the *Groups* panel, select the  button in the *Groups* panel and select *Letter* from the pull-down menu

A legend to the Group numbers can be obtained with *File > Export database fields* in the *Comparison* window.

Alternatively, a selected information field can be displayed instead of the key:

4.1.12.24 In the *Comparison* window, click on the information field which you would like to display as key (e.g. 'Strain number').

4.1.12.25 Select *Layout > Use field as key* from the menu in the *Comparison* window. The strain number is now displayed in the 'Key' field and in e.g. a PCA plot, rendered tree, etc.

4.1.13 Cluster significance tools

A dendrogram tells you something about the groups among a selection of entries, but nothing about the *significance*, i.e. the reliability or the trueness of these groups. Therefore, the software offers a range of methods that express the stability or the error at each branching level. The simplest indication of the significance of branches is showing the average similarities of the dendrogram branches (see 4.1.11.8).

The *Standard Deviation* of a branch is obtained by reconstructing the similarity values from the dendrogram branch and comparing the values with the original similarity values. The standard deviation of the derived values versus the original values is a measure of the reliability and internal consistence of the branch.

4.1.13.1 Right-click on **RFLP1** in the *Experiments* panel, and select *Show dendrogram*.

4.1.13.2 Select *Clustering > Calculate error flags*.

An error flag is drawn on each branch. The average similarity and the exact standard deviation is shown at the position of the cursor (see Figure 4-12). The smaller this error flag, the more consistent a group is. For example, the *Perdrix* group has a small error flag, meaning that this group is very consistent. This group will for example not disappear by incidental changes such as tolerance settings, adding or deleting entries, etc.

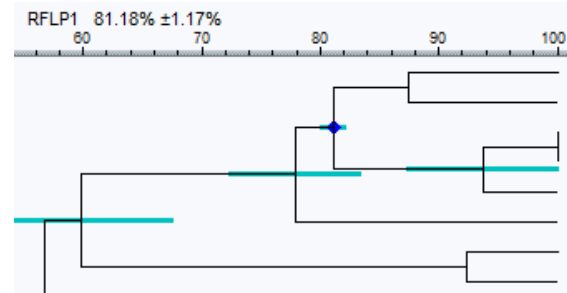


Figure 4-12. Dendrogram with error flags, detail. The average similarity and standard deviation is shown at the cursor's position (top).

4.1.13.3 Select *Clustering > Calculate error flags* again to remove the error flags.

The *Cophenetic Correlation* is also a parameter to express the consistency of a cluster. This method calculates the correlation between the dendrogram-derived similarities and the matrix similarities. The value is usually calculated for a whole dendrogram, to have an estimation of the faithfulness of a cluster analysis. In InfoQuest FP, the value is calculated for each cluster (branch) thus estimating the faithfulness of each subcluster of the dendrogram. Obviously, you can obtain the cophenetic correlation for the whole dendrogram by looking at the cophenetic correlation at the root.

4.1.13.4 Select *Clustering > Calculate cophenetic correlations*.

The cophenetic correlation is shown at each branch (Figure 4-13), together with a colored dot, of which the color ranges between green-yellow-orange-red according to decreasing cophenetic correlation. This makes it is easy to detect reliable and unreliable clusters at a glance.

*Bootstrap analysis*¹ measures cluster significance at a different level. Instead of comparing the dendrogram to its similarity matrix, it directly measures the influence of characters on the obtained dendrogram. The concept is very simple: "sampling with replacement", i.e. characters are randomly left out from the character set and are replaced with others². For each sampling case, the

1. Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Statist.* 7:1-26

2. Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791

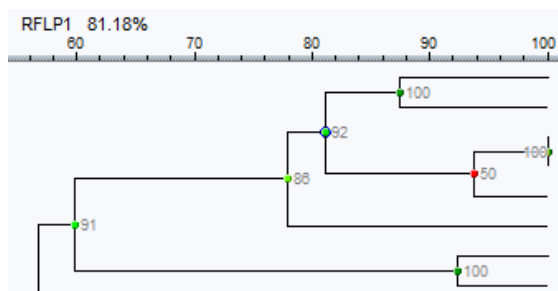


Figure 4-13. Dendrogram showing cophenetic correlation values, detail.

dendrogram is recalculated, and the relative number of dendrograms in which a given cluster occurs is a measure of its significance. This method requires the characters to be independent and equally important.

Since bootstrap analysis requires a closed character set, the method can only be performed on aligned sequences and character type data. In case of fingerprint type data, a band matching needs to be performed first (Section 4.9).

NOTES:

(1) Bootstrap analysis of character type data can only be performed on a composite data set (see Section 3.7).

(2) To be able to perform bootstrap analysis on your data, do NOT choose **Average from experiments** in the Composite data set comparison settings dialog box (see Section 4.8). The bootstrap analysis option is only accessible when using a similarity coefficient.

4.1.13.5 Select the composite data set **All-Pheno** (if not already present, see Section 3.7 on how to set up a composite data set) in the *Experiments* panel and select **Clustering > Calculate > Cluster analysis (similarity matrix)**.

4.1.13.6 In the *Composite data set comparison* dialog box, select **Pearson correlation** with **Standardized characters** and **UPGMA** as clustering method. Press **<OK>** to calculate the dendrogram.

4.1.13.7 When the dendrogram appears, select **Clustering > Bootstrap analysis** and enter the number of simulations (samplings) to perform. A reasonable number of samplings is 100.

4.1.13.8 Press **<OK>** and wait until the sampling and calculation process is finished. No need to explain that calculating 100 matrices and dendrograms can take some computing time.

The bootstrap values are shown in a similar way as the cophenetic correlation values (see Figure 4-13).

Another way of looking at dendrograms is to try to delimit, by objective means, the relevant clusters from the non-relevant clusters. The simplest and most arbi-

trary method is to draw a vertical line through the dendrogram in a way that it cuts most homogeneous clusters from most heterogeneous clusters. However, there are more statistically founded methods to draw either straight lines, or to evaluate cluster by cluster and delimit relevant clusters at different similarity levels. The *Cluster Cutoff method* in InfoQuest FP is one of these statistical methods. The method draws a line through the dendrogram at a certain similarity level, and from the resulting number of clusters defined by that line, it creates a new, simplified, similarity matrix, in which all within-cluster values are 100%, and all between-cluster values are 0%. Then, the *Point-biserial correlation* is calculated, i.e. the correlation between this new matrix and the original similarity matrix. The same is done again for other cutoff similarity levels, and the level offering the highest *Point-biserial correlation* is the one offering the most relevant groups.

In InfoQuest FP, this standard method is even further refined, as the cutoff values can be different per cluster, allowing even more reliable clusters to be defined.

4.1.13.9 Select **Clustering > Calculate cluster cutoff values**. The branches that were found to be below the cluster cutoff value are shown in dashed lines.

4.1.14 Matrix display functions

The similarity matrix is displayed in the *Similarities* panel, located at the right hand side of the *Comparison* window (see Figure 4-5).

4.1.14.1 If the similarity matrix is not shown for the selected experiment, you can display it with **Layout > Show matrix**. This option is only available when a dendrogram was calculated for the selected experiment. A dendrogram can be calculated as described in 4.1.9.4 to 4.1.9.6.

NOTE: It is also possible to show the average similarities for the branches directly on the dendrogram; see 4.1.11.8.

4.1.14.2 It may be necessary to reduce the space allocated for the image and for the information fields, in order to increase the space for the matrix panel, by dragging the separator lines between the panels.

Initially, the matrix is displayed as differentially shaded blocks representing the similarity values. The interval settings for the shadings is graphically represented in the caption of the *Similarities* panel (Figure 4-14).



Figure 4-14. Adjustable similarity shading scale.

There are two ways to change the intervals for shading:

4.1.14.3 Drag the interval bars on the scale; the matrix is updated instantly.

4.1.14.4 Select *Layout > Similarity shades* in the menu. The maximum/minimum values for each interval can be entered as numbers.

4.1.14.5 To show the similarity values in the matrix, select *Layout > Show similarity values*. If it is difficult to read the values on the shaded background, you can remove the shades with *Layout > Similarity shades* and entering 100% for each interval.

4.1.14.6 With the option *Layout > Show matrix rulers* (default enabled), a set of horizontal and vertical rulers appear on the similarity cell where clicked. These rulers connect the two entries from which the similarity value is derived.

If you want to find the similarity value on the matrix between two entries in the comparison, click first on the point on the diagonal of the matrix corresponding to the first entry, and then on the second entry inside the *Information fields* panel (Figure 4-15). The similarity value is the intersection between the horizontal and the vertical rulers.

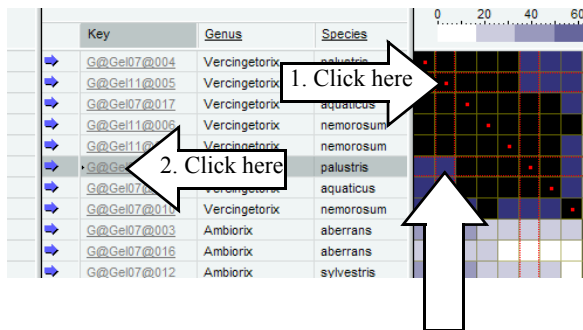


Figure 4-15. Workflow for finding a similarity value between two entries.

4.1.14.7 By double-clicking on a similarity block or value, you can pop up the detailed comparison between the two entries (4.1.2).

4.1.14.8 To export a tab-delimited text file of the similarity matrix, select *File > Export similarity matrix*.

This text file contains the entry keys as descriptors. You can export a text file which contains the same descriptors with the corresponding information fields:

4.1.14.9 Export the information fields with *File > Export database fields*.

4.1.15 Group statistics

As mentioned in 4.1.12, a number of statistical functions will need the presence of Groups. These Groups statis-

tics functions are based upon the Groups the user has defined. We have explained earlier how to define Groups (see 4.1.12.2 to 4.1.12.10), and if you have gone through the dendrogram display functions (4.1.11), the Groups are already present on the dendrogram.

•K-means partitioning

One function to let the software automatically determine Groups is the mathematical function *K-means partitioning*. The user first creates Groups based upon one or more strains (e.g. type strains). Then, the program automatically calculates for each entry of the cluster analysis in which group it fits best. This fitting can be based upon *Average* similarity with the group, upon the highest similarity (*Nearest neighbor*), or upon the lowest similarity (*Furthest neighbor*). Obviously, the partitioning process must be iteratively executed, since by adding an entry to a group, the average similarity of the group as well as the highest and lowest similarities with entries may change.

4.1.15.1 To illustrate the partitioning method, we select **RFLP1**.

4.1.15.2 Select the root and select all entries on the dendrogram with CTRL + left-click.

4.1.15.3 Remove all group assignments with *Groups > Assign selection to > None*.

4.1.15.4 Select one or a few entries per cluster, each time assigning a different group color to them (see Figure 4-16 for an example). Do not forget to unselect all entries before you start defining a next group.

4.1.15.5 In the menu, select *Groups > Partitioning of groups*, which allows you to choose between the three options described above (Figure 4-17).

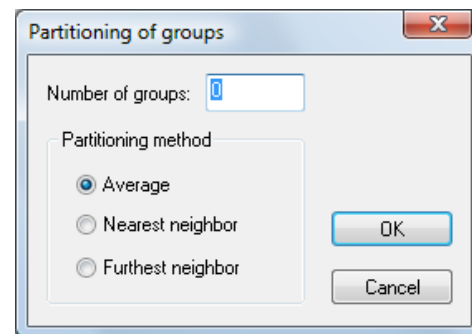


Figure 4-17. Partitioning of groups dialog box.

4.1.15.6 Leave *Number of groups* on zero so that the program will only use the groups we have defined manually.

4.1.15.7 Select *Nearest neighbor*, which will place a new entry in the group containing the highest individual similarity with that entry.

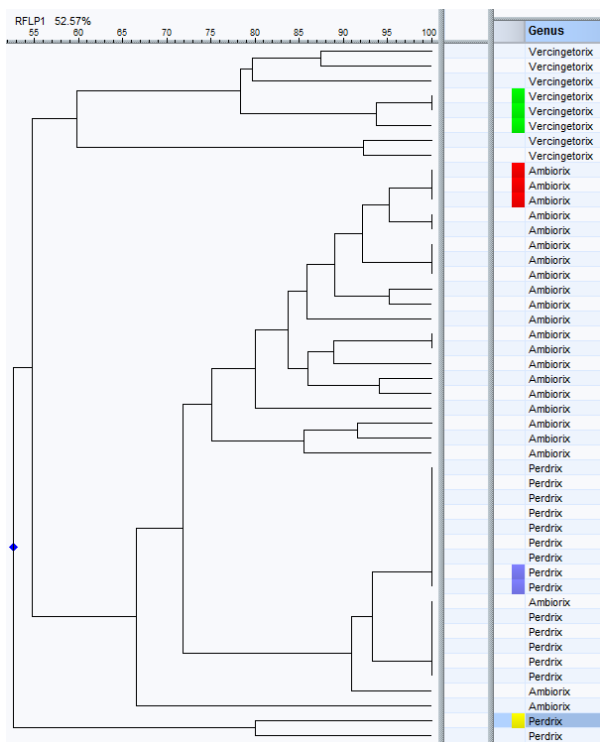


Figure 4-16. Example of manual group assignment in preparation of a partitioning process.

4.1.15.8 Press **<OK>** to execute the partitioning. After partitioning, all entries belong to one of the defined groups.

Note that these groups do not necessarily correspond exactly to the visual clusters on the dendrogram. This can be the case if the clusters on the dendrogram are not well-defined or inconsistent. A second cause of aberrations is the oversimplification of complex matrices by the UPGMA algorithm.

4.1.15.9 As an alternative, you can also select **Groups > Partitioning of groups**, specifying a predefined number of groups, e.g. 3.

4.1.15.10 Press **<OK>** to partition into 3 groups. The program now has defined the 3 most relevant groups in the comparison.

• Group separation statistics

These statistical methods determine the stability of the defined groups, whether they are defined manually, derived from clusters, using K-means partitioning, or created from an information field. They involve the *Jackknife* method and the “*Group violations*” measurement.

4.1.15.11 With **Groups > Group separation**, the separation between the defined groups are investigated.

The *Group separation settings* dialog box is shown, allowing a number of choices to be made (Figure 4-18).

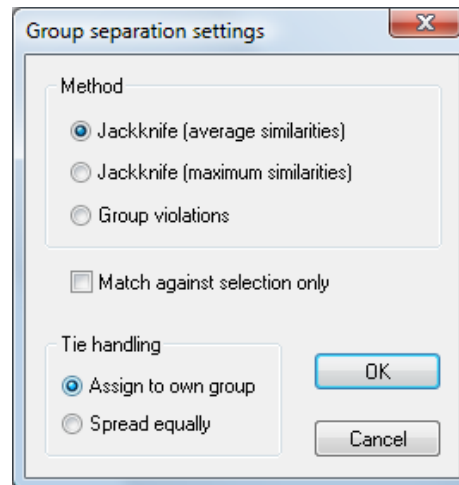


Figure 4-18. *Group separation settings* dialog box.

The principle of the *Jackknife* method is to take out one entry from the list, and to identify this entry against the different groups. This can be done by calculating the *Average similarities* with each group, or finding the *Maximum similarities* with each group. This is done for all entries (unless *Match against selection only* is checked). The percentage of cases that entries are identified to the group they were originally assigned to, is a measure of the internal stability (significance) of that group. The percentage of cases that entries are identified to another group than originally assigned to, is indicative of lack of internal stability.

Using *Match against selection only*, you can let the program calculate the matches against a selection you made in the comparison, rather than against all entries of the groups.

In cases where an entry has an equal match with a member of its own group and a member of another group (a “tie”), there are two equally valid interpretations possible. The program can handle such ties in an ‘optimistic’ way, i.e., by always assigning equal matches to their own group, or in a ‘realistic’ way, by spreading ties equally between the own and the other groups.

The way ties are handled can be chosen in the *Settings* dialog box under *Tie handling*. This includes two options, *Assign to own group* and *Spread equally*.

4.1.15.12 Click **<OK>** with the default settings to display the *Group separation statistics* window (Figure 4-19).

Note that the values in the matrix are not reciprocal, i.e., the matrix is not symmetric! The number of misidentifications for members of group 1 are given in column 1 (Figure 4-19), for members of group 2 in column 2, etc. In Figure 4-19 for example, 0% of group 1 members are identified as group 2, but 4.7% of group 2 members are identified as group 1.

The overall quality of the Group separation is indicated in the status bar of the window; it is the average of the

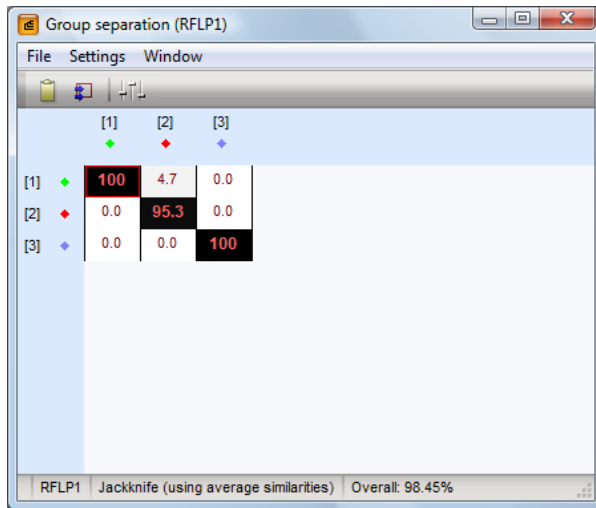




Figure 4-19. Group separation statistics window.

diagonal, i.e. the total percentage of correct identifications.

When the Jackknife method is used, a value (or cell) in the group separation matrix can be selected, and with

the  button or *File > Select cell members*, the entries contributing to this cell will be selected in the *Comparison* window. The method is useful to identify entries that fit well or do not fit well in their assigned groups.

*NOTE: The interpretation of matching and non-matching entries is less easy when the **Spread equally** function has been chosen, since in that case, some entries may fall outside their group “unexpectedly” when they have an equally high score with another group.*

4.1.15.13 Click  or select *Settings > Statistics* to call the *Settings* dialog box again.

4.1.15.14 Under *Method*, select *Group violations*. Figure 4-19 is based on group violations between three groups partitioned as above (**RFLP1**).

The group violations method compares all the similarity values within a group with those between a group and the other groups. All the values occurring in the overlap zones (see Figure 4-20) are considered “violations” of the integrity of the group.

The percentage of group violations for group A is the number of external entries scoring higher than the lowest internal values over the total number of similarity values considered. The percentages seen in the diagonal of the matrix are the percentages of **non-violations**.

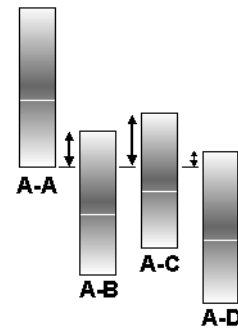


Figure 4-20. Schematic representation of internal similarity range of group A (A-A), and similarity ranges with other groups (A-B, A-C, and A-D). The overlapping values are group violations.

4.1.15.15 The *Group statistics* can be copied to the clipboard using *File > Copy to clipboard* or by pressing the




button.



4.1.16 Printing a cluster analysis

When printing from the *Comparison* window, InfoQuest FP first shows a print preview. This print preview shows the same information as is shown in the panels of the *Comparison* window: for example, a dendrogram, one or more images from different experiments, metrics scale, etc. One exception is the similarity matrix: the print preview does not print matrices unless you explicitly select it in the print preview. The preview looks exactly as it will look on printed pages. You can edit the layout of the print preview by adjusting the space allowed for the different items (dendrogram, image(s), information fields), by changing the size of the figure to fit on one or more pages, etc.



NOTE: The print preview will always display dendrogram, experiment image and database information arranged from left to right, irrespective of the configuration of the Comparison window.

4.1.16.1 In the *Comparison* window, select *File > Print preview* or press the  button, which opens the *Comparison print preview* window (Figure 4-21).



The *Comparison print preview* window is divided in two panels, which are both dockable (see 1.6.4 for display options of dockable panels). The *Overview* panel shows an overview of the pages that will be printed, with the actual page in yellow. In the *Print preview* panel, the actual page is shown.


4.1.16.2 With the PgUp and PgDn keys or *Edit > Previous page*  and *Edit > Next page* , you can thumb through the pages that will be printed out.

4.1.16.3 It is possible to zoom in and out on a page using

Edit > Zoom in  and **Edit > Zoom out**  (shortcuts CTRL+PgUp and CTRL+PgDn on the keyboard) or by using the zoom slider (see 1.6.7) in the *Preview* panel.

4.1.16.4 When zoomed, the horizontal and vertical scroll bars allow you to scroll through the page.

4.1.16.5 The whole image can be enlarged or reduced with **Layout > Enlarge image size**  or **Layout > Reduce image size** .

4.1.16.6 If a similarity matrix is available, it can be shown and printed with **Layout > Show similarity matrix** or .

4.1.16.7 With **Layout > Show comparison information**, the name of the comparison (if already saved) and the number of entries are indicated on top of the first page.


4.1.16.8 It is possible to display a header line with the database field names when **Layout > Show field names** is selected.

On top of the preview page, there are a number of small yellow slide bars (Figure 4-21). These slide bars represent the following margins, respectively:

- Left margin of the whole image;
- If dendrogram shown, right margin of dendrogram;
- If image shown, right margin of image;
- If groups are defined, right margin of groups;
- Right margin of entry keys or group codes (if not hidden);
- Right margins of different information fields (except the hidden fields);
- If the similarity matrix is shown, right margin of matrix.

Left on the first preview page, there are two slide bars: representing the top margin of the whole figure and the lower margin of the header, respectively. Left on the last page, there is one slide bar representing the bottom margin of the image.

Each of these slide bars can be shifted individually to reserve the appropriate space for the mentioned items. The image is printed exactly as it looks on the preview.

4.1.16.9 You can preview and print the image in full color with **Layout > Use colors** or .

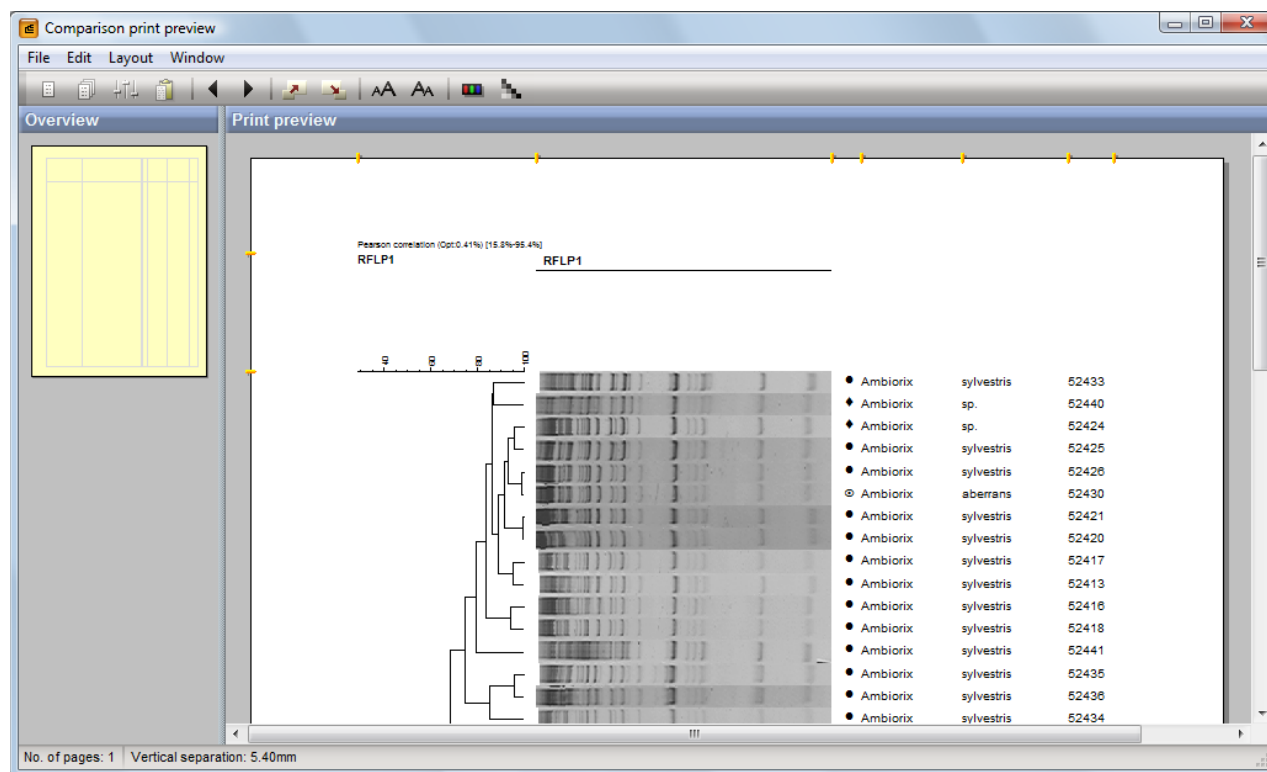






Figure 4-21. The *Comparison print preview* window.

4.1.16.10 In addition, the menu command **File > Printer setup** or  allows you to set the paper orientation, the margins, and other printer settings for the default printer.

4.1.16.11 If the preview is covering more than one page, you can click on a specific page in the *Overview* panel to select a page from the range.

4.1.16.12 With **File > Print this page** or , the current page is printed.

4.1.16.13 Use **File > Print all pages** or  to print all pages at once.

4.1.16.14 If you want to export the image to another software package for further editing, use **File > Copy page to clipboard** or .

This function provides a choice between the Windows *Enhanced Metafile* format, i.e. the standard clipboard exchange format between native Windows applications (default), or a *bitmap* file with 75 dpi, 150 dpi, 300 dpi or 600 dpi resolution. Many software applications, although supporting the enhanced metafile format, are unable to properly import some advanced InfoQuest FP clipboard files that make use of mixed vector, bitmap and (rotated) text components. If you experience such problems, you should select a bitmap file to be exported, or use another software application (or a more recent version of the same software) to import the graphical data.

With the *Copy page to clipboard* function, only the current page is copied to the clipboard. If you want the whole image to be copied to the clipboard, first reduce the size of the image (4.1.16.5).

*NOTE: When preferred, the image of a fingerprint type can be shown and printed with a space between the gelstrips. To do so, open the Experiment type window in the program's InfoQuest FP main window (under **Fingerprint types**) and select **Layout > Show space between gelstrips**.*

4.1.16.15 Select **File > Exit** to close the *Comparison print preview* window.

4.1.17 Exporting rendered trees

In publications and presentations, particularly in a phylogenetic context, a dendrogram is sometimes represented as a real tree with a stem and branches. Such representations can be achieved in InfoQuest FP using the *rendered tree* option in the *Comparison* window (Figure 4-22). This option should be used with care, as it will only produce acceptable pictures from a very

limited number of entries and with fairly equidistant members.

Rendered trees can be created from a standard rooted tree in the *Comparison* window as well as from unrooted phylogenetic trees (Parsimony and Maximum Likelihood).

4.1.17.1 In the *Comparison* window, create a dendrogram containing a small and not too heterogeneous group of entries (e.g. 10 entries).

4.1.17.2 Select **Clustering > Rendered tree export**. A *Rendered tree settings* dialog box (Figure 4-23) prompts for a number of settings:

- **Hide branches if shorter or equal to** allows all entries that are very similar to be grouped together at one branch tip. This allows simpler trees to be produced and may avoid starlike branch tips to occur.
- **Hide distance labels if shorter or equal to** sets a minimum for the distance values to be shown on the branches. If many short distances occur, the labels may overlap, which can be avoided by only allowing larger distance to be shown. If the value is set to 100 (or more), no distance labels will be shown.
- **Tree type** can be *Rooted* or *Unrooted*. If the dendrogram is rooted by nature (e.g. UPGMA) it makes no sense to export an unrooted tree from it.
- **Display type** can be *Rendered*, i.e. with a thicker stem and smooth, gradually narrowing branches, or *Outlines*, where stem and branches are represented by straight lines.
- **Display field** is a pull-down listbox where one of the available database fields can be chosen.

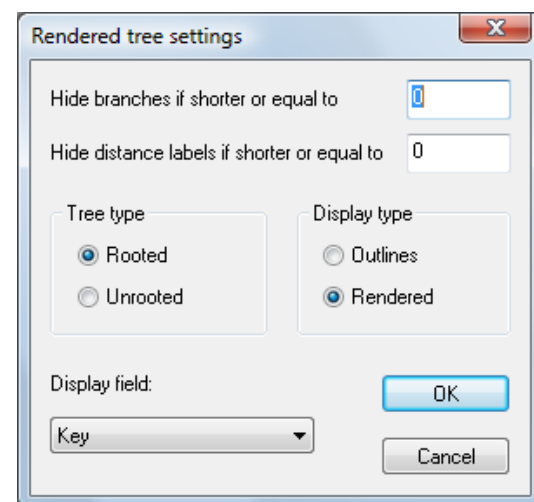


Figure 4-23. *Rendered tree export settings* dialog box.

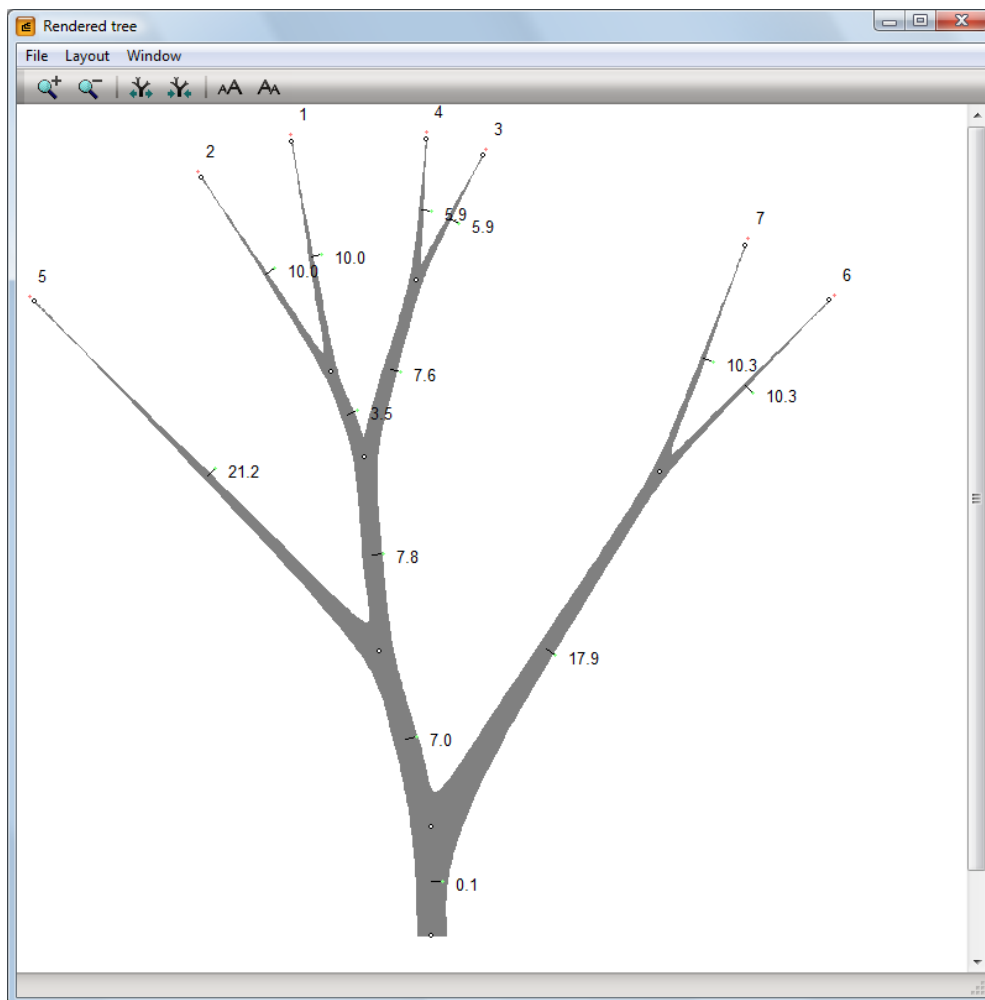


Figure 4-22. The Rendered tree window.

NOTE: In case the information fields are too long, it is possible to replace them by a group code, if groups are defined (4.1.12). The use of group codes is explained in 4.1.12.23.

4.1.17.3 Rendered trees can also be exported from parsimony and maximum likelihood trees (see Section 4.9). If a *rooted* rendered tree is exported, the highlighted branch of the unrooted tree will be used to create the root.

4.1.18 Analysis of the congruence between techniques

As soon as multiple techniques are used to study the relationships between organisms, the question arises how congruent the groupings obtained using the different techniques are. It is also interesting to compare the techniques by the level at which they discriminate the entries, in other words, the taxonomic depth of the techniques.

An evident way to perform such a study is by comparing the similarity matrices obtained from the

different experiment types used. By plotting the corresponding similarity values in an X-Y coordinate system, one can easily observe the kind and degree of concordance at a glance. InfoQuest FP even calculates a regression curve through the plot.

4.1.18.1 In the *InfoQuest FP* main window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as STANDARD (see 4.1.3.2 to 4.1.3.4).

4.1.18.2 Check whether a cluster analysis is present for each experiment type (except the composite data sets) by right-clicking on the experiment name in the *Experiments* panel. A floating menu appears.

If the menu items *Show dendrogram* and *Show matrix* display in black (enabled), a cluster analysis is present for the experiment and no further calculation is needed.

4.1.18.3 If the menu items *Show dendrogram* and *Show matrix* display in gray (disabled), it means that no cluster analysis is present for the experiment. In that case, select *Calculate cluster analysis (similarity matrix)* from the floating menu.

4.1.18.4 In the *Comparison* window's menu, select *Clustering > Congruence of experiments*.

The *Experiment congruence* window (Figure 4-24) shows both a matrix of congruence values between the techniques (experiment types) and a dendrogram derived from that matrix.

The default method to calculate the congruence between two experiment types is by using the Pearson product-moment correlation coefficient. An alternative coefficient is *Kendall's tau*. The principle of Kendall's tau is as follows: if value A is higher than value B in experiment 1, then corresponding value A of experiment 2 should also be higher than corresponding value B of experiment 2. The less infringements on this statement, the more congruent the techniques are. Kendall's tau has the advantage over Pearson correlation that non-linear correlations still have significant scores. In addition, the significance of the correlation between techniques is shown (green values) when the Kendall's tau is selected, as well as the standard deviation on the values.

4.1.18.5 The congruence matrix can be exported as an enhanced metafile using *File > Copy image to clipboard*, or printed directly with *File > Print image*.

4.1.18.6 A tab-delimited text export of the congruence matrix can be achieved with *File > Export similarity values*. In case Kendall's tau is used, this will export the errors and significance values as well.

4.1.18.7 Select *Calculate > Experiment correlations*.

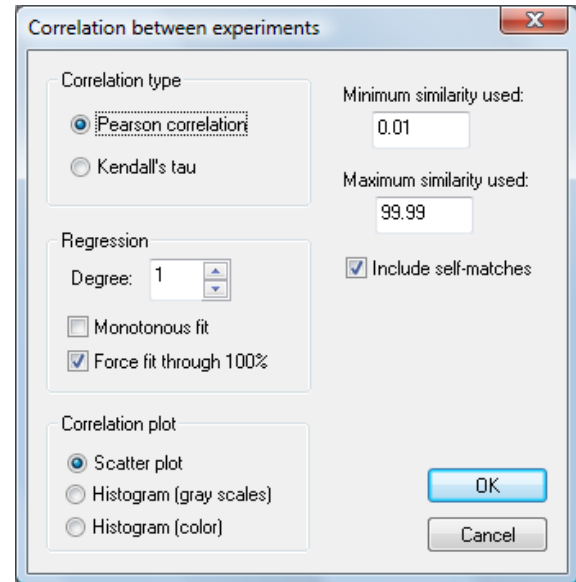


Figure 4-25. *Correlation between experiments settings dialog box*.

The *Correlation between experiments* dialog box that pops up shows the settings for the calculation of correlation between the experiment types (Figure 4-25).

4.1.18.8 In the *Correlation type* box, select *Kendall's tau*.

The *Minimum similarity used* and *Maximum similarity used* allow a range of similarity values to be specified within which the analysis is done. Normally, one can

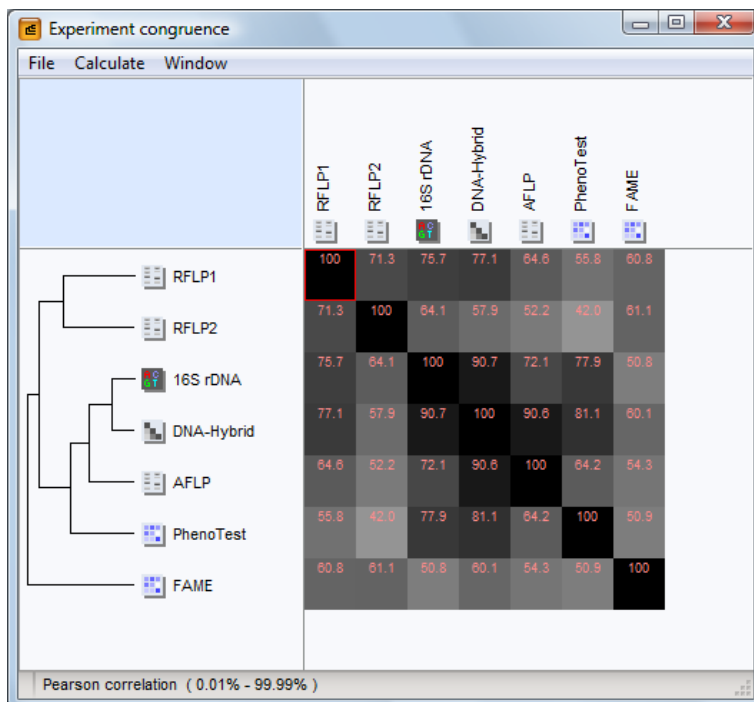


Figure 4-24. *Experiment congruence window*.

enter 0% and 100%, respectively, for these values. **Include self-matches** is an option which gives the user the choice whether to include entries compared with themselves. Obviously, self-matches are always 100% and may thus influence the correlation obtained between two experiment types.

4.1.18.9 Enter zero and 100% respectively for minimum and maximum similarity used, and uncheck (disable) **Include self-matches**.

The **Regression** determines the kind of best-fitting curve that is calculated through a *Similarity plot* of two experiment types (see 4.1.18.11). You can enter the **Degree** of the regression (first degree is linear, second degree is a quadratic function, etc.). If there is any concordance between techniques, one should expect that the function increases monotonously; with **Monotonous fit**, only such functions are allowed. With **Force through 100%**, the program will force the regression curve to pass through 100% for both techniques. In other words, if entries are seen identical in one technique, you would expect that they are seen identical in another technique as well. If you do not wish to see the regression in the similarity plot, enter 0 as **Degree**.

Under **Correlation plot**, you can choose **Scatter plot** to plot each pair of similarity values as one dot in a *Similarity plot* between two experiment types (see 4.1.18.11). Especially for very large data sets resulting into dense scatter plots, it can be useful to average the number of dots in a given area and represent that average rather than the individual pairs. This can be achieved with **Histogram (gray scales)** and **Histogram (color)**. When color is chosen, a multi color scale is used that ranges continuously from white over blue, green, yellow, orange, and red to black.

4.1.18.10 Select a 3rd degree, **Monotonous fit**, and **Force through 100%**. Choose **Scatter plot** as correlation plot type and press <OK>.

4.1.18.11 Click on a value in the similarity matrix and **Calculate > Similarity plot**.

The similarity plot between the two selected experiments appears (Figure 4-26), with a third degree regression drawn through it. Excluded values (due to **Minimum similarity used** and **Maximum similarity used**) are shown in gray.

4.1.18.12 You can click on any dot in the similarity plot to pop up a detailed pairwise comparison between the two entries (see 4.1.2).

4.1.18.13 While hovering the mouse pointer over the plot, it becomes a lasso selection tool. You can make a selection of dots by dragging the mouse over the plot. The corresponding database entries are selected. The

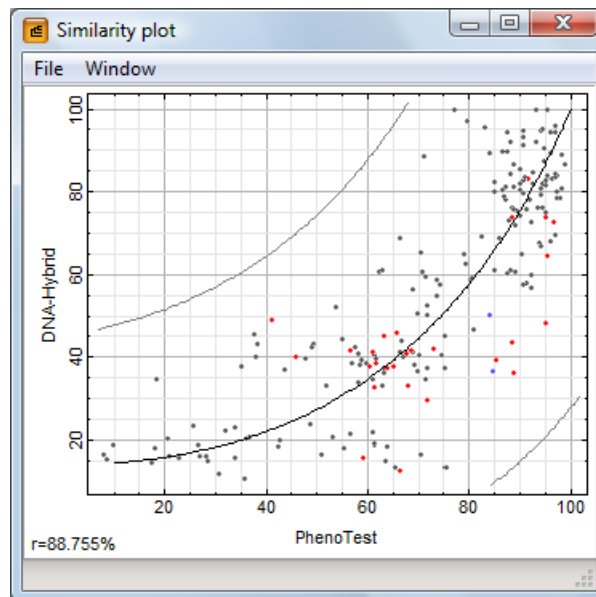


Figure 4-26. Similarity plot between two experiments.

selection status of the database entries is displayed as follows on the plot:

- Dots resulting from two selected entries are displayed in blue;
- Dots resulting from one selected entry and one non-selected entry are displayed in red;
- Dots resulting from two non-selected entries are displayed in black.

NOTE: In a comparison of n entries, every entry forms a dot with $n-1$ other entries. As such, if you select one dot, $n-1$ other dots will appear in red. For the same reason, a restricted selection of dots on the plots can easily result in all dots becoming either red or blue (and all entries being selected).

It is also possible to display the defined groups on the similarity plot (see 4.1.12. for working with groups). First we define groups for this comparison as follows:

4.1.18.14 In the *Comparison* window, right-click in the header of the 'Genus' column and select **Create groups from database field** from the floating menu.

4.1.18.15 In the next dialog box, make sure **Subdivide existing groups** in unchecked and press <OK>.

4.1.18.16 In the *Similarity plot* window, select **File > Show group colors**.

Since each dot is composed of two entries, the dots may be composed of two colors (see Figure 4-27).

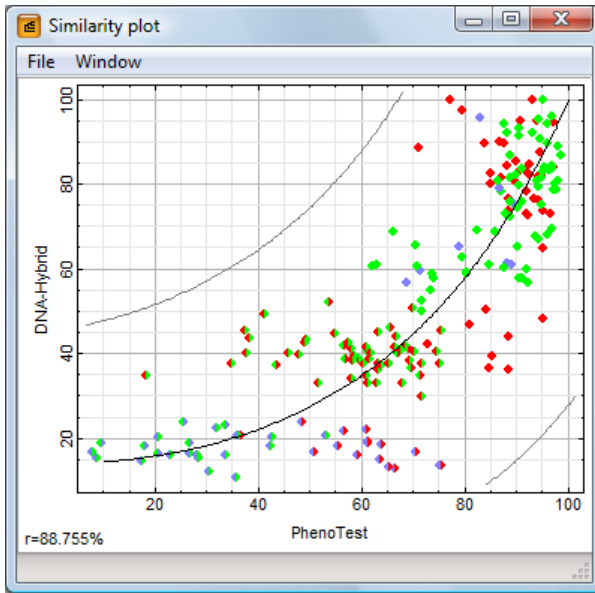


Figure 4-27. Similarity plot of experiments with group colors shown.


4.2 Cluster analysis of fingerprints CL FP


4.2.1 Fingerprint comparison settings

The comparison settings of fingerprint types will be illustrated using the **DemoBase** database.

4.2.1.1 Open the database **DemoBase**.

4.2.1.2 Open the comparison **All** if already existing. Alternatively, select all entries except STANDARD (see 4.1.3.2 to 4.1.3.4) and create a new comparison (4.1.3.4).

4.2.1.3 Select **RFLP1** in the *Experiments* panel and show the normalized gel image by pressing the  button.

4.2.1.4 Select **Clustering > Calculate > Cluster analysis (similarity matrix)**. You can also press the  button, in which case the following menu pops up (Figure 4-28).

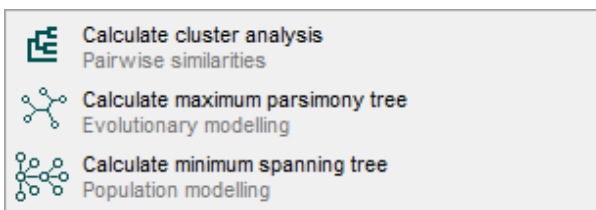


Figure 4-28. Cluster analysis menu popped up from the dendrogram button.

The *Comparison settings* dialog box allows you to specify the similarity coefficient to calculate the similarity matrix, and the clustering method to be applied (see Figure 4-29).

Two coefficients provide similarity based upon densitometric curves; the Pearson product-moment correlation (*Pearson correlation*) and the *Cosine coefficient*.

Four different binary coefficients measure the similarity based upon common and different bands:

1. The *Jaccard* coefficient

$$S_J = \frac{N_{AB}}{N_A + N_B - N_{AB}}$$

2. The *Dice* coefficient

$$S_D = \frac{2N_{AB}}{N_A + N_B}$$

3. The *Jeffrey's X* coefficient

$$S_X = \frac{1}{2} \left(\frac{N_{AB}}{N_A} + \frac{N_{AB}}{N_B} \right)$$

4. The *Ochiai* coefficient

$$S_O = \frac{N_{AB}}{\sqrt{N_A N_B}}$$

A fifth coefficient, *Different bands*, is essentially a distance coefficient as it simply counts the number of different bands in two patterns. It is converted into a similarity by subtracting this distance value from 100. If you select one of these binary coefficients, you can enable the *Fuzzy logic* option: instead of a yes/no decision whether two bands are matching or not, the program lets the matching value gradually decrease with the distance between the bands. The *Area sensitive* option makes the coefficient take into account differences in area between two matching bands: if for each matching band the areas on both patterns are exactly the same, the coefficient reduces to a normal binary coefficient; the more the areas differ, the lower the similarity will be. The *Relaxed doublet matching* option allows a single band to match with two bands of a doublet, on condition that both bands of the doublet fall within the tolerance window from the single band.

Among the dendrogram types, the program offers the Unweighted Pair Group Method using Arithmetic averages (*UPGMA*), the *Ward* algorithm, the *Neighbor Joining* method, and two variants of UPGMA, namely *Single linkage* and *Complete linkage*. The option *Advanced* is explained in Section 4.10.

4.2.1.5 Select *Dice* and *UPGMA*.

The *Position tolerances* button allow you to specify the maximum allowed distance between the positions of two bands on different patterns, for which these bands can be considered as matching.

4.2.1.6 Press the **<Position tolerances>** button.

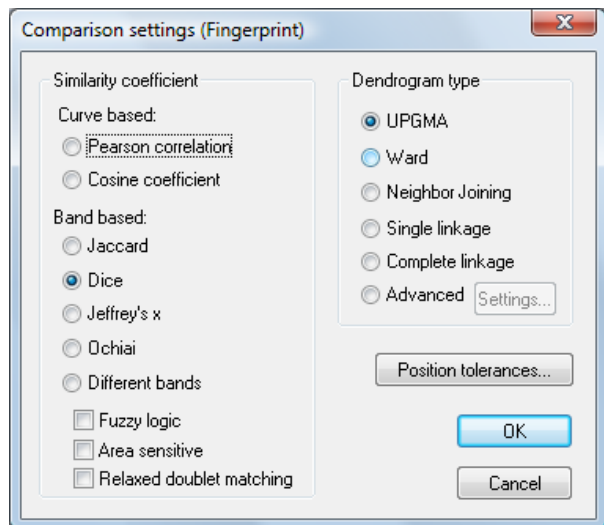


Figure 4-29. The *Comparison settings* dialog box.

The *Position tolerance settings* dialog box for the fingerprint type is popped up (Figure 4-30).

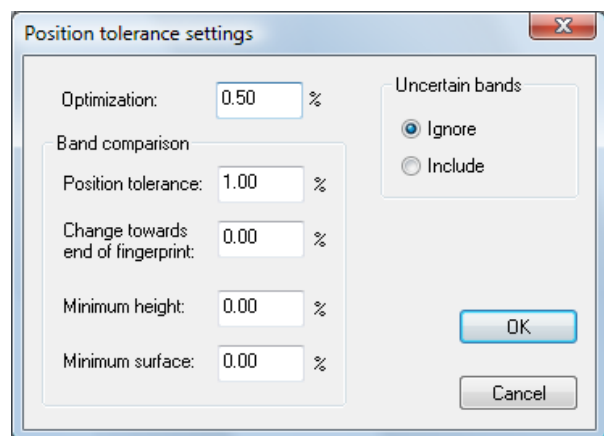


Figure 4-30. *Position tolerance settings* dialog box of a fingerprint type.

The *Position tolerance* is the maximal shift allowed (in percentage of the pattern length) between two bands to consider them as matching. This parameter only applies to band matching coefficients. With *Change towards end of fingerprint*, you can specify a gradual increase or decrease in tolerance. In 4.2.4, we discuss how to have the program automatically calculate the optimal position tolerance settings for your fingerprint type.

The *Optimization* is a shift that you allow between any two patterns and within which the program will look for the best possible matching. This parameter applies for both curve-based and band matching coefficients. To understand the utility of optimization in addition to tolerance, see the example in Figure 4-31.

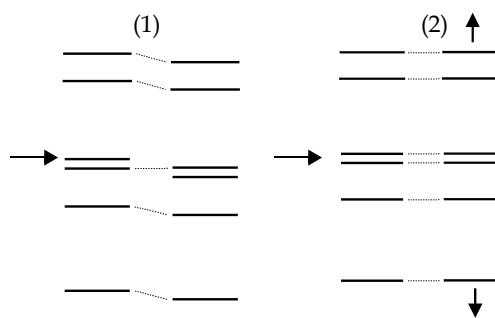


Figure 4-31. Effect of position tolerance (1) and optimization (2) on the matching between patterns.

In 4.2.4, we also discuss how the program can automatically find the best optimization value for your fingerprint type.

With minimum height and minimum surface, you can exclude weak or irrelevant bands.

NOTE: Both the Comparison settings and the Position tolerance settings are stored along with the fingerprint type. The same dialog boxes can be called from the Experiment type window settings ().

The *Uncertain bands* option allows you to either include uncertain bands or ignore them (see 3.2.6.1). When *Ignore* is chosen, uncertain bands are not taken into account. This means that in a pairwise comparison, an uncertain band is not penalized if there is no matching band on the other pattern. Conversely, if there is a band on the other pattern that matches an uncertain band, it will also be ignored in that comparison. When *Include* is chosen, uncertain bands are treated in the same way as certain bands, which means that an uncertain band which is not complemented by a band in the other pattern, is penalized.

NOTE: The Ignore option will only work when both Fuzzy logic and Area sensitive are disabled in the Comparison settings dialog box (Figure 4-29).

4.2.1.7 Enter a position tolerance of 1%, an optimization of 1%, a change of 0%, and a minimum height and minimum surface of 0%, and press <OK>.

4.2.1.8 Press <OK> again in the *Comparison settings* dialog box to start the cluster analysis.


When finished, the dendrogram and the similarity matrix are shown. For more information about the panels in the *Comparison* window, see 4.1.3.


4.2.2 Fingerprint display functions

For fingerprint types, additional information can be shown in the *Experiment data* panel and data can be


exported as text file. These features will be illustrated with the fingerprint type **RFLP1**.

4.2.2.1 Make sure the image of **RFLP1** is shown in the *Experiment data* panel (see 4.1.3.10).


4.2.2.2 Press the  button or **Layout > Show metrics scale** to display the molecular weight scale of the selected fingerprint type.

4.2.2.3 Select **Layout > Show bands** or press  to show or hide the band positions in the *Experiment data* panel.

4.2.2.4 One can also show only the band positions in the *Experiment data* panel without showing the actual image.

Select **Layout > Show image** or press  to show or hide the image.

4.2.2.5 When bands are shown on the image, they can be exported as a tab-delimited file with **File > Export bands**. The export file, popped up as **result.txt** in Notepad, contains the key of the entry, and a list of band positions as relative run lengths (in percent) and molecular weight (in case a regression curve is calculated for the reference system used; see 3.2.9).

4.2.2.6 Select **Layout > Show densitometric curves** or press  to show or hide small densitometric curves in the *Experiment data* panel. One can also show only the curves without showing the actual image (see 4.2.2.4).

4.2.2.7 When densitometric curves are shown on the image, they can be exported as a tab-delimited file with **File > Export densitometric curves**. The export file, popped up as **result.txt** in Notepad, contains the list entry keys separated by tabs, and a list of densitometric curves, of which the curves are listed as columns, separated by tabs.

In case only densitometric curves are available (e.g. in case of profiles from automated sequencers), it can be useful to display the curves as pseudo gelstrips (reconstructed images). This option is selected in the *Fingerprint type* window as follows:

4.2.2.8 In the *InfoQuest FP main* window, double-click on a fingerprint type (e.g. **RFLP1**) in the *Experiments* panel.

4.2.2.9 In the *Fingerprint type* window, select **Layout > Show curves as images**. If densitometric curves are now shown in a comparison (4.2.2.6), they will be displayed as pseudo gelstrips.

In case densitometric curves have different intensities, the densitometric curves can be rescaled so that each curve fills the full available intensity range specified for the fingerprint type. This can be achieved as follows:

4.2.2.10 In the *Fingerprint type* window, select **Layout > Rescale curves**. If densitometric curves are now shown in a comparison (4.2.2.6), they will all be displayed with equal intensity.

4.2.2.11 The image of patterns can be shown with a space between the gelstrips. To do so, open the *Fingerprint type* window in the program's *InfoQuest FP main* window and select **Layout > Show space between gelstrips**.

4.2.3 Defining 'active zones' on fingerprints

When clustering fingerprints, one is not necessarily interested in comparing complete patterns. For example, when the loading well or the loading dye is comprised within the fingerprints, it may be better to exclude such a region from the cluster analysis.

For each fingerprint type, it is possible to define *excluded regions* which are applied for all comparisons using this fingerprint type.

4.2.3.1 Select any entry in the **DemoBase** database that contains a fingerprint of **RFLP1**.

4.2.3.2 Open the *Fingerprint type* window for **RFLP1** in the *Experiments* panel.

At the bottom of the window, the fingerprint of the selected database entry is shown (Figure 4-32).

4.2.3.3 To exclude a region for comparison, hold the left mouse button and the SHIFT key simultaneously while dragging the mouse pointer over the fingerprint.


The excluded region becomes cross-hatched in red. In the bottom part of the window, the parts of the fingerprints that are included for comparison are shown as percentages (see Figure 4-32).

4.2.3.4 To include a region, hold the left mouse button (without holding the SHIFT key), while dragging the mouse pointer over the fingerprint.

4.2.3.5 You can for example exclude the top 15% and the end 15% of the fingerprints.

NOTE: You can exclude / include multiple regions. The defined regions apply both to comparisons based on densitometric curves and to comparisons based on band matching. Bands falling within an excluded region will not be considered for cluster analysis and band matching analysis.

4.2.3.6 You can specify the exact start and end of the active zone(s) using a script available on the Bio-Rad website. The scripts can be launched from the *InfoQuest FP main* window, using the menu **Scripts > Browse**

Internet, or , and then selecting *Fingerprint related tools* > *Set active zones*.

4.2.3.7 Back in the *Comparison* window, select **RFLP1** and *Clustering* > *Calculate* > *Cluster analysis (similarity matrix)*, to recalculate the dendrogram using the excluded regions.

4.2.4 Calculation of optimal position tolerance optimization and settings

InfoQuest FP possesses a very interesting option to calculate automatically the optimal settings for position tolerance and optimization for a given fingerprint type. The principle is as follows: the user selects a number of entries which he or she wants to cluster into a comparison. The program will calculate similarity matrices with varying position tolerances. Within a limited range, the optimal position tolerance value yields the matrix with the highest group contrast: scores as high as possible within groups and as low as possible between groups. This translates in the highest standard deviation on the matrix of similarity values. The same process can be launched to find the best optimization range. Given the principle of the method, it is important to select entries belonging to different groups or showing enough heterogeneity.

The best way to proceed is to create a comparison with *Groups* (see 4.1.11) already defined, e.g. based upon cluster analysis or partitioning (see 4.1.15). The program

will then optimize the intergroup separation based upon these groups. If no groups are defined, the standard deviation of the whole matrix is optimized, which also works in case the comparison contains some groups of more related patterns.

In case you choose a correlation coefficient based on densitometric curves, only the optimization value is needed, and the program will calculate this value. However, in case you apply a band matching coefficient, for example Dice or Jaccard, both the tolerance and optimization values are important. Therefore, the program can also calculate the optimal setting for both values in combination with each other. If n matrices are to be calculated for the tolerance value, and n matrices for the optimization, the combined process requires $n \times n$ matrices to be calculated. In addition, each value from each matrix is to be calculated a number of times within the tolerance/optimization boundaries, in order to find the highest value. No need to argue that this process is extremely time-consuming; it should only be executed on very small numbers of entries. Alternatively, both parameters can be calculated separately.

Given the time needed to calculate $n \times n$ matrices with increasing tolerance applied, we recommend to first calculate the optimization value using Pearson coefficient, and then, using this value, calculate the optimal position tolerance setting. This is done as follows:

4.2.4.1 In the *InfoQuest FP main window* with **DemoBase** loaded, open comparison **All**, or create a comparison containing all entries except the STANDARDS (see 4.1.9).

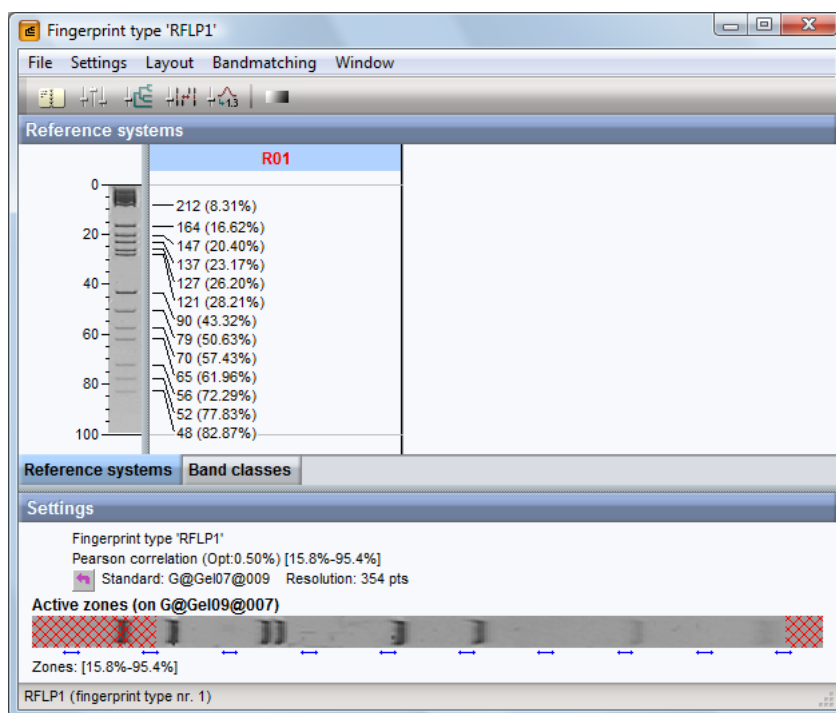


Figure 4-32. *Fingerprint type* window with excluded regions defined (see arrows).

4.2.4.2 In case no groups are defined, select the **Genus** database field and *Groups > Create from database field*.

4.2.4.3 Select **RFLP1** in the *Experiments* panel, and *Clustering > Tolerance & optimization analysis*.

The *Comparison settings* dialog box appears (see Figure 4-29) where you can select the coefficient and clustering method. Only the coefficient will influence the calculation of the optimization.

4.2.4.4 Select *Pearson correlation* under *Similarity coefficient* and press <OK>.

The program now calculates the best optimization value. When finished, the *Position tolerance analysis* window appears (Figure 4-33) showing the group separation in function of the allowed optimization in the right diagram.

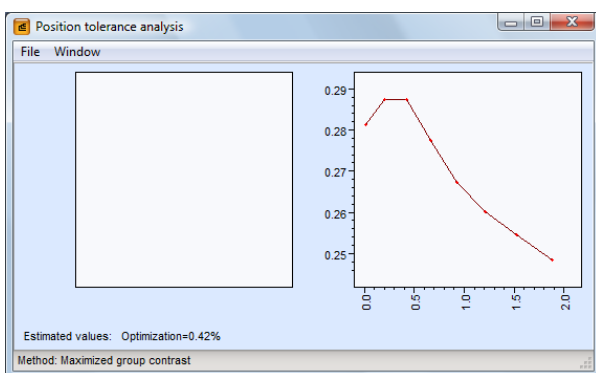


Figure 4-33. Position tolerance analysis. Optimization analysis shown for a curve-based coefficient.

The ideal optimization value is shown (bottom) and is automatically saved in the settings for the experiment type.

4.2.4.5 Close the window with *File > Exit* and select *Clustering > Tolerance & optimization analysis* again.

4.2.4.6 This time, select the *Dice* coefficient and press <OK>.

4.2.4.7 The program asks "*Do you wish to estimate the optimization parameter?*". Answer <No>.

The program now calculates the best position tolerance value for band matching. When finished, the *Position tolerance analysis* window (Figure 4-34) shows the group separation in function of the allowed band matching tolerance in the left diagram.

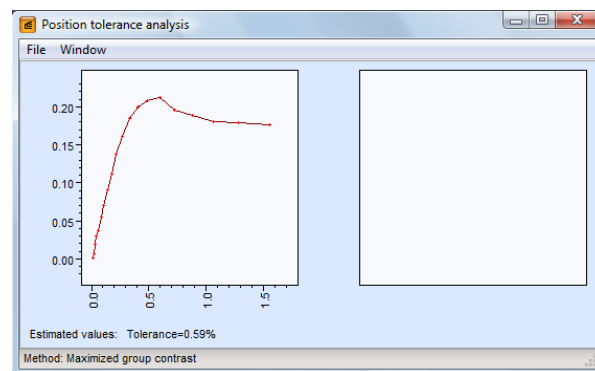


Figure 4-34. Position tolerance analysis. Position tolerance analysis shown for a band matching coefficient.

The position tolerance value is shown (bottom) and is automatically saved in the settings for the experiment type.

4.2.4.8 Close the window with *File > Exit*.

4.3 Band matching and polymorphism analysis

4.3.1 Introduction

Band matching is a comparison function which applies only to fingerprint types. It can be executed on any selection of entries from the database. In a first step, InfoQuest FP divides all the bands found among the selected patterns into *classes of common bands* (1 to 8 in Figure 4-35). As such, every band of a given pattern belongs to a class, and conversely, every band class is represented by a band on one or more patterns. The result is shown in Figure 4-35.

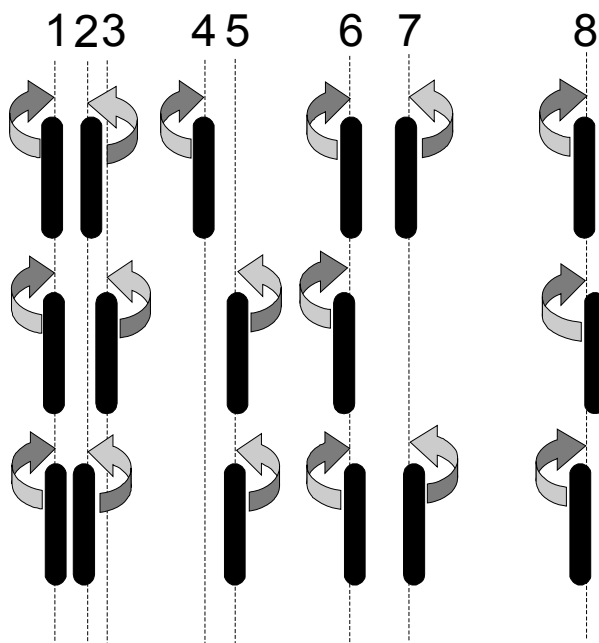


Figure 4-35. Comparative quantification: bands are assigned to classes.

Clearly, the number of band classes distinguished will depend on the *optimization* and the *position tolerance* that is allowed between bands considered as matching. For example, when a larger position tolerance is specified, more bands will be grouped in the same class than when a small position tolerance is chosen. In Figure 4-35, taking a larger position tolerance would have resulted in the merging of band classes 2 and 3, whereas a smaller position tolerance would have resulted in two separate classes for band class 8.

For each pattern, a particular band class can have two states: present or absent. This is the basis for *polymorphism analysis*, a tool which allows comparative binary (+/-) tables to be generated, displaying polymorphic bands between the selected patterns. These tables,

created as text or tab-delimited files, are ready for export to other specialized software for statistics, genetic mapping or other further analysis. The binary table for the above example (Figure 4-35) is shown in Figure 4-36.

	1	2	3	4	5	6	7	8
Pattern 1	+	+	-	+	-	+	+	+
Pattern 2	+	-	+	-	+	+	-	+
Pattern 3	+	+	-	-	+	+	+	+

Figure 4-36. Binary presence/absence table of banding patterns.


Instead of using binary (+/-) data, the same tables can be generated using band intensities obtained from the curves (band heights or surfaces) or from the two-dimensional pattern contours (volumes or concentrations).


The use of band matching tables is obvious: it provides a binary or numerical character table for fingerprint type patterns, which allows a number of statistical techniques to be applied, including Minimum Spanning Trees (Section 4.11), Maximum Parsimony trees (Section 4.9), dimensioning techniques such as Principal Components Analysis and related techniques (Section 4.12), and bootstrap analysis on dendrograms (4.1.13).

To visualize a band matching table as a character matrix (binary or quantitative), it is necessary that a composite data set is associated with the fingerprint type. Therefore, the use of composite data sets is described here in association with band matching tables. However, it is possible to apply the techniques mentioned in the previous paragraph directly on the fingerprint type without having a composite data set associated to it.

4.3.2 Creating a band matching

4.3.2.1 In the *InfoQuest FP* main window with **DemoBase** loaded, open comparison **All**, or create a comparison containing all entries except the STANDARDS (see 4.1.9).

4.3.2.2 Select **RFLP1** in the *Experiments* panel and press the  button or *Layout > Show image*.

4.3.2.3 Choose *Bandmatching > Perform band matching* or press .

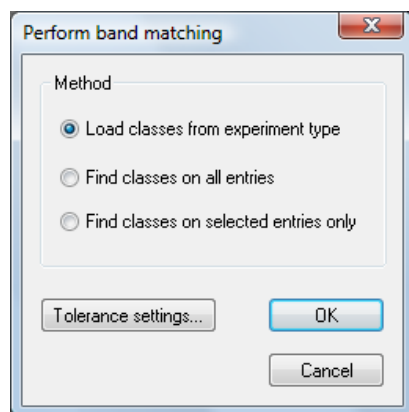


Figure 4-37. *Perform band matching dialog box of a fingerprint type.*

The *Perform band matching* dialog box pops up (see Figure 4-37), listing three different band matching options.

- **Load classes from experiment type:** the band classes stored with the experiment type are loaded (see 4.3.5). This way you can have perfect control on what bands to use in the analysis.
- **Find classes on all entries:** a band matching is performed on all entries within the comparison.
- **Find classes on selected entries only:** a band matching is performed on the currently selected entries only.

4.3.2.4 Press **<Tolerance settings>** to open the *Position tolerance settings* dialog box (see Figure 4-38).

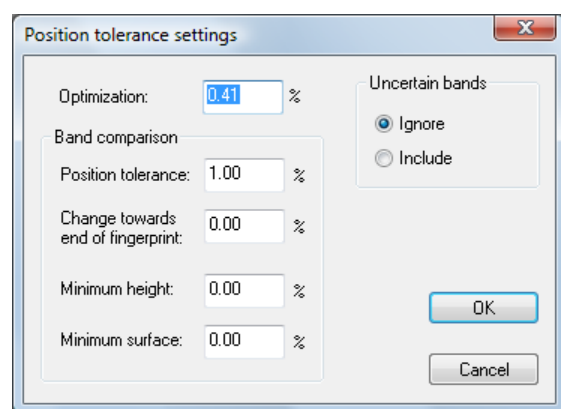


Figure 4-38. *Position tolerance settings dialog box of a fingerprint type.*

The *Position tolerance* is the maximal shift allowed (in percentage of the pattern length) between two bands allowed to consider them as matching. With *Change towards end of fingerprint*, you can specify a gradual increase or decrease in tolerance.

The *Optimization* is a shift that you allow between any two patterns and within which the program will look for

the best possible matching. To understand the utility of optimization in addition to position tolerance, see the example in Figure 4-31.

With *Minimum height* and *Minimum surface*, you can exclude weak or irrelevant bands.





The *Uncertain bands* option allows you to either include uncertain bands or ignore them. When *Ignore* is chosen, uncertain bands are ignored. This means that in composing a band matching table, the software will omit the uncertain bands, considering them as characters that are unknown. When *Include* is chosen, uncertain bands are treated in the same way as certain bands, which means that uncertain bands will contribute to the band classes of a band matching tables in the same way as certain bands.

4.3.2.5 For this example, enter a position tolerance of 1%, an optimization of 1%, a change of 0%, and a minimum height and minimum surface of 0%, and press **<OK>**.


4.3.2.6 Because no band classes are yet defined for **RFLP1** (see 4.3.5), select *Find classes on all entries* in the *Perform band matching* dialog box and press **<OK>**.

The program has now defined the band classes and has associated each band with a class. The band classes are shown as blue lines (Figure 4-39) and the bands are linked to a class in red.

NOTE: Band classes are only defined within active zones of the fingerprint type. Active zones can be set in the Fingerprint type window of the corresponding fingerprint type (see 4.2.3).

4.3.2.7 Zoom in on the image as necessary using the zoom functions  and  (*Layout > Zoom in* and *Layout > Zoom out*) or by using the zoom sliders (see 1.6.7 for instructions on how to use the zoom sliders). The latter option allows you to zoom separately in the horizontal () and vertical () direction. Horizontal zooming can also be achieved via *Layout > Stretch (X dir)* (keyboard shortcut CTRL+SHIFT+PgUp) and *Layout > Compress (X dir)* (keyboard shortcut CTRL+SHIFT+PgDn).

Zooming in the horizontal direction only can be an interesting option for long patterns with numerous small bands, such as **AFLP** in the **DemoBase**. This causes the image to be enlarged in the horizontal direction only, so that sharp bands become better visible, without losing the overview of a large number of patterns.

4.3.2.8 Press the  button or *Layout > Show metrics scale* to display the molecular weight scale of the fingerprint type.

After having performed a band matching, all band classes are labelled with a band class label. The band

class labels are listed on top of the image. If a band class is selected, its label is highlighted (see Figure 4-40).

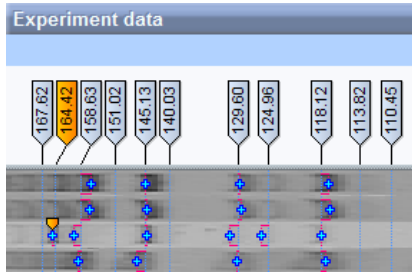


Figure 4-40. Band class labels.

If a regression curve is calculated for the reference system(s) of the selected fingerprint type, the metric positions of the band classes are displayed in the labels (e.g. 167.62; 164.42; ...).

4.3.2.9 Double-click on a band class label, or select **Band-matching > Band class information**, or press CTRL+I to open the *Band class information* dialog box.

The *Band class information* dialog box contains detailed information on the band class (see Figure 4-41):

- **Name:** If a regression curve is calculated for the reference system(s), the default name of the band class label is the % normalized position of the band class. This is the average position of all bands belonging to that band class, expressed as a percentage of the normalized track length. Each band class name is

editable and can be changed to any name of your choice. Band class names can be changed in the *Band class information* dialog box, but also from within the *Fingerprint type* window (see 4.3.5). When changing the band class names, the metric positions of the band classes remain present in the database (see *Position (metrics)* column in Figure 4-44).

- **Position:** The position reflects the relative position of the band class, derived from the regression curve.
- **Occurrence:** The occurrence corresponds to the relative occurrence of bands in the band class, expressed as a percentage of the run length.
- **Position spread box:** The position spread box lists the standard deviation of the bands to the band class and the scores of the 50th, 90th and 98th percentile. The scores are the relative positions below which respectively 50, 90 and 98% of the bands are found.

4.3.2.10 Close the *Band class information* dialog box by pressing <OK>.

4.3.3 Manual editing of a band matching

Due to shape or distribution, the program does not always assign the bands to the correct class. Therefore, you can manually correct the assignments.

For the manual band matching editing tools, a multi-level undo and redo function is available. The undo

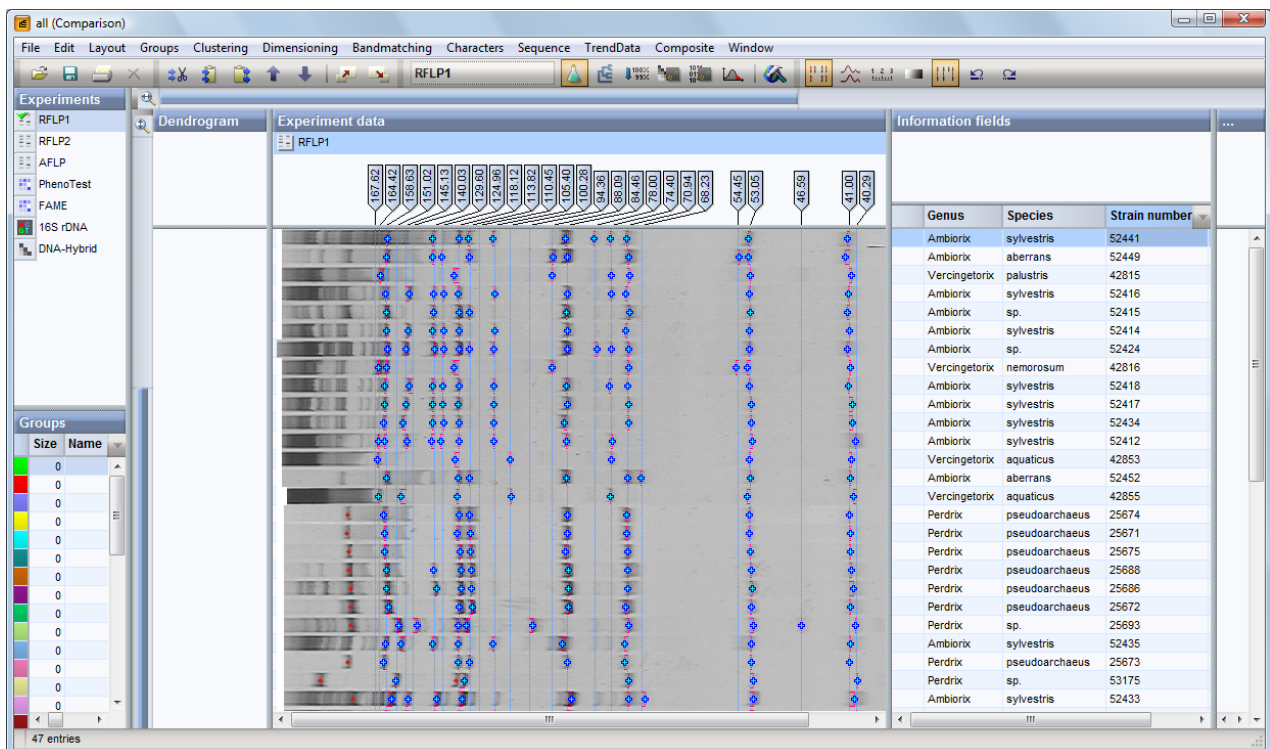


Figure 4-39. Band matching analysis in the *Comparison* window.

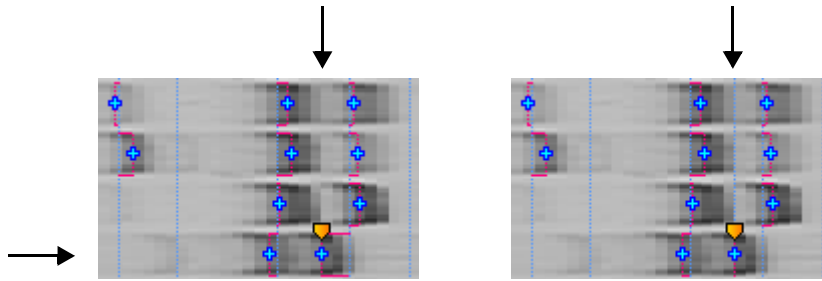


Figure 4-43. Splitting up a band class into two band classes.

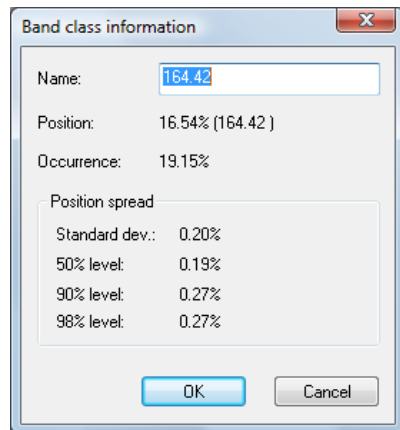




Figure 4-41. The *Band class information* dialog box.

function can be accessed with *Bandmatching > Undo* or CTRL+Z or the  button. The redo function is accessible through *Bandmatching > Redo* or CTRL+Y or the  button.

In Figure 4-42, the band marked with the arrow is assigned to the left of two close classes, whereas it should be assigned to the right class.

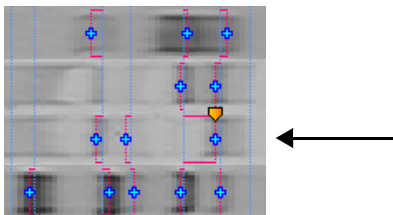


Figure 4-42. Detail of band class assignments.

NOTE: You can easily see which bands belong to a given band class by double-clicking on the vertical blue dotted line that represents the class: all bands that belong to the class are selected with a green flag.

Reassigning a band to another class can be done with a simple drag-and-drop procedure:

4.3.3.1 Select the band that was wrongly assigned. While pressing the mouse button, drag it to the band class

where it should be assigned to and release the mouse button.

If you do not wish to use a single band in a band matching analysis, you can undo its assignment as follows:

4.3.3.2 Click on the band that you want to unassign and drag it outside of the gelstrip.

A whole band class is deleted as follows:

4.3.3.3 Click on a band belonging to the band class.

4.3.3.4 Right-click on the band, and select *Band classes > Remove band class* from the floating menu. You can also press SHIFT+DEL on the keyboard to remove the selected band class.

If different bands are incorrectly assigned to the same class, you can create a second class as follows (Figure 4-43):

4.3.3.5 Select a band which should belong to a new class (see Figure 4-43).

4.3.3.6 Right-click on the band, and select *Band classes > Add new band class* from the floating menu or press SHIFT+ENTER on the keyboard.

The program asks “Do you want to auto assign bands to the new class?”. If you press <No>, the new band class will contain only the selected band. If you select <Yes>, all bands that are closer to the new band class are automatically reassigned to that new class. In order to reassign bands to the other class, follow the drag-and-drop procedure explained in 4.3.3.1.

If bands are incorrectly assigned to different classes, you can merge the classes as follows (Figure 4-43):

4.3.3.7 Choose a band which occurs quite in the middle of the two classes.

4.3.3.8 Right-click on the band, and select *Band classes > Add new band class* from the floating menu or press SHIFT+ENTER on the keyboard. Press <No> when the program asks to auto assign bands to the new class.

4.3.3.9 Choose a band which belongs to the left class.

4.3.3.10 Right-click on the band, and select **Band classes** > **Remove band class** from the floating menu, or press SHIFT+DEL.

4.3.3.11 Choose a band which belongs to the right class (left-click).

4.3.3.12 Right-click on the band, and select **Band classes** > **Remove band class** from the floating menu, or press SHIFT+DEL.


4.3.3.13 Select the new band class to which all the bands should belong (left-click). The band class label becomes highlighted.

4.3.3.14 Right-click, and select **Band classes** > **Auto assign all bands to class** from the floating menu.




After reassigning bands, removing and adding bands etc. the band class position may not be the center anymore. You can correct the position of the band class:

4.3.3.15 Select the band class (left-click) and call the floating menu (right mouse button) to select **Band classes** > **Center class position**.

NOTE: These commands are also accessible from the main menu, but they are much easier using the floating menu.


4.3.3.16 If all assignments are corrected, you can save the band matching with **File** > **Save** or .

NOTE: A band matching is saved along with the comparison. When a comparison is opened and a band matching is available for the experiment type selected,

*the  button shows up . The graphical representation of the band matching can be displayed again by **Layout** > **Show bands** or by pressing the  button.*

4.3.4 Adding entries to a band matching


Since a band matching analysis and the associated table can be saved, it should be possible to delete entries from, or add entries to the band matching at any time.


4.3.4.1 To delete some entries, simply select some entries and **Edit** > **Cut selection** or .

If entries are added however, it is possible that those new entries contain bands that are not defined as a band class yet. If you have performed some editing work to the band classes already, it would be beneficial to preserve the existing band classes, and simply associate the bands of the new entries to the existing classes, and introduce new classes in those cases where the new

entries have bands that do not fit in any of the existing classes. This is achieved as follows:

4.3.4.2 Select a few entries in the database. If you have executed the previous step (4.3.4.1) there are still some entries selected and placed on the clipboard.

4.3.4.3 [In case you would have copied something else in the meantime, select **Edit** > **Copy selection** or  in the *InfoQuest FP* main window.]

4.3.4.4 With **Edit** > **Paste selection** or  in the other comparison, the selected entries are placed back in the band matching.

4.3.4.5 Select **Bandmatching** > **Search band classes**. The *Perform band matching* dialog box as in Figure 4-37 is shown. Select **Find classes on all entries** and press <OK>.

The program now asks “Remove existing band classes?”.

4.3.4.6 In order to preserve the existing band matching, it is important to answer <No> to this question.

4.3.5 Saving band classes to a fingerprint type

After having defined band classes in the *Comparison* window, you can save the band classes to the corresponding fingerprint type.

4.3.5.1 Select **Bandmatching** > **Save band classes to experiment type**.

4.3.5.2 The program asks for confirmation, select <Yes>.

4.3.5.3 Open the *Fingerprint type* window for **RFLP1** and select the *Band classes* panel (see Figure 4-44).

All band classes defined for **RFLP1**, together with their relative and metric positions, are listed in the *Band classes* panel. The names of the band classes can be edited by clicking twice in the information fields.

If band classes are saved for a fingerprint type, the band classes can be loaded when checking **Load classes from experiment type** in the *Perform band matching* dialog box (see Figure 4-37).

The ability to edit, save and load band classes has several interesting implementations:

- Perfect control on what bands to use in the analysis.
- Every new set of fingerprint profiles can easily be compared based upon the predefined set of band classes.

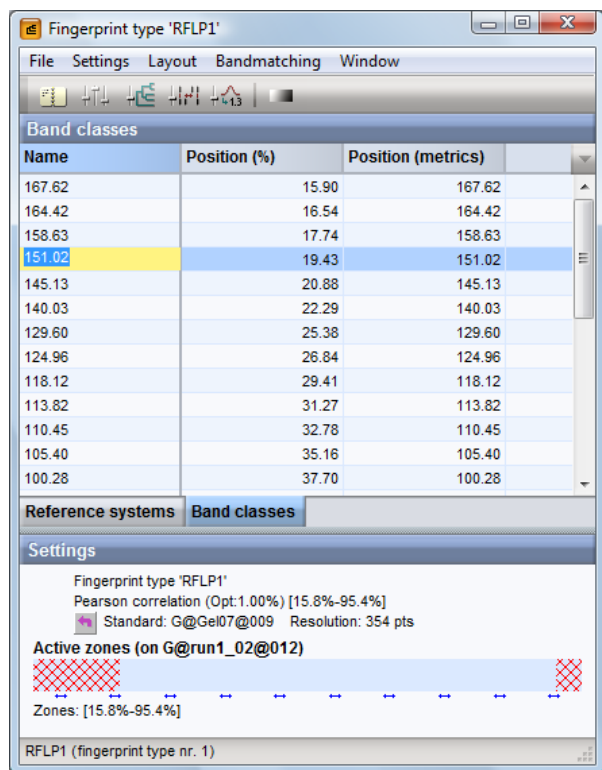


Figure 4-44. The *Band classes* panel.

- Band matching tables can be reduced to the relevant information, e.g. bands that carry genetic marker information.

4.3.5.4 To add a band class to the list select *Bandmatching > Add new band class* in the *Fingerprint type* window.

4.3.5.5 Select *Bandmatching > Remove band class* to remove a band class from the list.

4.3.6 Band and band class filters

When searching bands in complex patterns, especially those for which the terminal step is a PCR reaction such as AFLP patterns, it is sometimes difficult to define objective criteria as to what is a band and what is not a band. However, when the user examines a set of patterns by eye, it often becomes easier to decide whether a band is valid or not, because the user automatically compares the band with those on neighboring patterns, thus obtaining information which cannot be obtained by inspecting the pattern alone. This is more or less the way the band filters work in the band matching application of InfoQuest FP: in a first step, band classes are defined over all patterns; then the relative areas of all bands of a given class are averaged, and if a band deviates more than a certain percentage from this average, it is not considered as being a matching band for this class.

Using this tool, it is possible to define more bands on the gels than one would usually do, without spending a lot of time deleting and adding bands manually. Using the

band matching filters, weak bands or artifacts that do not reflect the expected intensity will be filtered out automatically, and the assignment of bands is often as reliable as after hours of band editing work.

4.3.6.1 In the band matching analysis created in 4.3.2, select *Bandmatching > Band class filter*.

This pops up the *Band filtering settings* dialog box (Figure 4-45). It exists of two parts: the upper part "*Remove all bands below...*" is to filter individual bands within a given band class, and the lower part "*Remove all band classes that have no bands exceeding...*" is to remove all band classes that do not contain any significant band.

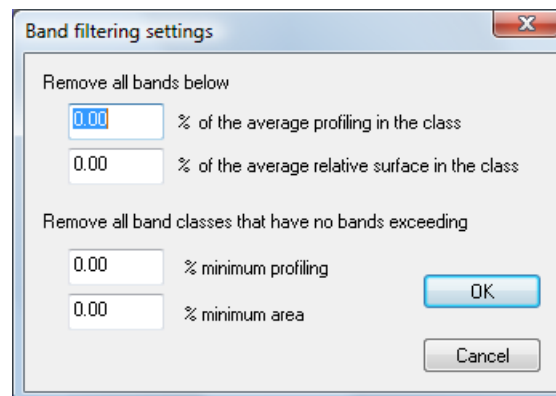


Figure 4-45. *Band filtering settings* dialog box for band matching.

Similar as for band searching, the band class filters consist of two separately working components: a *profiling* component, which is the height of the band or class, and an *area* component, which is the relative area (surface).

4.3.6.2 Within a band class, you can *Remove all bands below* a certain % of the average profiling in the class.

If you enter 80%, this means that, if the height of a band is lower than 80% of the average profiling calculated for its class, it will not be matched with that class, and the band will be recorded negative in the band matching table. Note that the profiling of a band is an absolute measure: if a pattern is rather weak, many of its bands may be excluded from the band matching just by this fact. In such cases, we recommend to take the surface as filtering factor.

4.3.6.3 Within a band class, you also can furthermore *Remove all bands below* a certain % of the relative surface in the class.

In this case, if you enter 80%, all bands that have a relative surface less than 80% of the average surface for the band class will not be matched with that class, and the bands will be recorded negative in the band matching table. Since the surface is relative to the total surface of a pattern, weak patterns in principle will not be treated differently compared to dark patterns.

In case of complex patterns such as AFLP, many band classes consist of just one weak band, spot or artifact and have no genetic or taxonomic relevance. Such band classes are just filling up the band matching table, and being treated equally important, they are disturbing the information provided by the band matching table. Therefore, InfoQuest FP offers the possibility to have all band classes excluded from the band matching table that do not contain at least one clear relevant band.

4.3.6.4 With *Remove all band classes that have no bands exceeding* a certain % *minimum profiling*, you can remove all irrelevant band classes based upon the minimum height of the bands included.

If you enter 20%, this means that a band class for which the highest band is less high than 20% of the OD range of the fingerprint type will be considered irrelevant and will be removed.

NOTE: This is again a non-relative parameter. If by incidence a band class is formed by a set of weak patterns, it may be excluded incorrectly. If this happens to be a problem, we recommend to use the more reliable feature of % minimum area only.

4.3.6.5 With *Remove all band classes that have no bands exceeding* a certain % *minimum area*, you can remove all irrelevant band classes based upon the minimum area of the bands included. The minimum area is defined as the area relative to the total area of a pattern.

If you enter 20% here, a band class that contains no band with an area bigger than 20% of its pattern's total area will be removed from the band matching table.

4.3.7 Exporting band matching information

Band matching information can be exported as a binary (presence/absence) table or as a quantitative character table.

4.3.7.1 In the band matching analysis created in 4.3.2, select *Bandmatching > Export band matching*. The program will ask "Export quantitative information?".

4.3.7.2 Press <No> to export the band matching information as a binary (presence/absence) table in tab-delimited format.

The exported band intensity values are based on the *Comparative quantification settings* for the used fingerprint type. This option can be defined in the *Fingerprint type* window, but it can also be changed in the *Comparison* window, as follows:

4.3.7.3 Select *Bandmatching > Comparative Quantification settings*. This opens the *Comparative Quantification settings* dialog box (Figure 4-46).

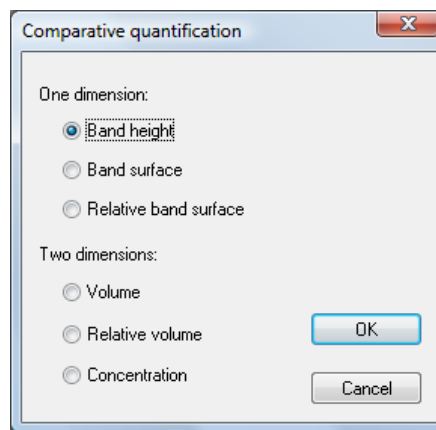


Figure 4-46. The *Comparative quantification settings* dialog box.


One dimension quantification is based on the densitometric curves extracted from the patterns: *Band height* is the height of the peak; *Band surface* is the area under the Gaussian curve approximating a band; *Relative band surface* is the same as band surface, but expressed as a percentage of the total band area of the pattern.

Two dimensions quantification is based on the band contours of the two-dimensional pattern images: *Volume* is the absolute volume within the contour; *Relative volume* is the same as a percentage of the total band volume of the pattern; *Concentration* is the physical concentration unit the user has assigned based upon regression through known calibration bands.


If no two-dimensional quantification is performed for the gels, it is obvious that one should select among the first three options.

4.3.7.4 Close the *Comparative quantification settings* dialog box with <OK> or <Cancel>.

4.3.8 Tools to display selective band classes

4.3.8.1 If present, remove the existing band matching analysis by selecting RFLP1 in the *Experiments* panel and *Bandmatching > Perform band matching* or . The program will ask for confirmation. Press <OK>.

4.3.8.2 Clear any selection made by pressing F4, and manually select some entries in the comparison.

4.3.8.3 Select *Bandmatching > Perform band matching* or press  to create a new band matching.

4.3.8.4 In the *Perform band matching* dialog box (see Figure 4-37), check the option *Find classes on selected entries only*.

With this option, the program will only create band classes for bands found on the entries in the selection.

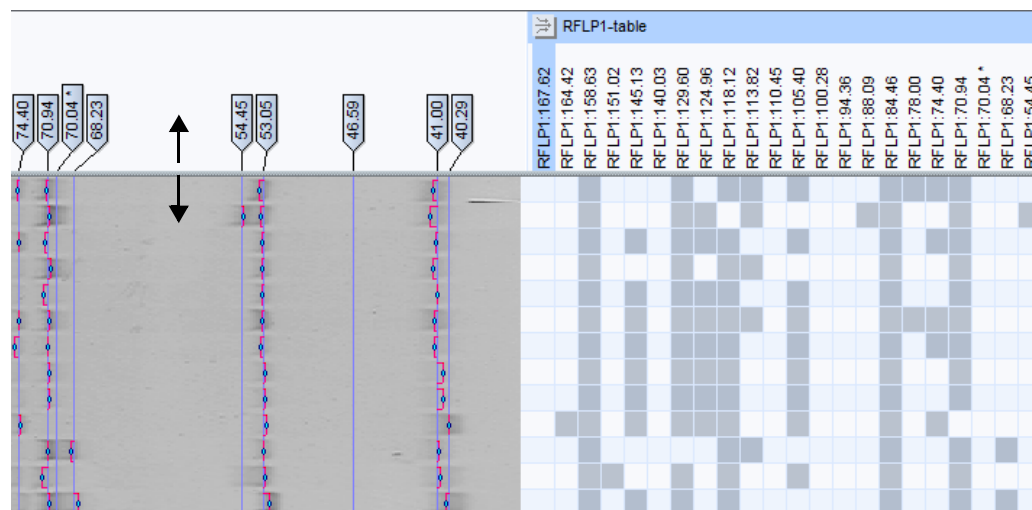




Figure 4-47. Binary band matching table; detail.

4.3.8.5 With *Bandmatching > Auto assign all bands to all classes*, you can let the program assign the bands of the non-selected entries to the corresponding band classes.

InfoQuest FP offers another interesting tool to display only the polymorphic bands. To make this tool as flexible as possible, the polymorphic bands are only looked for within the selection list. For genetic mapping purposes, the user can select the patterns from two (or more) parent entries, and have the program display only the polymorphic band classes between these two patterns. This reduces the size of the band matching table to contain only the polymorphic bands of interest. Of course, the user can add or delete band classes afterwards, as desired (see 4.3.3 for manual editing of band classes).

4.3.8.6 Clear any list of selected patterns with F4.

4.3.8.7 First, remove the existing band matching analysis by selecting **RFLP1** in the *Experiments* panel and *Bandmatching > Perform band matching* or .

4.3.8.8 Select *Bandmatching > Perform band matching* or press  to create a new band matching, including all band classes.

4.3.8.9 Select two entries having a few different bands.


4.3.8.10 Select *Bandmatching > Polymorphic bands only (for selection list)*. Only the band classes that are polymorphic between the selected two patterns are now displayed.

4.3.9 Creating a band matching table for polymorphism analysis

Before a presence/absence table as shown in Figure 4-36 can be displayed in InfoQuest FP, you will need to define a *composite data set*, containing the fingerprint type as input. A composite data set is a character table that contains all the characters of one or more experiment types. Such a character table is necessary to convert the band classes and represent them as presence/absence tables.

4.3.9.1 If not already available, define a composite data set for the two RFLP techniques (**RFLP-combined**) as described in .

When a comparison is opened after the composite data set **RFLP-combined** had been defined, **RFLP-combined** is listed in the *Experiments* panel the *Comparison* window. Since we defined **RFLP1** and **RFLP2** as being the experiment types used in this composite data set, the band matching values for **RFLP1** as calculated in the previous paragraphs (4.3.2 to 4.3.8) are automatically filled in as character values.

4.3.9.2 In the *Experiments* panel, with the band matching for **RFLP1** shown, press the  button of **RFLP1-combined**. The binary band matching table appears as in Figure 4-47.

4.3.9.3 In order to reveal the complete information on the band classes, it may be necessary to drag the separator line between the table and its header (see Figure 4-47) downwards.

NOTES:

(1) You can scroll between the image of gel patterns and the character table using the scroll bar at the bottom of the image panel. Once the character table is present, it is

RFLP1:94.36
 RFLP1:88.09
 RFLP1:84.46
 RFLP1:78.00
 RFLP1:74.40
 RFLP1:70.94
 RFLP1:68.23
 RFLP1:54.45
 RFLP1:53.05
 RFLP1:46.59
 RFLP1:41.00
 RFLP1:40.33

HEADER:
 Band classes

0.00 0.00 25.27 3.48 2.99 8.95 0.00 0.00 8.31 0.00 13.31 0.00
 0.00 10.39 0.00 0.00 8.85 10.46 0.00 0.00 11.78 0.00 10.23 0.00
 0.00 13.49 25.32 0.00 0.00 15.51 0.00 9.28 7.11 0.00 5.05 0.00
 0.00 0.00 19.79 0.00 0.00 13.71 0.00 0.00 5.13 0.00 5.55 0.00
 0.00 0.00 21.67 2.36 2.42 6.39 0.00 0.00 6.29 0.00 6.61 0.00
 0.00 0.00 25.06 0.00 0.00 3.92 0.00 0.00 6.33 0.00 5.27 0.00
 0.00 11.89 0.00 0.00 0.00 10.34 0.00 5.47 7.44 0.00 3.69 0.00
 0.00 0.00 26.11 0.00 0.00 4.97 0.00 0.00 5.85 0.00 3.65 0.00
 0.00 0.00 26.20 0.00 3.76 2.59 0.00 0.00 5.95 0.00 4.56 0.00

TABLE:
 Rows=entries

Figure 4-50. Numerical band matching character table exported from InfoQuest FP (space-delineated).

still possible to edit the band class assignments on the patterns. The character table is updated automatically.

(2) Band classes that have been created by the user are marked with an asterisk (*).

4.3.9.4 Use *Composite > Export character table* to export a space or tab-delineated text file of the binary band matching table.

When the program asks “Use tab-delineated fields”, you should answer <Yes> to produce a tab-delineated text file. The tab-delineated table looks as shown in Figure 4-48 and is in fact very similar to the one obtained via the command *Bandmatching > Export band matching*.

	RFLP1:167.62	RFLP1:164.42	RFLP1:158.63	RFLP1:151.02
G@Gel07@004	0	1	0	0
G@Gel11@005	0	1	1	0
G@Gel07@017	0	1	0	0
G@Gel11@006	0	1	1	0
G@Gel11@011	1	1	0	0
G@Gel08@016	0	1	1	0
G@Gel07@015	0	1	0	0
G@Gel07@010	0	1	1	0
G@Gel08@003	0	0	1	0
G@Gel08@006	0	0	1	0
G@Gel08@015	0	0	1	0

Figure 4-48. Binary band matching character table exported from InfoQuest FP (tab-delineated).

In the tab-delineated format, the band classes (header) and the band presence/absence table are given in columns separated by tabs. This format is the easiest to import in spreadsheet or database software packages.

4.3.9.5 To show the intensity of the bands, choose *Composite > Show quantification (colors)*.

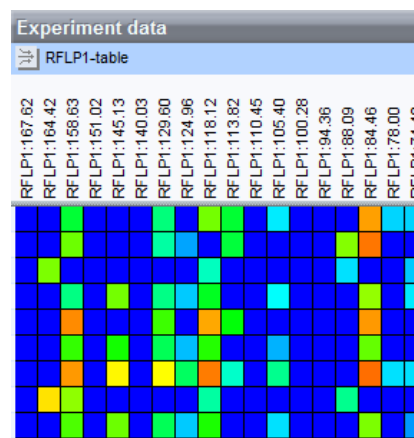


Figure 4-49. Intensity of bands shown in color.

The color ranges from blue (weakest bands) over cyan, green, yellow, orange to red (darkest bands) (see Figure 4-49). The intensity is based upon the *Comparative quantification settings* for the used fingerprint type. This option can be defined in the *Fingerprint type* window or in the *Comparison* window (see 4.3.7.3).


4.3.9.6 Make sure that **RFLP1-combined** is selected in the *Comparison* window. Select *Composite > Show quantification (values)* to display the numerical intensities of the bands.

4.3.9.7 With *Composite > Export character table*, a numerical band matching table is created in text format, separated by tabs or spaces (Figure 4-50).

4.3.10 Finding discriminative bands between entries

The use of a composite data set allows discriminative bands to be searched for in a band matching table.

4.3.10.1 In database **DemoBase**, have a *Comparison* window open with all non-“STANDARD” entries selected (e.g. comparison **All**, see 4.1.9) and the composite data set **RFLP1-combined** shown (see 4.3.2 and 4.3.6.3).

4.3.10.2 Make sure that the image of the composite data set is shown, by pressing the  button of **RFLP1-combined** in the *Experiments* panel.

4.3.10.3 Minimize or reduce the *Comparison* window so that the *InfoQuest FP main window* (at least the menu and toolbar) becomes visible.

4.3.10.4 Press F4 to make sure that no entries are selected.

4.3.10.5 In the *InfoQuest FP main window*, select **Edit > Search entries** (F3), enter *Vercingetorix* in the **Genus** field and press <Search>.

All *Vercingetorix* entries are selected in the *Database entries* panel of the *InfoQuest FP main window* and in the *Information fields* panel of the *Comparison* window.

4.3.10.6 To group the selected entries, choose **Edit > Bring selected entries to top** in the *Comparison* window or press CTRL+T on the keyboard.

4.3.10.7 Select **Composite > Discriminative characters**.

The characters (bands) are reorganized in such a way that those characters positive for the selected entries and negative for the other entries occur left, and those characters negative for the selected entries and positive for the other entries occur right (see Figure 4-51).

In a composite data set, it is possible to list the entries according to the value of a selected character. In case of banding patterns, the entries will be ordered by the intensity of a selected band. This feature allows for a particular band the entries to be found in which the band is present or not.

4.3.10.8 Show the band table as intensity table with **Composite > Show quantification (colors)**.

4.3.10.9 Click on a band class in the band classes header (Figure 4-51) and **Composite > Sort by character**.

The entries are now sorted by increasing intensity of the selected band class.

Furthermore, it is possible to perform a transversal (or two-way) clustering of a band matching table. See 4.8.4 for a detailed description of the transversal clustering of composite data sets.

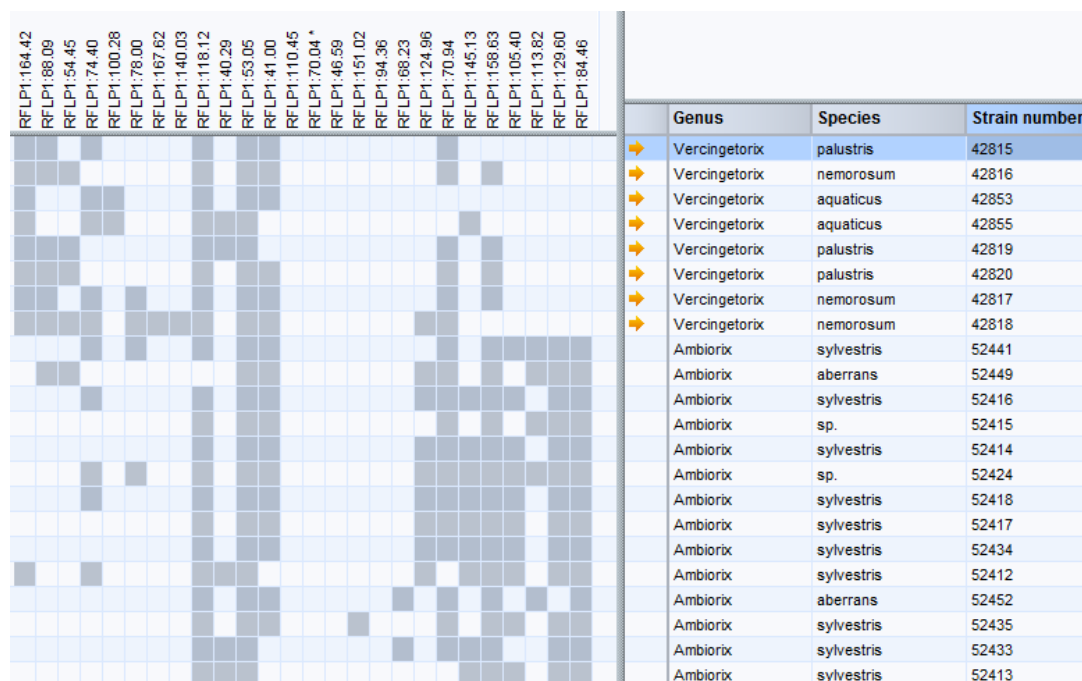


Figure 4-51. Discriminative bands for selected entries, positive discrimination left, negative discrimination right.

4.4 Cluster analysis of characters CL CH

4.4.1 Character comparison settings

In terms of parameter settings, character sets are the simplest class of data to analyze. The various types of character sets that exist, however, require a large number of coefficients to be available for analyzing character tables.

4.4.1.1 In the *InfoQuest FP* main window with **DemoBase** loaded, open comparison **All**, or create a comparison containing all entries except the STANDARDS (see 4.1.9).

4.4.1.2 In the *Experiments* panel of the *Comparison* window, select a character type to analyze, for example **PhenoTest**.

4.4.1.3 Select **Clustering > Calculate > Cluster analysis (similarity matrix)**. The *Comparison settings* dialog box appears (Figure 4-52).

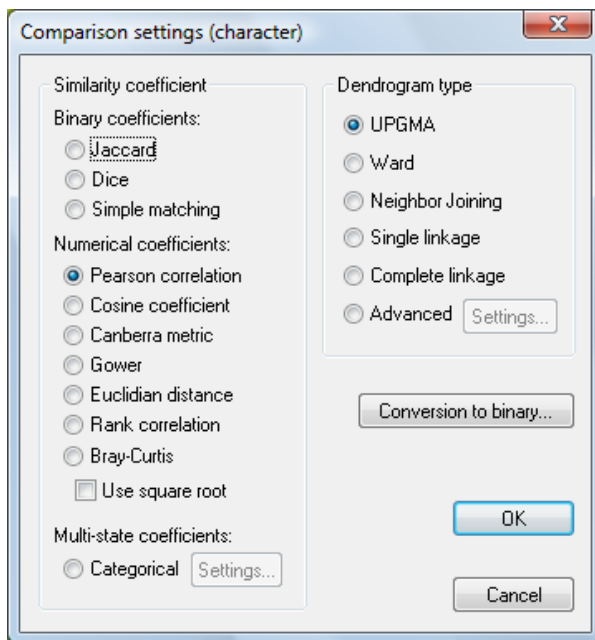


Figure 4-52. *Comparison settings* dialog box for character data.

The left panel lists the available similarity coefficients. *Binary coefficients* include *Jaccard*, *Dice*, and *Simple matching*.

1. The *Jaccard* coefficient

$$S_J = \frac{N_{AB}}{N_A + N_B - N_{AB}}$$

2. The *Dice* coefficient

$$S_D = \frac{2N_{AB}}{N_A + N_B}$$

3. The *Simple matching* coefficient

$$S_{SM} = \frac{N_{AB} + N_{ab}}{N_{AB} + N_{ab} + N_{Ab} + N_{aB}}$$

In the above notations, uppercase A or B means that a character is positive for entry A or B, whereas lowercase a or b means that the character is negative for the entry.

Dice and Jaccard are very related to each other whereas Simple matching is more fundamentally different. The Jaccard and Dice coefficients only consider "scoring characters" being two positive characters in both data sets, whereas the Simple matching coefficient also considers two negative characters as scoring.

When dealing with a non-binary (numerical) data set, a conversion needs to be done from numerical values to binary values (positive or negative), before one of these coefficients can be applied.

4.4.1.4 The **<Conversion to binary>** button lets you specify how this conversion is done (Figure 4-53).

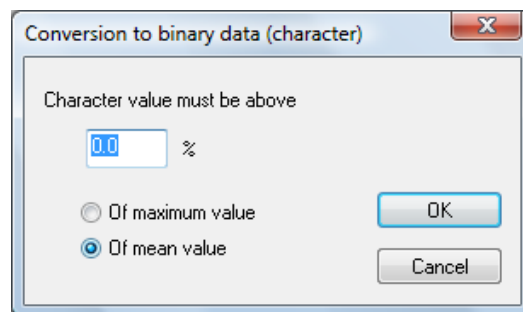


Figure 4-53. *Conversion to binary* dialog box.

By default, every character that has a value above zero will be converted to positive. Alternatively, one can specify a certain percentage of either the *maximum value* or the *mean value* from the experiment.

Numerical coefficients include *Pearson correlation* (or Pearson product-moment correlation) and the related *Cosine coefficient*, the *Canberra metric* and *Gower coefficients*, *Euclidean distance*, *Rank correlation*, and *Bray-Curtis*.

1. The *Pearson correlation* coefficient

$$r_{j,k} = \frac{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)(x_{i,k} - \bar{x}_k)}{\sqrt{\sum_{i=1}^n (x_{i,j} - \bar{x}_j)^2 \sum_{i=1}^n (x_{i,k} - \bar{x}_k)^2}}$$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{i,j}$$

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n x_{i,k}$$

2. The *Cosine* coefficient

$$C_{j,k} = \frac{\sum_{i=1}^n x_{i,j} x_{i,k}}{\sqrt{\sum_{i=1}^n x_{i,j}^2 \sum_{i=1}^n x_{i,k}^2}}$$

3. The *Canberra Metric* coefficient

$$D_{CANB,j,k} = \frac{1}{n} \sum_{i=1}^n \frac{|x_j - x_k|}{|x_j + x_k|}$$

4. The *Gower* coefficient

$$D_{G,j,k} = \frac{\sum_{i=1}^n w_{i,j,k} S_{i,j,k}}{\sum_{i=1}^n w_{i,j,k}}$$

5. The *Euclidean Distance* coefficient

$$\Delta_{j,k} = \frac{1}{n} \sqrt{\sum_{i=1}^n (x_{i,j} - x_{i,k})^2}$$

6. The *Bray-Curtis* coefficient

$$D_{BC,j,k} = \frac{\sum_{i=1}^n |x_{i,j} - x_{i,k}|}{\sum_{i=1}^n |x_{i,j} + x_{i,k}|}$$

7. The *Rank correlation* (or *Spearman rank-order correlation*) is a special purpose coefficient, which, for each entry, converts the character arrays into ranks based on intensities, and uses these rank arrays to calculate correlations. The coefficient is very robust, but not sensitive to reveal details. The *Bray-Curtis* coefficient (also called *Sorensen* coefficient) is a widely used general purpose coefficient for quantitative comparisons.

For comparisons between highly related organisms, it can be useful to check the option *Use square root*, especially when using Pearson correlation or Euclidean distance. This has the effect that narrow branches on a dendrogram are stretched out relatively more than distant links.

The *Categorical* coefficient is neither binary nor numerical, since it treats each different value as a different *state*. This coefficient is useful for analyzing *multistate* character sets, for example colors (red, green, blue etc.) represent each a categorical state. Typical multistate characters used in typing, taxonomy and phylogeny are phage typing, Multilocus Sequence Typing (MLST), Variable Number Tandem Repeats (VNTR) typing. The types or categories assigned to the different phage reactions, allele numbers, or repeat numbers, respectively, are good examples of categorical or multistate data which can be analyzed using the categorical coefficient.

As an extra option for the categorical coefficient, one can specify a certain *tolerance* for values to be considered as belonging to the same category. This makes it possible to treat non-discrete (non-integer) values as categorical. Certain data types, such as VNTR fragment lengths, are in origin not categorical, but have to be converted into categories before the analysis. Using the tolerance setting, one can analyze VNTR data directly based upon fragment lengths.

NOTE: InfoQuest FP offers a number of dedicated plugins, to work with e.g. MLST and VNTR data.

The *Categorical similarity settings* can be found in a dialog box which is called by pressing **<Settings>** under *Categorical* (Figure 4-54). By default, the *Tolerance* value is set to zero, which means that no tolerance is allowed, i.e. the values must be identical to be considered the same category. With *Fuzzy logic* checked, the coefficient will score each character match decreasingly with increasing distance between the values, between full match (zero distance) and no match (distance = tolerance).

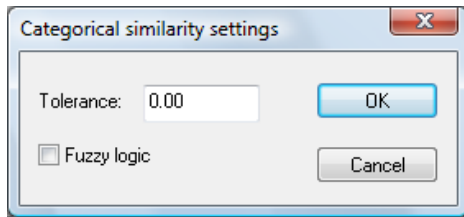



Figure 4-54. *Categorical similarity settings* dialog box.

4.4.2 Character display functions

4.4.2.1 In the comparison **All**, press the  button of **PhenoTest**.

The pattern images are displayed in the *Experiment data* panel. Initially, the character values are displayed as colors according to the color scale defined for each character.

4.4.2.2 Select **Characters > Show values** to show the corresponding character values for all entries in the *Experiment data* panel.

4.4.2.3 If a mapping was defined for the character type, the mapped names for each character value can be displayed in the *Experiment data* panel with **Characters > Show mapping**. If no mapping is present or if the character values do not fall within the ranges of the predefined criteria, a '<?>' is displayed.

4.4.2.4 To export the character values select **Characters > Export character table**.

It is possible to list the entries according to the character values of a selected character.

4.4.2.5 Select a character in the header of the *Experiment data* panel (e.g. c2) and select **Characters > Sort by character value**. The entries are now sorted by increasing value of the selected character.

4.4.3 Advanced analysis of massive character sets using GeneMaths XT

The analysis of huge data sets such as produced by gene chips or high-density gene arrays (micro-arrays) requires special clustering algorithms capable of processing many thousands of entries or characters. In addition, the successful exploration of such data sets also depends on the ability to associate certain clusters of characters (observations) with groups of entries (samples). Although these features are available in the Comparison functions of InfoQuest FP, the flexibility of handling and clustering extremely large matrices as well as some sophisticated statistical functions are provided in a separate program, GeneMaths XT (or its predecessor, GeneMaths). The GeneMaths XT program is capable of clustering data sets of up to a million characters per entry.

GeneMaths XT is available as a standalone program, but can also be added as a module to the InfoQuest FP software. In the latter case, InfoQuest FP provides the database tools, and a data matrix for comparison is first created in InfoQuest FP. The menu command **File > Analyze with GeneMaths** in the InfoQuest FP *Comparison* window then automatically launches GeneMaths XT (or GeneMaths; whichever is last installed) with the current selection of entries and experiments.

4.4.3.1 To run GeneMaths XT as a module of InfoQuest FP, create a comparison in InfoQuest FP with an appropriate large character set. If you do not have gene array data available, create a comparison in InfoQuest FP' **DemoBase** containing all entries that have the experiment **FAME** available, and click on **FAME** in the *Experiments* panel of the *Comparison* window. This character set contains some 60 characters.

4.4.3.2 In the InfoQuest FP *Comparison* window, select **File > Analyze with GeneMaths**.

This will launch the GeneMaths XT program with its main window. Full descriptions of the GeneMaths XT software is available in a separate manual.

When a connected database is defined, characters can be described by more than one information field. GeneMaths XT is launched with all the character field information available, and is able to display these multiple character fields together, whenever the characters are chosen as rows.

Character as well as entry information fields can be edited directly from GeneMaths XT and changes are saved in the InfoQuest FP database. Selecting entries in InfoQuest FP and GeneMaths XT is also synchronized.

4.5 Multiple alignment and cluster analysis of sequences CL SQ

4.5.1 An introduction to sequence analysis

Among all types of experimental data, cluster analysis of sequence data is by far the most complex in steps and possibilities. The fact that sequences need to be *aligned* before one can estimate similarity requires a number of additional steps before a dendrogram can be obtained. Furthermore, sequence data are a suitable substrate for a number of phylogenetic clustering algorithms which can rarely be applied to other types of data.

There are two ways to obtain a dendrogram from sequence data: by aligning the sequences *pairwise* (steps 1-2 in Figure 4-55), or by obtaining a *multiple alignment* of all sequences (steps 1-6 in Figure 4-55).

The best multiple alignments that can be achieved, particularly for large numbers of sequences, involve the following steps.

1. Pairwise alignment and calculation of similarity of all possible pairs of sequences, resulting in the *Pairwise alignment similarity matrix*.

2. Construction of a *UPGMA dendrogram* based on the similarity matrix obtained.
3. Determination of *consensus sequences* at each linkage node of the dendrogram, down to the root.
4. Alignment of all sequences based on the local and the root consensus sequences.
5. Calculation of a similarity matrix based on the aligned sequences, the *Multiple alignment similarity matrix*.
6. Construction of a Neighbor Joining dendrogram based on the multiple alignment similarity matrix.

In step 1, each individual sequence is aligned with each other sequence, and for each pair of aligned sequences, the similarity value is calculated and registered in a similarity matrix. The obtained matrix (*pairwise alignment similarity matrix*) will serve as the basis for cluster analysis by the Unweighted Pair Group Method using Arithmetic averages (UPGMA) (step 2). Neighbor Joining or other algorithms resulting in unrooted dendrograms would not be suitable here, as in such dendrograms, the closest

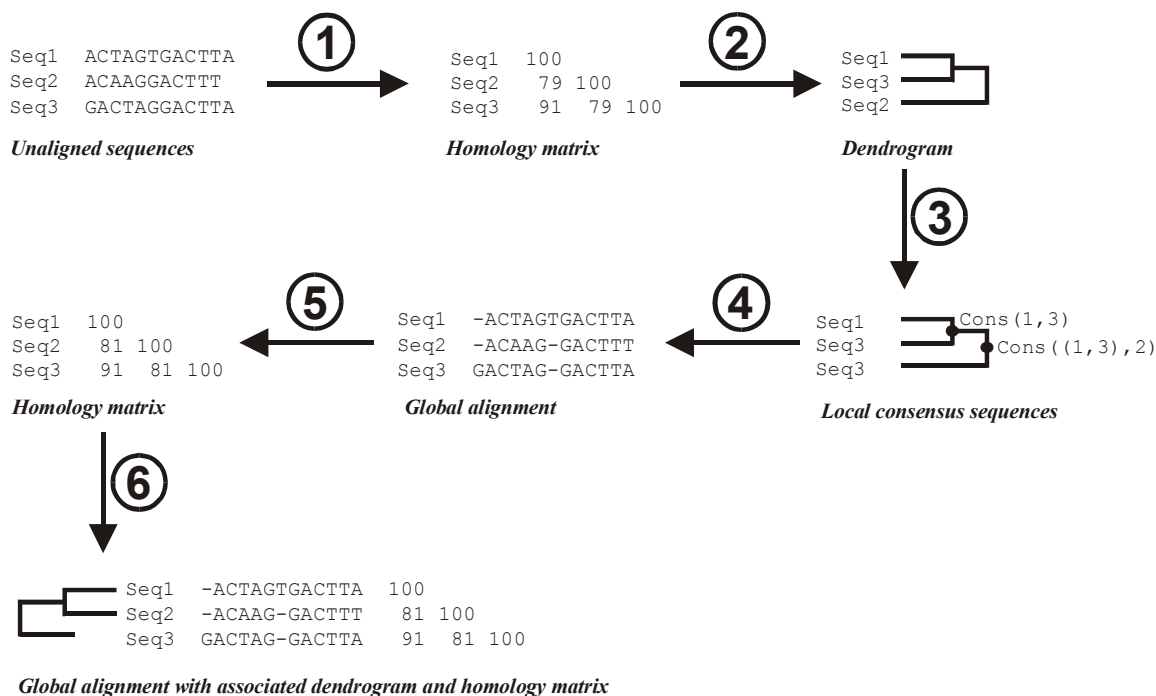


Figure 4-55. Steps in a cluster analysis of sequences: dendrogram based on pairwise alignment (steps 1 to 2), and dendrogram based on multiple alignment (steps 1 to 6).

linked sequences are not necessarily the most related ones. This is a requirement for step 3, discussed below.


Steps 3 and 4 are very important for obtaining a meaningful global alignment. Each linkage node on the UPGMA dendrogram represents a *local alignment* of the sequences linked at the node, resulting in a *local consensus*. These local consensus sequences are calculated downwards, i.e. starting from the highest related sequences down to the dendrogram root (step 3). In the above example, the highest linkage observed is between sequences 1 and 3, leading to consensus (1,3). The next linkage level is the branch that links sequences 1 and 3 with sequence 2. At this node, the consensus (1,3) is aligned with sequence 2. This results in a consensus ((1,3),2), which will, in turn, be aligned with the consensus of another group linked to this one. For each sequence or local consensus, the program keeps track of the positions of the gaps that are introduced to align it with the branch it is linked to. Finally, a *global consensus* for the whole dendrogram is inferred.

The program now introduces to each individual sequence all the gaps that were introduced on the subsequent consensus sequences following the path from the sequence itself down to the global consensus (step 4). This results in a *global* or *multiple alignment*.

The multiple alignment in turn can be used as the basis for the calculation of a similarity matrix. Now, instead of aligning each sequence with each other sequence to determine their mutual similarity, the multiple alignment is used to calculate the *multiple alignment-based similarity* between each pair of sequences (step 5). Once the multiple alignment is present, this step is extremely fast. The *multiple alignment-based similarity matrix* can be used for Neighbor Joining or UPGMA clustering, or other clustering algorithms (step 6).

4.5.2 Calculating a cluster analysis based on pairwise alignment

4.5.2.1 In the *InfoQuest FP* main window with **DemoBase** loaded, open comparison **All**, or create a comparison containing all entries except the STANDARDS (see 4.1.9).

4.5.2.2 Select **16S rDNA** in the *Experiments* panel and press the  button (or select *Layout > Show image* from the menu).

Initially, the sequences are not aligned and no similarity matrix exists.

NOTE: It is possible that a dendrogram (and a matrix) are still displayed in the Comparison window. This is the dendrogram of the last clustered experiment, which

*you can remove with **Layout > Show dendrogram** and **Layout > Show matrix**.*

4.5.2.3 The similarity matrix is calculated with *Clustering > Calculate > Cluster analysis (similarity matrix)*

or the  button.

The *Pairwise comparison settings* dialog box appears (Figure 4-56), showing three groups of settings: the *Pairwise alignment settings*, the settings for *Similarity calculation*, and the *Clustering* method.

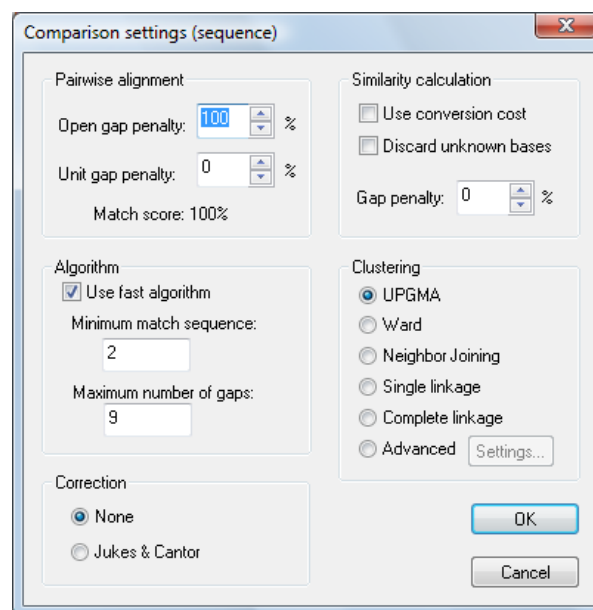


Figure 4-56. The *Pairwise comparison settings* dialog box.

The *pairwise alignment settings* involve an *Open gap penalty* and a *Unit gap penalty*. A mismatch between bases on two sequences, e.g. A with G, is considered as 100% score. The open gap penalty is the percentage cost of that score if one single gap is introduced in one of both sequences. The unit gap penalty is the percentage cost of that score to increase the gap by one base position. The default setting is 100% open gap penalty and 0% unit gap penalty, which means that introducing a gap in one of both sequences has the same cost as a mismatch, whereas there is no extra cost for gaps of multiple positions. It should be emphasized that the pairwise alignment settings will only determine the way the alignment is done: if a large unit gap cost is set (e.g. 350%), the program would not easily introduce gaps between sequences; for example, the program would rather allow three successive mismatches than one single gap. If no gap cost is chosen (0%) the program would introduce gaps to match every single base. The pairwise alignment settings have no direct influence on the similarity values, but of course, if the obtained alignments differ, the similarity values may differ too.

Use fast algorithm offers an interesting accelerated algorithm, with two adjustable parameters: the *Minimum match sequence* and the *Maximum number of gaps*. The program creates a *lookup table* of groups of bases for both sequences. The *minimum match sequence* is the size of such a group. The smaller the groups are, the more precise the alignment will be, but the longer the alignment will take. The parameter can be varied between 1 and 5, with 2 as default. The *maximum number of gaps* is the maximum number of possible gaps that you allow the algorithm to introduce in one of both sequences. The values can be varied between 0 and 99 with 9 as default. The larger the number, the more gaps the program can create to align every two sequences, but the longer the alignment will take. If zero is selected, no gaps at all would be introduced. Thus, you can custom-define its accuracy between very fast and fairly rough, to slow and very accurate.

Contrary to the pairwise alignment settings, the *Similarity calculation* parameters will not influence the alignments, but determine the way the similarity is calculated. The *Gap penalty* is a parameter which allows to specify the cost the program uses when one single gap is introduced. This cost is relative to the score the program uses for a base mismatch, which is equal to 100%. The program uses 0% as default. When *Discard unknown bases* is disabled, the program will use a predefined cost table for scoring uncertain or unknown bases. For example, N with A will have 75% penalty, as there is only 25% chance that N is A. Y and C will be counted 50% penalty because Y can be C or T with 50% chance each. If this setting is disabled, all uncertain and unknown bases will not be considered in calculating the final similarity. *Use conversion cost* is a parameter which makes calculation of the pairwise similarity matrix faster. Both described alignment methods work in two steps: first they determine the total maximal conversion score to convert one sequence into the other (given the current alignment settings) and then they realize the alignment using the minimal gap cost and maximal matching score. If *Use conversion cost* is enabled, the calculated conversion cost is transformed into a similarity value. This method is two times faster than the usual similarity calculation, but the obtained values cannot be described as real "similarity".

Under *Correction*, one can select the one parameter correction for evolutionary distance, as calculated from the number of nucleotide substitutions as described by *Jukes and Cantor* (1969)¹. The resulting dendrogram displays a distance scale which is proportional to an evolutionary time, rather than a similarity scale.


As *Clustering* method, one can choose between *UPGMA*, *Ward*, *Neighbor Joining*, *Single linkage* and *Complete linkage*.

4.5.2.4 Select an *Open gap penalty* of 100, a *Unit gap penalty* of 0, *Minimum match sequence* of 2, *Maximum number of gaps* of 9, enable *Discard unknown bases*, with a *Gap penalty* of 0 for similarity calculation, *None* for correction, and select *UPGMA* as clustering method.

4.5.2.5 Press <OK> to calculate the matrix and the dendrogram.

When the calculations are finished, the dendrogram and the matrix are shown. The sequences are still unaligned since no multiple alignment is calculated yet.

4.5.3 Calculating a multiple alignment

4.5.3.1 Select *Sequence > Multiple alignment* or .

The *Global alignment settings* dialog box (Figure 4-57) appears.

When a multiple alignment is calculated, individual sequences and local consensus sequences are aligned pairwise, down to the root, to obtain a global consensus (see steps 3 and 4 on Figure 4-55). It is this pairwise alignment of local consensus sequences that uses the same two parameters as explained before: the *Open gap penalty* and the *Unit gap penalty*.

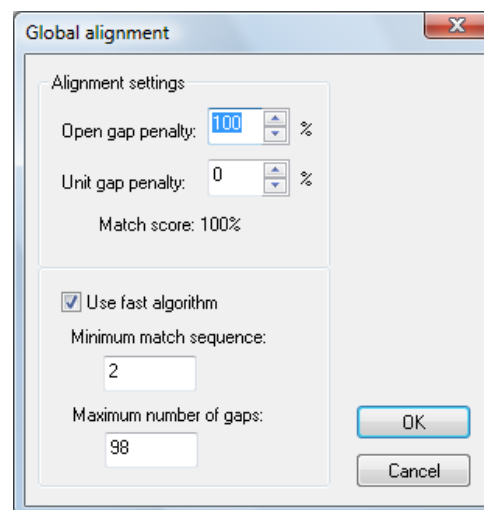


Figure 4-57. *Global alignment settings* dialog box.

The significance of the open and unit gap penalties is the same as explained for pairwise alignment: they are the percentage of the mismatch cost to create a gap, and to increase the gap by one base position, respectively. The default setting is 100% *Open gap penalty* and 0% *Unit gap penalty*, which means that introducing a gap in one of both sequences has the same cost as a mismatch, whereas there is no extra cost for gaps of multiple positions. These pairwise alignment settings will only determine the way the alignment of the local consensus sequences is done: if a large unit gap cost is set (e.g. more than 100%), the program would not easily introduce gaps between sequences. If no gap cost is chosen

1. Jukes, T.H. and C.R. Cantor. 1969. In "Mammalian Protein Metabolism III" (H.N. Munro, ed.), p. 21. Academic Press, New York.

(0%) the program would introduce gaps in order to match single bases. The pairwise alignment settings have no direct influence on the similarity values obtained from a global alignment, but if the eventual multiple alignment differs, the derived similarity values may differ too.

Use *fast algorithm* is an algorithm with two adjustable parameters: the *Minimum match sequence* and the *Maximum number of gaps* (see also under pairwise alignment). The *Minimum match sequence* can be varied between 1 and 5, with 2 as default. The *Maximum number of gaps* can be varied between 0 and 198 with 98 as default. The smaller the first number and the larger the second number, the more accurate the multiple alignment should be. If the default values are not satisfactory, e.g. for very diverse sequences, some experimenting is recommended.

Note that the *Global alignment settings* dialog box does not contain settings for similarity calculation, unlike the *Pairwise alignment settings* dialog box. The similarity matrix based upon the global alignment is not calculated automatically by the program, but requires a further command by the user (see 4.5.11; illustrated in step 5 of Figure 4-55).

4.5.3.2 Press **<OK>** to start the multiple alignment. When the calculations are done, all sequences are aligned in the *Experiment data* panel.

4.5.4 Multiple alignment display options

With *Sequence > Display settings*, the general display options such as colors and symbols, can be changed. These settings are specific to the sequence type and can therefore also be accessed from the *Sequence type* window (see 4.5.14).

In order to facilitate visual interpretation of multiple alignments there are three methods to highlight homologous regions.

Select *Sequence > Neighbor blocks* to show the *Neighbor match* representation.

This representation shows bases as blocks (highlighted) if at least one of the neighboring sequences has the same base at the corresponding position. Between two different groups of consensus, a small white line is drawn (Figure 4-58).

```

CCGCGTATTTACCGGATGG
CCGCGTA-TTGCCGGATGG
CCGCGTA-TTGCCAGATGG
CCGCGTATTTGCCAGATGG
CCGCGTATTTGTCAGATGG
CCGCGTATTTGTCAGATGG
CCGCGTATTTATCAGATGG
CCGTGTATTTATCAGATGG
  
```

Figure 4-58. Neighbor match representation.

The *Consensus match* first requires a consensus sequence to be present. A consensus sequence is defined from one or more sequences, and in case a user-defined percentage of the sequences have the same base at a given position, this base will be written in the consensus. Usually, one will select the root to calculate the consensus from. This method highlights bases (shown as blocks) on the aligned sequences if they are the same as on the consensus sequence.

4.5.4.1 Select the root and *Sequence > Create consensus of branch*. A dialog box prompts *Enter minimum consensus percentage*. You can for example enter 50, which means that a base at a given position will only be shown in the consensus sequence at least 50% of the sequences have that base at the given position. A consensus sequence of the root is now shown on the header of the *Experiment data* panel.

NOTE: A consensus sequence cannot be obtained from an advanced tree (see Section 4.10).

4.5.4.2 Select *Sequence > Consensus blocks* to show the consensus match representation (Figure 4-59).

```

CCGCGTATTTACCGGATGG
CCGCGTA-TTGCCGGATGG
CCGCGTA-TTGCCAGATGG
CCGCGTATTTGCCAGATGG
CCGCGTATTTGTCAGATGG
CCGCGTATTTGTCAGATGG
CCGCGTATTTATCAGATGG
CCGTGTATTTATCAGATGG
  
```

Figure 4-59. Consensus match representation.

The *Consensus difference* also displays the consensus sequence in the editor caption, and only shows bases that differ from the consensus while bases that are the same as the consensus are shown as |.

4.5.4.3 Select *Sequence > Consensus difference*. The consensus difference representation is as in Figure 4-60.

```

CCGCGTATTTACCGGATGG
| | | | | - | G | | |
| | | | | - | G | A |
| | | | | | | G | A |
| | | | | | | G T | A |
| | | | | | | G T | A |
| | | | | | | T | A |
| | | | | | | T | A |
| | | | | | | T | A |
  
```


Figure 4-60. Consensus difference representation.

4.5.4.4 A consensus sequence can be copied to the clipboard with *Sequence > Copy consensus to clipboard*.

Bases for which there is a consensus in at least 50% of the sequences are named, the other bases are unnamed (N).



4.5.5 Editing a multiple alignment

A multiple alignment can be edited manually and is saved along with the comparison.

4.5.5.1 Select *File > Save* or  to save the multiple alignment.



In order to rearrange the multiple alignment as desired, any sequence can be moved up or down:

4.5.5.2 Left-click on the entry you want to move up or down.

4.5.5.3 Press the  button to move the entry up, or the  button to move it down.

4.5.5.4 To move a sequence to the top or the bottom of the alignment, hold the SHIFT key and press the up or down button, respectively.

Note that, as soon as an entry is moved up or down, the dendrogram disappears: a dendrogram imposes a certain order to the entries, which is not compatible with freely moving sequences up or down. You can display the dendrogram again using *Layout > Show dendrogram*, however, this will reorder the entries again so that all manual changes you made to the sequence order are lost.

4.5.5.5 A number of manual alignment editing tools are described below. For these editing tools, the multiple alignment editor contains a multilevel undo and redo function. The undo function can be accessed with *Sequence > Edit alignment > Undo* or the  button (shortcut CTRL+Z on the keyboard). The redo function is accessible through *Sequence > Edit alignment > Redo* or the  button (shortcut CTRL+Y on the keyboard).

The undo/redo function works for the following sequence editing functions: drag-and-drop realignments (4.5.6), inserting and deleting gaps (4.5.7), removing common gaps (4.5.8), and changing sequence bases (4.5.9). The undo/redo function also works for all automatic alignment functions, including full multiple alignment (4.5.3) and partial alignments obtained with one of the following commands *Align internal branch*, *Align external branch*, and *Align selected sequences* (4.5.12).

4.5.6 Drag-and-drop manual alignment

4.5.6.1 A cursor, visible as a black rectangle can be placed on any base of any sequence, and can be moved up, down, left, and right using the arrow keys.

4.5.6.2 The cursor can also be extended to cover a range of bases both in the vertical and the horizontal direction.

This can be achieved by holding down the SHIFT key while pressing the arrow keys. The result is that blocks of bases can be selected as shown in Figure 4-61. By

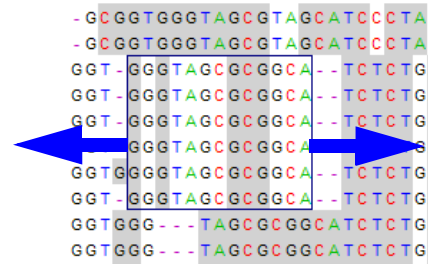



Figure 4-61. Selecting blocks of bases for drag-and-drop manual alignment.

dragging the mouse towards the left or the right (see Figure 4-61), the block of bases can be realigned within the alignment. While moving the block it remains displayed so that the user can see the resulting alignment at each position. The realignment is executed as soon as the mouse button is released. If necessary, the block can be moved over other bases at the left or right side. This will then force a gap to be introduced in the sequences up and down from the block, in order to both preserve the original alignments left and right from the block, and align the block the way the user has forced it to.

A useful tool to select a group of identical bases at once is to click on one of the bases and choose *Sequence > Edit alignment > Highlight identical positions* or CTRL+SHIFT+E on the keyboard.

4.5.7 Inserting and deleting gaps

Besides the easy drag-and-drop realignment tool described above (4.5.6), a number of buttons (and corresponding keyboard shortcuts) are available to manually edit a multiple alignment. Using the editing tools listed below, all changes made to a sequence, i.e. inserting gaps or deleting gaps, cause shifts no further than the next gap. You can consider an aligned sequence as a series of blocks with some space in between (the gaps), just like carriages on a railway: if one block is shifted to the right, it will move alone until it touches the next block, which will then move together, until they touch the next block etc. The following manual alignment editing tools are available:


 Inserts a gap at the position of the cursor, by shifting the block right from the cursor position to the right. This function can be used on a gap as well as on a base. In the latter case, the base at the cursor position will also shift to the right, i.e. the gap will be inserted left from it. Keyboard: INSERT.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCG-GT--ACC-TCTGGCTG-TGGTCCTTA
```

Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG-GT--ACC- TTCTGGCTG-TGGTCCTTA
```


 Inserts a gap at the position of the cursor, by shifting the block left from the cursor position to the left. This function is similar to the previous function. Keyboard: HOME.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG-GT--ACC- TTCTGGCTG-TGGTCCTTA
```

Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG-GT--ACCT TTCTGGCTG-TGGTCCTTA
```


 Inserts gaps at the position of the cursor, by shifting the block right from the cursor position to the right, until it closes up with the next block. Keyboard: SHIFT+INSERT.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG- ACC--TTCTGGCTG-TGGTCCTTA
```

Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG- ACC TTCTGGCTG-TGGTCCTTA
```

 Inserts gaps at the position of the cursor, by shifting the block left from the cursor position to the left, until it closes up with the next block. Keyboard: SHIFT+HOME.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG--TACC -TTCTGGCTG-TGGTCCTTA
```

Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGGTACC -TTCTGGCTG-TGGTCCTTA
```


 Deletes a gap by shifting the block right from the gap to the left. Keyboard: CTRL+DEL.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG--TACC TTCTGGCTG-TGGTCCTTA
```

Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG--TACC TTCTGGCTG--TGGTCCTTA
```


 Deletes a gap by shifting the block left from the gap to the left. Keyboard: END.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG---TACC TTCTGGCTG-TGGTCCTTA
```

Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG---TACC TTCTGGCTG-TGGTCCTTA
```


 Deletes all gaps right from, and including the cursor, by shifting the block right from the gap to the left. Keyboard: SHIFT+DEL.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG- -TACC-TTCTGGCTG-TGGTCCTTA
```

Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG- ACC--TTCTGGCTG-TGGTCCTTA
```

 Deletes all gaps left from, and including the cursor, by shifting the block left from the gap to the right. Keyboard: SHIFT+END.

Example:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TCTTACCGG- -TACC-TTCTGGCTG-TGGTCCTTA
```


Result:

```
TTGGACCGGGAAAAAATTTCTTTCTGATAGTCCTTA
TATCTTACCGG -TACC-TTCTGGCTG-TGGTCCTTA
```

4.5.7.1 To insert and delete gaps or move blocks of a group of sequences as a whole, it is possible to lock a branch on the dendrogram by selecting the branch on the dendrogram (click on the dendrogram node) and *Sequence > Lock/unlock dendrogram branch*.

Locked branches are displayed in red and can be unlocked using the same command.

When no dendrogram is present for a set of aligned sequences, it is also possible to create groups of locked sequences, as follows.

4.5.7.2 Make sure no dendrogram is present with the multiple alignment. If a dendrogram is present, right-click in the *Dendrogram* panel and select  *Show dendrogram*.

4.5.7.3 Select a consecutive group of entries using CTRL + left-click or SHIFT + left-click. When selected, the entries are marked with colored arrows.

4.5.7.4 In the *Sequence* menu, select *Create locked group*. Locked sequences are connected by a red brace in the left panel.

4.5.7.5 To unlock locked groups of sequences, click on any of the entries within the group, and select *Sequence > Unlock group*.

Note that locked groups are not the same as locked branches on the dendrogram (4.5.7.1). When the dendrogram is shown, the locked groups will not be seen anymore, whereas clusters on the dendrogram that were locked previously, become visible and active. When the dendrogram is removed again, the locked groups become visible and active again.

Locked groups have the advantage over locked dendrogram branches that the sequences within a locked group are not restricted to clusters from the dendrogram. One can rearrange the sequences in the multiple alignment as desired (4.5.5.3), and then create groups of locked sequences.

Locked groups or locked branches will not react as a group on sequence editing functions.

4.5.8 Removing common gaps in a multiple alignment

4.5.8.1 After a series of manual realignments, it may be possible that the multiple alignment contains one or more common gaps, i.e. gaps that occur over all sequences. Instead of having to remove those gaps for all sequences separately, the user can let the software find and remove all common gaps automatically.

4.5.8.2 To remove common gaps automatically, select *Sequence > Edit alignment > Remove common gaps* or press CTRL+SHIFT+G on the keyboard.

4.5.9 Changing sequences in a multiple alignment

In some cases, it is possible that ambiguous positions in certain sequences can be filled in when a multiple alignment of highly homologous sequences is present. InfoQuest FP offers the possibility to change bases in sequences within a multiple alignment.

4.5.9.1 Place the cursor on any base in the multiple alignment.


4.5.9.2 Hold the CTRL key and type a base letter, a space (gap) or any letter corresponding to the IUPAC nucleotide naming code.

The sequence is now changed in the multiple alignment, but not yet in the InfoQuest FP database.

4.5.9.3 In order to reload the original sequence, select *Sequence > Reload sequence from database*. The existing alignment will be preserved.

4.5.9.4 To save the changed sequence to the database, select *Sequence > Save changed sequences*.

As an alternative, a base in a sequence can also be changed by double-clicking on that base, which will pop up the experiment card of the sequence, with the clicked base selected. To change the base, simply type in another base from the keyboard. Upon exiting the experiment card, the software will ask to save the changes. These changes are immediately updated in the multiple alignment.

NOTE: If the sequence was assembled from trace files using the Assembler program, it is NOT recommended to modify the sequence in the Comparison window or experiment card. Instead, open Assembler by pressing the  button in the caption of the sequence experiment card and change the base there. When the assembly is saved, the experiment card and the Comparison window are automatically updated.

4.5.10 Finding a subsequence

In order to find certain subsequences in a sequence from a multiple alignment, e.g. restriction sites, primer sequences, repeat patterns etc., you can perform a subsequence search.

4.5.10.1 First, select a sequence within the multiple alignment (white rectangular cursor).

4.5.10.2 The *Subsequence search* dialog box (Figure 4-62) is popped up with *Sequence > Find sequence pattern*.

You can enter any sequence including unknown positions, which are entered as a question mark. You also can allow a number of mismatches to occur in matching subsequences, by specifying a number under *Mismatches allowed*.

For rare subsequences which you do not expect to occur more than once, select *Complete sequence*. For frequently occurring subsequences, you can place the cursor at the start of the sequence, and check *Right from cursor*. By successively pressing *Find*, all subsequent matching patterns will be shown. Similarly, *Left from cursor* shows the first matching pattern left from the cursor, whereas *Closest to cursor* only shows the matching pattern closest to the cursor, in any direction.

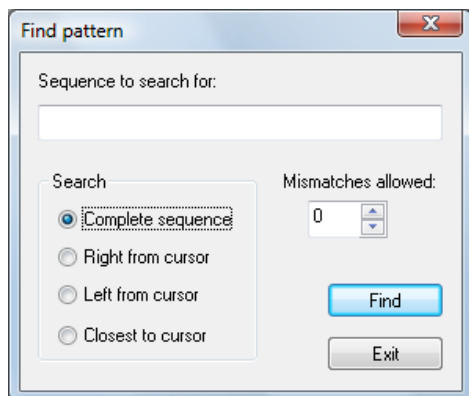



Figure 4-62. Subsequence search dialog box.

4.5.11 Calculating a clustering based on the multiple alignment (steps 5 and 6)

The mutual similarities between all the sequences are calculated from the aligned sequences as present in the multiple alignment.

4.5.11.1 Select *Sequence > Calculate global cluster analysis* or ; the *Global alignment similarity* dialog box is shown (Figure 4-63):

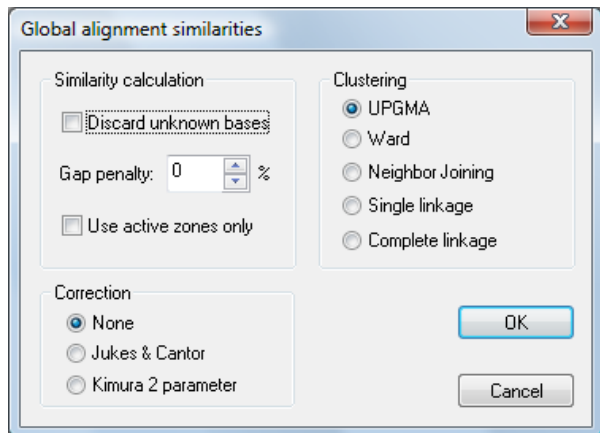


Figure 4-63. Global alignment similarity dialog box.

These settings determine the way the similarity is calculated between the pairs of sequences. The *Gap penalty* is a parameter which allows you to specify the cost the program uses when one single gap is introduced. This cost is relative to the score the program uses for a base matching, which is equal to 100%. The program uses 0% as default. When *Discard unknown bases* is disabled, the program will use a predefined cost table for scoring uncertain or unknown bases. For example, N with A will have 75% penalty, as there is only 25% chance that N is A. Y and C will be counted 50% penalty because Y can be C or T with 50% chance each. If this setting is enabled, all uncertain and unknown bases will not be

considered in calculating the final similarity. To obtain a dendrogram, you can choose between the available clustering algorithms.

The checkbox *Use active zones only* is only applicable when a reference sequence is defined, and when certain zones on this reference sequence are excluded for analysis (see 4.5.17 for more information on excluding regions for comparison).

Under *Correction*, one can select the *Jukes and Cantor (1969)*¹ correction, a *one parameter* correction for the evolutionary distance as calculated from the number of nucleotide substitutions. Alternatively, the *Kimura 2 parameter* correction (Kimura, 1980)² can be selected. In either case, the resulting dendrogram displays a distance scale which is proportional to an evolutionary time, rather than a similarity scale.

4.5.11.2 Check *Discard unknown bases*, select a *Gap penalty* of 0, and leave *Use active zones only* unchecked. Apply no *Correction* and select *Neighbor Joining* as clustering method.

4.5.11.3 Press <OK> to start calculating the multiple alignment-based dendrogram. This calculation is usually fast.

4.5.11.4 Select *File > Save* or  to save the clustering based on the multiple alignment.

4.5.12 Adding entries to and deleting entries from an existing multiple alignment

The feature of InfoQuest FP that makes it possible to add entries to (or delete entries from) an existing cluster analysis also applies to sequence clustering: it is not necessary to recalculate the complete similarity matrix because the program can calculate the similarity of the new sequence(s) with each of the other sequences and will add these new similarity values to the existing matrix. Particularly in case of sequence clustering, this feature is extremely time-saving and causes no degeneration of the clustering.

In case a multiple alignment exists, the problem is slightly more complex. As soon as sequences are added, the program will have to recalculate the multiple alignment (steps 3 and 4 of scheme in Figure 4-55) to find the optimal alignment again for the new set of sequences. This could cause corrections in the alignment made by the user to be lost each time sequences are added. Therefore, the program offers some additional features to add sequences to existing multiple alignments without

1. Jukes, T.H. and C.R. Cantor. 1969. In "Mammalian Protein Metabolism III" (H.N. Munro, ed.), p. 21. Academic Press, New York.
2. Kimura, M. J. 1980. Mol. Evol. 16: 111.

affecting the existing alignment, including manual corrections.

4.5.12.1 In database **DemoBase**, open comparison **All**, and display the previously created multiple alignment (see 4.5.3).

4.5.12.2 If a dendrogram based on the multiple alignment is shown, select **Sequence > Show global cluster analysis** to undo displaying the global cluster analysis.

4.5.12.3 The pairwise dendrogram appears; if not, choose **Layout > Show dendrogram**.

4.5.12.4 Select some entries and cut them from the analysis with **Edit > Cut selection**.

The dendrogram based on pairwise similarities is recalculated immediately, and the multiple alignment is preserved, since deleting entries does not influence the multiple alignment.

NOTE: A dendrogram based on a multiple alignment (global cluster analysis) is not displayed any more when entries are removed from or added to an existing multiple alignment.

4.5.12.5 Paste the selection (which is still on the clipboard) again with **Edit > Paste selection**.

The matrix based upon pairwise alignments and the corresponding dendrogram are now being updated, and when finished, the pasted sequences are shown in the multiple alignment. However, the program has NOT aligned them.

4.5.12.6 Inspect where the pasted sequences are inserted in the dendrogram (colored arrows).

4.5.12.7 If the pasted sequences constitute one single branch, select that branch on the dendrogram, and **Sequence > Align internal branch**.

4.5.12.8 The sequences within the branch are now being aligned internally, and once this is finished, you can select **Sequence > Align external branch** to align the sequences from the rest of the dendrogram with the selected branch.

The advantage of this approach is that by using the **Align internal branch** feature, only the sequences within the selected branch are aligned. This is useful to update a part of a multiple alignment without affecting the non-selected branches. With the **Align external branch** feature, the selected branch is aligned to the rest of the dendrogram as a whole: all sequences within the branch are treated as one block, and all the other sequences are treated as another block. The two blocks are aligned to each other. These features give the user full control over how new sequences are added to a multiple alignment without affecting any editing.

4.5.12.9 A similar result can be obtained with the **Align selected sequences** function (see 4.5.13).

4.5.13 Automatically realigning selected sequences

4.5.13.1 With the function **Sequence > Align selected sequences**, any set of selected sequences can be realigned within an existing multiple alignment. The function preserves any automatic and manual alignment that exists between all the non-selected sequences, which are treated as one block. The selected sequences are aligned one by one to the non-selected sequences. The difference with the method described in 4.5.12 is that the new sequences are not first aligned among each other, which may produce a slightly different result.

4.5.14 Sequence display and analysis settings

A number of settings related to a sequence type are stored as initial settings. These include display settings as well as alignment, clustering, and conversion settings. The initial settings can be changed in the **Sequence type** window (Figure 4-67).

4.5.14.1 To open the **Sequence type** window, double-click on a sequence type in the **Experiments** panel of the **InfoQuest FP main** window (or select **Experiments > Edit experiment type**).

4.5.14.2 With **Settings > Comparison settings**, you can edit the pairwise comparison settings as explained in 4.5.2.

4.5.14.3 Using **Settings > Global alignment settings**, the settings for calculating a global alignment (see 4.5.3) can be edited.

4.5.14.4 With **Settings > Global alignment comparison settings**, the settings for calculating cluster analysis from a multiple alignment can be edited, as explained in 4.5.11.

4.5.14.5 The menu **Settings > Character conversion settings** allows the parameters to be set for converting bases into categorical characters (see 4.5.16).

4.5.14.6 **Settings > Display settings** allows the color and viewing settings in the multiple alignment editor to be specified.

4.5.14.7 The **Sequence display settings** window (Figure 4-64) provides two defaults for color settings: the **White** default, which corresponds to the most widely used colors for the bases on a white background, and the **Black** default, which uses a black background in the multiple alignment editor, using the base color scheme of earlier versions of InfoQuest FP.

4.5.14.8 Apart from the two defaults, every item can be assigned a specific color using the slide bars for the **Red**, **Green** and **Blue** components. Characters can be chosen to indicate gaps and consensus positions.

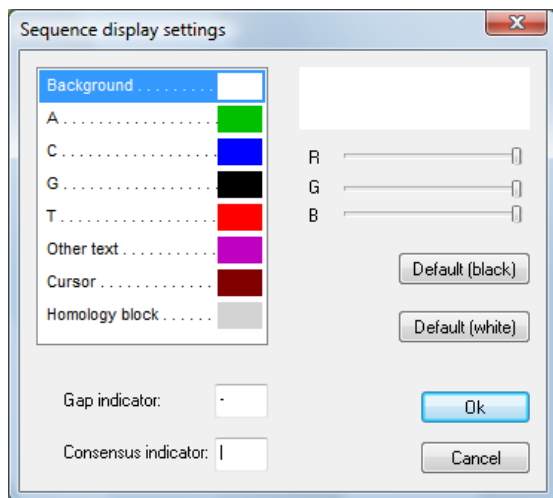


Figure 4-64. The *Sequence display settings* window.

4.5.15 Exporting a multiple alignment

4.5.15.1 Aligned or unaligned sequences in a comparison can be exported as text file with the command *File > Export sequences (tabular)*.

The program now asks "Do you want to export the database fields?".

4.5.15.2 Answer <Yes> to export tab-delimited database fields along with the sequences.

Next, the program asks "Do you want to include regions with gaps?".

4.5.15.3 Answer <Yes> if you want to preserve the gaps introduced in the multiple alignment.

This allows aligned sequences to be exported from InfoQuest FP to other software applications. Gaps are represented as spaces.

4.5.15.4 A more advanced way to export alignments is provided by the *Formatted sequence export* tool, using *File > Export sequences (formatted)*.

This tool allows alignments to be exported in blocks of a defined number of bases, so that multiple blocks can be presented underneath each other on the same page (Figure 4-65).

4.5.15.5 Initially, the window is empty; the view is updated by pressing the <Display> button.

The *Formatted sequence export* window has the following options:

- **Export format** can be *Raw text* or *Rich text*. The latter will export the alignment as RTF, which allows text to be formatted. When *Raw text* is chosen, some options

such as consensus blocks and base coloring do not apply.

- **Number of bases per line** determines the number of base positions to be displayed in one line block. If the total alignment is longer than the number specified, subsequent blocks of the same length will be displayed beneath each other.

- **Include field** allows an information field to be displayed left from the alignment.

- **Numbering** displays a position numbering above the alignment blocks. With *Decimals*, only multipliers of 10 are indicated, whereas with *All numbers*, multipliers of 10 are indicated as an upper line, and unit numbering is indicated in a lower line.

- **Consensus indication** allows the *Consensus blocks* or the *Consensus difference* (if the consensus is calculated in the *Comparison* window) to be indicated (rich text only).

- With *Include consensus sequence*, the consensus sequence is shown on top of the alignment blocks.

- **Use base coloring** will display the bases in color (only rich text mode).

- **Exclude inactive regions** is to exclude regions on the reference sequence that are marked as inactive (see 4.5.17).

- **Selected region only** will only display and export the region in the alignment that is selected using the block selection tool (see 4.5.6). Only the horizontal selection is taken into account (base positions); all sequences are exported regardless of the vertical selection (entries).

4.5.15.6 The view is only updated after pressing the <Display> button.

4.5.15.7 With <Copy to clipboard>, the current view is copied to the clipboard of your operating system and can be pasted in other programs. Depending on the choice, the alignment will be exported as RTF or flat text.

4.5.16 Converting sequence data to categorical character sets

DNA sequence data can be converted into categorical character data, whereby each base is represented by an integer number: A = 1, C = 2, G = 3, and T = 4. The converted categorical data can be visualized and analyzed as a composite data set (4.3.2). The possibility to convert bases into categorical characters requires that a multiple alignment is calculated from the sequences, and also, that a composite data set exists which includes the sequence type (exclusively or in combination with other sequence types).

In addition to the four above states displayed in the composite data set, a fifth state (zero) can optionally be assigned to a gap position. As another option, it is possible to consider only the mutating positions, i.e. the positions that differ in at least one sequence from the others.

NOTE: If sequences containing IUPAC nomenclature are converted into categorical character sets, each of the ambiguous bases (e.g. M, R, Y, etc.) becomes an additional state.

4.5.16.1 The settings for converting sequences into character data can be changed in the *Sequence type* window (see 4.5.14). Choose *Settings > Character conversion settings* to open the *Character conversion settings* dialog box.

4.5.16.2 With the option *Exclude non-mutating positions*, only those base positions in a multiple alignment that do not contain the same for all sequences will be included in the character set.

4.5.16.3 With the option *Exclude positions with gaps*, those positions where one or more sequences have a gap in the multiple alignment will be excluded from the

character set. If gaps are not excluded, a fifth state is assigned to gaps (zero).

Converting sequences into character data can have several useful applications:

Minimum Spanning Trees can be calculated from the composite data set (see Section 4.11), thus allowing sequence data to be analyzed using MSTs.

Different genes, each represented in a separate sequence type, can be combined in one composite data set, so that the information from the different genes can be condensed in one single dendrogram. The option *Exclude non-mutating positions* (4.5.16.2) thereby offers the possibility to reduce the amount of information to only those base positions that are polymorphic in the entries analyzed.

In addition to clustering the entries, it is also possible in the composite data set to cluster the base positions, using the Transversal Clustering method (see 4.8.3). The result looks like in Figure 4-66, where groups of bases are clustered together according to their discriminatory behavior between groups of entries.

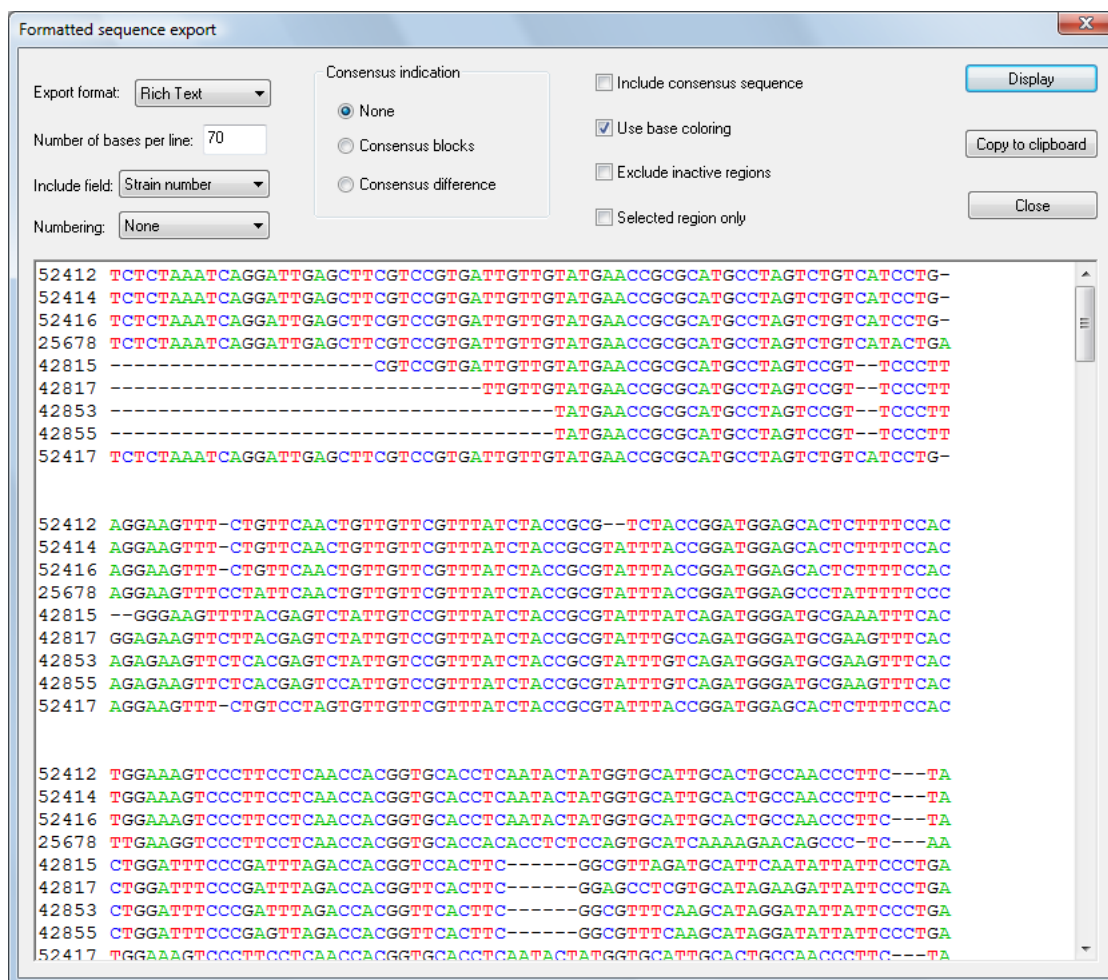


Figure 4-65. The *Formatted sequence export* window.

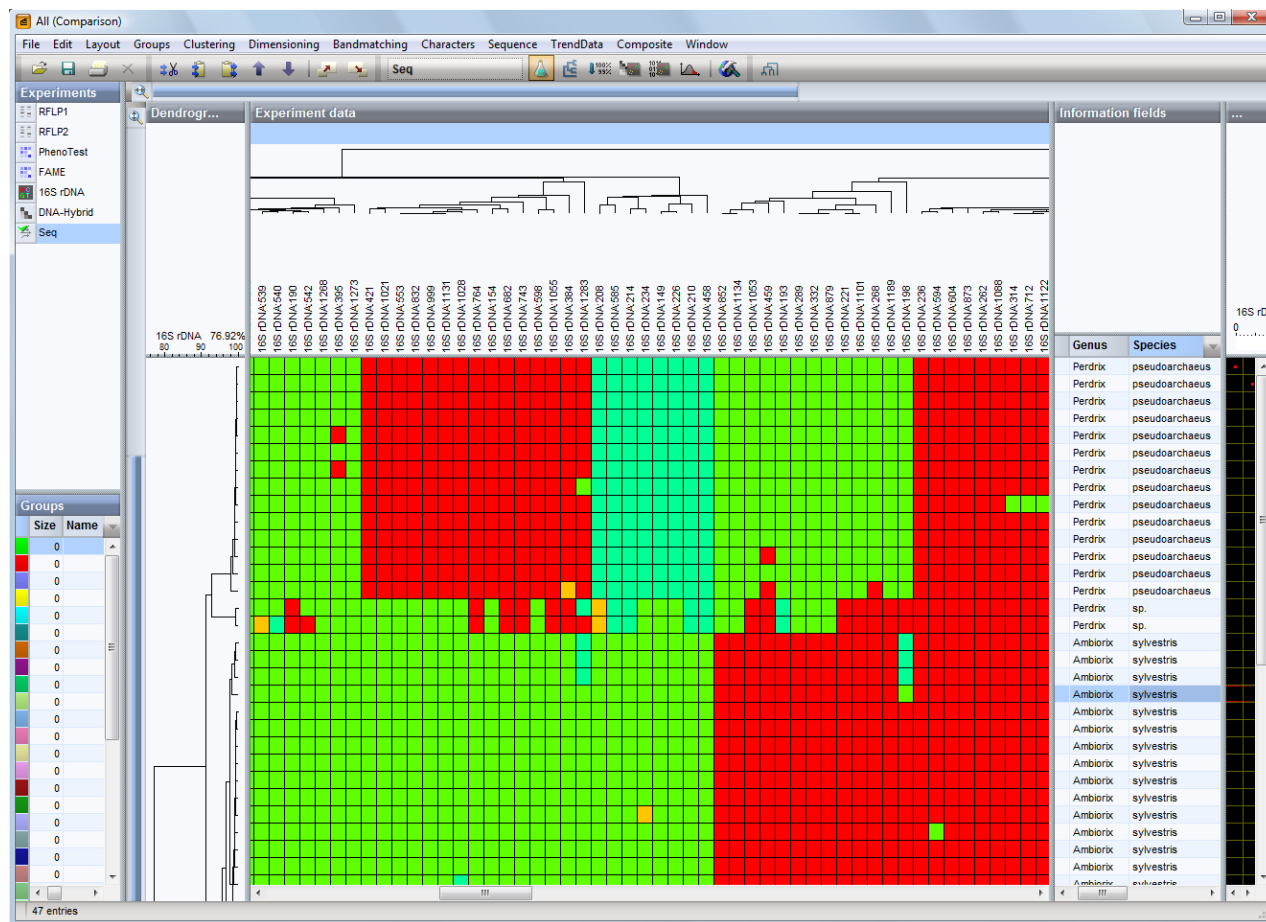


Figure 4-66. Comparison window showing composite data set generated from a sequence type. Bases were converted into categorical characters and clustered in both directions.



4.5.16.4 In this view (Figure 4-66), the bases can be shown as letters (default, to be obtained with *Composite > Show presence/absence*), as colors (using *Composite > Show quantification (colors)*), or as numbers (with *Composite > Show quantification (numbers)*). In the color view, "A" is shown in magenta, "C" in green, "G" in orange and "T" in red; a gap is blue. In the numbered view, "A" is 1, "C" is 2, "G" is 3, and "T" is 4; a gap is zero.


4.5.17 Excluding regions from the sequence comparisons

Similar as for the comparison of fingerprints, it is possible to exclude regions from the sequences to be clustered. First, one needs to define a reference sequence, and next, one can indicate the zones to be excluded and included on the reference sequence. The exclusion of regions is only possible when calculating a cluster analysis based upon globally aligned sequences (multiple alignment) and when the reference sequence is included in the multiple alignment. Only then, the program can introduce a consistent base numbering based on the reference sequence, which makes it



possible to specify the same exclude/include settings for different multiple alignments within the sequence type.

4.5.17.1 In the *InfoQuest FP main window*, open the *Sequence type* window of **16S rDNA** by double-clicking on **16S rDNA** in the *Experiments* panel.

Initially, there is no reference sequence present. A *link arrow*  allows you to link a reference sequence to a database entry, by clicking on the arrow and dragging it onto a database entry, and then releasing the mouse button. When the experiment is linked, its link arrow is purple: .

4.5.17.2 Drag the link arrow to database entry *Vercingetorix palustris* strain no. 42819: as soon as you pass over a database entry, the cursor shape changes into .

4.5.17.3 Release the mouse button on database entry *Vercingetorix palustris* strain no. 42819.

This entry is now defined as reference sequence, and the arrow in the *Sequence type* window has become purple  instead of gray . The reference sequence is shown in the *Sequence type* window (Figure 4-67).

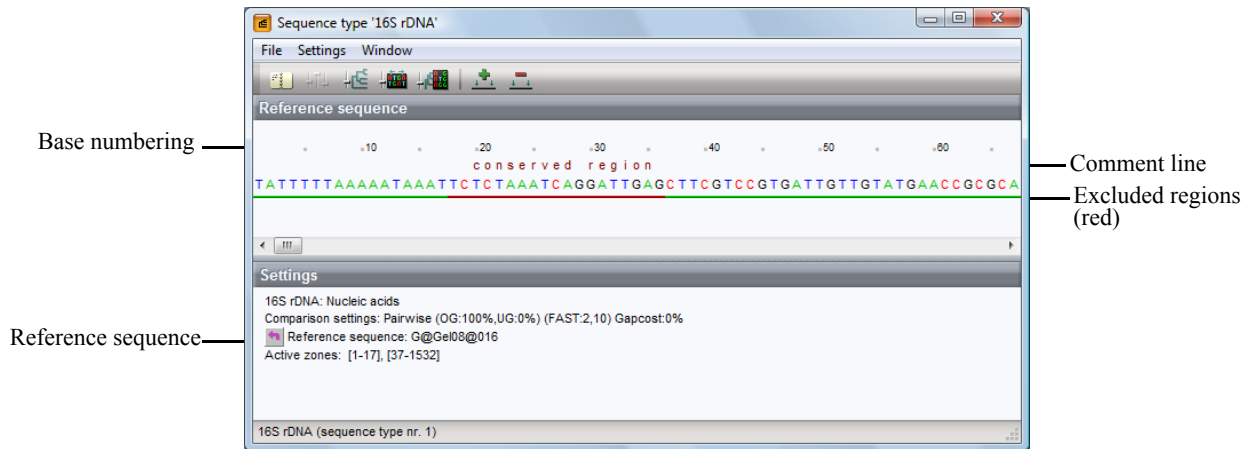




Figure 4-67. Sequence type window with reference sequence defined, region excluded, and comments added.

4.5.17.4 Select *Settings > Exclude active region* or  to exclude a region for comparison.

4.5.17.5 Enter start and end base number of the region to be excluded, and press the **<OK>** button.

4.5.17.6 The included regions are marked with a green line whereas the excluded regions are marked with a red line (Figure 4-67).

4.5.17.7 In order to remove all excluded regions at a time, select *Settings > Include active region* or . Enter 1 as *From* number, and enter the length of the sequence, or a number which certainly exceeds the sequence's length, as *To* number.

4.5.17.8 Open comparison **All** (or another comparison which you have saved with the aligned sequences).

4.5.17.9 Show the image of the aligned sequences.

4.5.17.10 Select the branch top of *Vercingetorix palustris* strain no. 42819 in the *Dendrogram* panel (the reference sequence) and *Sequence > Create consensus of branch*.

By creating a consensus of a single sequence, you can display the reference sequence in the consensus sequence line (Figure 4-67). At the same time, the excluded and included regions are indicated, and the base numbering, according to the reference sequence appears (Figure 4-67).

4.5.17.11 In order to see the base numbering it may be necessary to drag the horizontal line that separates the header from the *Experiment data* panel downwards.

4.5.18 Writing comments in the alignment

In order to mark special regions on the reference sequence or on the multiple alignment, a simple comment editor allows you to add any comment to the comparison. The comments can only be added when a reference sequence is present, when the sequences are aligned, and when a consensus sequence is shown. The comment line is saved along with the sequence type, and new comments can be added at any time.

4.5.18.1 Click the cursor on one of the aligned sequences in the image. At the position of the cursor, you can start writing comments.

The comments appear in the image header, above the consensus sequence (Figure 4-67). Any character input is supported. A's, C's, G's and T's are written in the colors of the bases.

4.5.18.2 To delete a comment, place the cursor on any sequence at the position of the first character of the comment and enter spaces.

4.6 Sequence alignment and mutation analysis

4.6.1 Introduction

The *Alignment* window is a convenient tool for the calculation of multiple sequence alignments, subsequence searches and mutation analysis. In this respect, it forms a more powerful alternative to the sequence analysis features available in the *Comparison* window. The *Alignment* window allows different views of the alignment to be displayed in different panels, for example a panel with the chromatograms (curves) and a panel with the bases. The cursor position is synchronized between the different panels. This feature, together with the fact that curves can be displayed for the trace files in the contigs, allows for a quick and reliable evaluation of the correctness of positions of interest, including mutations and SNPs.

The features of the *Alignment* window will be illustrated using an example dataset, which is available on the installation CD-ROM or from our website. The trace files originate from influenza A virus strains and represent partial sequences of the haemagglutinin (HA) and neuraminidase (NA) genes.

These publicly available trace files were downloaded from the NCBI Trace Archive (<http://0-www.ncbi.nlm.nih.gov.catalog.llu.edu/Traces/trace.cgi?>).

In the InfoQuest FP Startup screen, create a new database. You can call it e.g. **DemoAlign**. Leave all settings to their defaults.

4.6.1.1 In the *Plugin installation* toolbox, install the **Batch sequence assembly** plugin and the **Import** plugin (see 1.5.3 for instructions on how to install plugins).

4.6.1.2 In the *InfoQuest FP main window*, select **File > Batch sequence assembly > Batch sequence assembly**.

The .SCF trace files for the partial HA and NA sequences are located in the **Sample and Tutorial data\Example_traces** directory on the installation CD-ROM. Alternatively, the files are available on the download page of the website (www.bio-rad.com/software-downloads).

4.6.1.3 Browse for the **Example_traces** folder, select all trace files and press **<Open>**.

4.6.1.4 In the *File parsing settings* dialog box that appears (see Figure 4-68), select SCF as file format and check **Fetch experiment from file parsing** and press **<Proceed>**.

4.6.1.5 In the *File parsing* dialog box, use [EXP]_[KEY]-* as parsing string and press **<Parse>**. The key and exper-

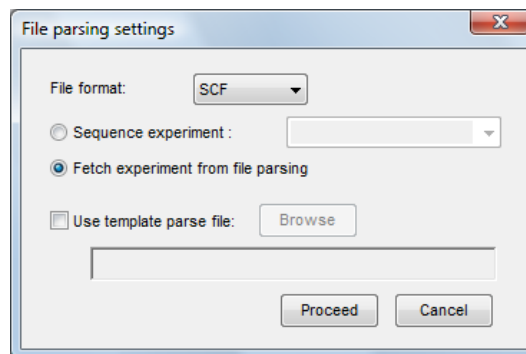


Figure 4-68. The *File parsing settings* dialog box from the **Batch sequence assembly** plugin.

iment name is parsed from the filename as shown in Figure 4-69. Press **<Proceed>** to continue.

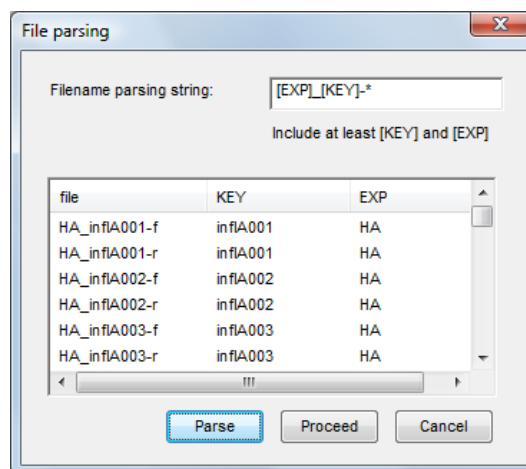


Figure 4-69. The *File parsing* dialog box from the **Batch sequence assembly** plugin.

4.6.1.6 In the *Experiments settings* dialog box that appears, **HA** and **NA** are listed under *Experiments missing in database*. Press **<Create>** to have the two sequence type experiments automatically created by the software.

4.6.1.7 Press **<Trimming settings>** to pop up the *Trimming settings* dialog box.

4.6.2 For the HA sequences in the example dataset, enter the trimming settings as specified in Figure 4-70 and press **<Save settings>**.

4.6.3 For the NA sequences in the example dataset, enter following trimming settings: start trim pattern "CCAGTAG", stop trim pattern "CGGGGTC" and stop

position offset -20. When completed, press <Save settings>.

4.6.3.1 Press <OK> to close the *Trimming settings* dialog box.

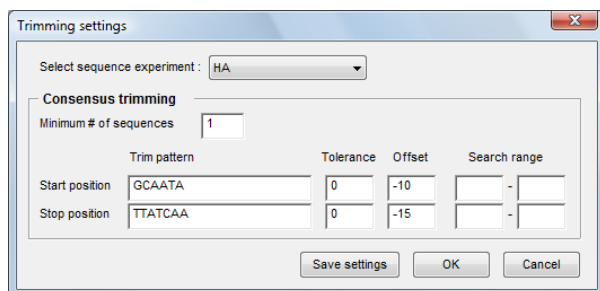


Figure 4-70. The *Trimming settings* dialog box, displaying trimming settings for the HA sequence example data.

For more information on the assembly settings, see the separate manual that comes with the Batch sequence assembly plugin.

4.6.3.2 Press <Proceed> and then <Assemble> to have the 19 sequences automatically assembled by the Batch sequence assembly plugin.

Database information fields can be imported from the separate text file **Strain_info.txt**, located in the **Sample and Tutorial data\Example_traces\Text_files** directory on the installation CD-ROM. Alternatively, the file is available on the download page of the website (www.bio-rad.com/softwaredownloads).

4.6.3.3 If not already installed, install the Import plugin from the *Plugin installation* toolbox. See 1.5.3 for more information on plugins.

4.6.3.4 In the *InfoQuest FP main window*, select **File > Import > Import fields and characters**.

4.6.3.5 In the *Import fields & characters* dialog box, check **Text file** and browse for the **Strain_info.txt** file.

4.6.3.6 Leave TAB selected as separator and press <OK>.

4.6.3.7 All external fields should be associated with a *Text field* and the *Link control field* should be 'Key'.

4.6.3.8 The external field 'Key' should be associated with the text field 'Key' in the database. For all other external fields, <Create new> should be selected.

4.6.3.9 When <OK> is pressed, the Import plugin will suggest information field names based on the external field names. Press <OK> to confirm the names for each *Database information field* dialog box that pops up.

Information fields for the five database entries are now imported and can be displayed in e.g. the *InfoQuest FP main, Comparison and Alignment* window.

In the *InfoQuest FP main window*, the *Alignments* panel is displayed in default configuration as tabbed view with the *Comparisons, Libraries and Decision Networks* panel in the bottom right part of the window. If desired, the configuration of the *InfoQuest FP main window* can be customized as described in . In the manual, however, the default configuration will be used.

4.6.4 Creating a new alignment project

4.6.4.1 Click on the *Alignments* tab to display the *Alignments* panel in the *InfoQuest FP main window*.

4.6.4.2 To create a new alignment project, select **Comparison > Alignments > Create new** or press the



button in the *Alignments* panel.

4.6.4.3 Enter a name for the new alignment project, e.g. **MyAlignment** and press <OK>.

The *Alignments* panel in the *InfoQuest FP main window* now contains one alignment project, called **MyAlignment** (see Figure 4-71). The date on which the alignment project was created and last modified is displayed in the default information fields 'Created' and 'Modified', respectively. When more than one alignment project is present, projects can be sorted and searched using the information present in the default or user-defined information fields. .

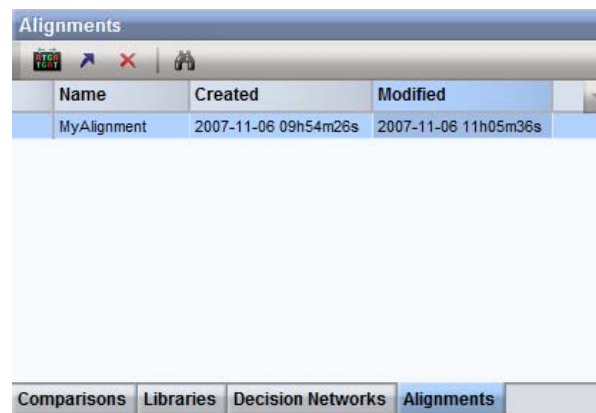





Figure 4-71. The *Alignments* panel in the *InfoQuest FP main window*, showing a single alignment project.

4.6.4.4 To delete an alignment project from the list, click on it in the *Alignments* panel (the alignment project now becomes highlighted) and press the  button.

NOTE: The first time an alignment project is opened, it will open with the currently selected entries in the *Database entries panel* (colored selection arrows). As soon as an alignment project has been saved, pressing the  button will open the alignment project with

the entries that were present when the alignment project was last saved.

4.6.4.5 Press CTRL+A on the keyboard to select all entries in the *Database entries* panel.

4.6.4.6 Press the  button in the *Alignments* window to open the newly created alignment project with the selected database entries.

The *Experiment types* dialog box opens, displaying a list of available sequence types (see Figure 4-72). From this list, the user can select the experiment types for which a sequence alignment should be created within the alignment project.

4.6.4.7 Leave the experiment types **HA** and **NA** selected in the list and press <OK> to display the alignment project in the *Alignment* window.

4.6.5 The *Alignment* window

The *Alignment* window (see Figure 4-73) consists of eight panels: the *Dendrogram*, *Sequence display 1*, *Information fields*, *Similarities*, *Sequence display 2*, *Mutation listing*, *Sequence search results*, and *Bookmarks* panel. All panels

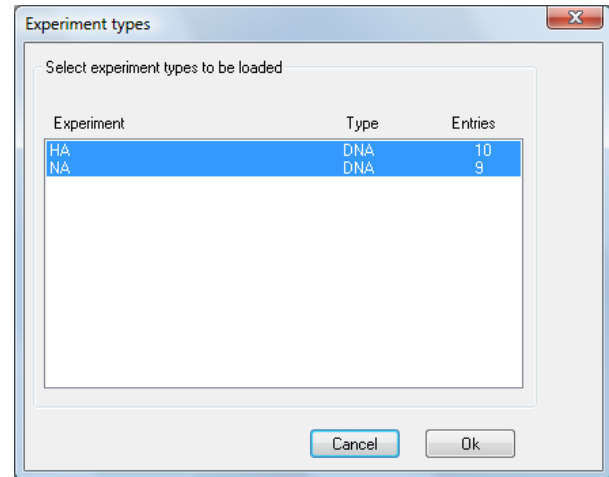


Figure 4-72. The *Experiment types* dialog box.

are dockable, which enables the user to customize the layout of the *Alignment* window according to personal preference and/or the type of analysis to be performed..

NOTE: The Dendrogram, Sequence display 1, Information fields and Similarities panels behave as a group, i.e. these panels cannot be docked outside this



Figure 4-73. The *Alignments* window with example data displayed.


group and they cannot be displayed in a separate window (undocked).



- **Dendrogram** panel: Displays an UPGMA or Neighbor Joining dendrogram calculated on the sequences present in the alignment project and for the sequence type selected. See 4.6.13 and 4.1.13 for dendrogram-related tools.
- **Sequence display 1** panel: Displays the nucleic acid sequences present in the alignment. Optionally, the corresponding curves and the translated (amino acid) sequences can be shown as well. See 4.6.10 for display options of the *Sequence display 1* panel.
- **Information fields** panel: This grid panel displays the entry information fields in tabular format, similar to e.g. the *Database entries* panel in the *InfoQuest FP main* window or the *Information fields* panel in the *Comparison* window.
- **Similarities** panel: Displays the matrix of similarity values, calculated on the sequences present in the alignment. This panel is similar to the *Similarities* panel of the *Comparison* window. See 4.1.14 for matrix display functions.
- **Sequence display 2** panel: Is nearly identical to the *Sequence display 1* panel except that it displays by default the curves (chromatograms) instead of the sequences. See 4.6.10 for display options of the *Sequence display 2* panel. Since two sequence display panels are available, sequences and curves can be displayed simultaneously, each in their respective panel.
- **Mutation listing** panel: Allows a search to be launched of a selection of sequences from the alignment against a consensus sequence. When a search is performed, it lists all mutations for that search. See 4.6.20 for detailed information.
- **Sequence search results**: Allows a search for a subsequence to be launched for a selection of sequences from the alignment. When a search is performed, it lists all occurrences of this subsequence. See 4.6.19 for detailed information.
- **Bookmarks** panel: Lists all bookmarks that are added to sequences in the alignment. See 4.6.21 for detailed information.

The upper part of the *Alignment* window contains the main menu and toolbars. The latter can be displayed or hidden according to your preferences.

4.6.5.1 You can drag the horizontal and vertical separator lines between the panels, in order to divide the space among the panels optimally.


4.6.5.2 In the *Information fields* panel, you can drag the separator lines between the information field columns to the left or to the right, in order to divide the space among the information fields optimally.

4.6.5.3 Clicking the column properties button () located on the right hand side in the information fields header in the *Information fields* panel gives access to functions to hide, freeze, or move information fields.

4.6.5.4 The zoom sliders indicated with  and  can be used to zoom selectively in the horizontal or vertical direction, respectively. See 1.6.7 for a detailed description of zoom slider functions.

4.6.6 General functions

When an alignment project is saved, all calculations done on the sequences it contains will be stored along. This includes similarity matrices and dendrograms for all sequence alignments. In a connected database, the alignment project will be saved by default within the connected database, so that it can be shared between users.

4.6.6.1 Select *File > Save project* or press  to save the alignment project (shortcut CTRL+S on the keyboard).

4.6.6.2 To reset the active sequence alignment, select *Alignment > Reset*. The *Dendrogram* panel and *Similarities* panel will be emptied and sequences in the alignment project will be unaligned. In case position-based search results are mapped on the alignment, e.g. sequence searches (see 4.6.19), mutation listings (see 4.6.20) or bookmarks (see 4.6.21), the program will warn about this under default general settings (see 4.6.6.8 for information about general settings). If you confirm the action, the position-based search results will be lost.

4.6.6.3 A copy of the active alignment can be created with *Alignment > Create copy*.

This option allows the user to keep the current alignment unaltered, while at the same time, alternative settings can be evaluated on a copy of the alignment. The active alignment can be selected from the drop-down list in the main toolbar of the *Alignment* window (see Figure 4-74).

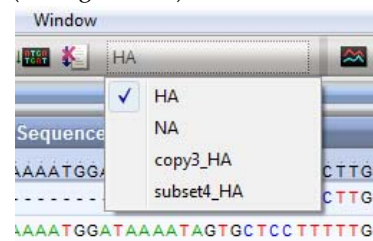


Figure 4-74. Drop-down list in the main toolbar of the *Alignment* window to select the active alignment.

Sometimes, the user is specifically interested in a certain region within an existing alignment. It is possible to

create a new alignment, containing only this specific region.


4.6.6.4 Using the mouse, select the region from the alignment (sequence block), in which you are interested.

4.6.6.5 Select *Alignment > Create subset*.

Similar as for a copy, the active alignment can be selected from the drop-down list in the main toolbar of the *Alignment* window (see Figure 4-74).

4.6.6.6 To delete an active alignment, select *Alignment > Delete*. If only one alignment is present (the active one), all entries will be removed from the alignment project upon executing this command.

Depending on the number of entries, the length of the sequences and the algorithm selected, sequence alignments and clusterings can take a long processing time. In order to speed up the calculations, or make multi tasking smoother, you may want to modify the calculation priority settings. See 4.1.10.1 to 4.1.10.2 for instructions on how to change these settings.

4.6.6.7 While the program is calculating, you can abort the calculations at any time using the  button.

4.6.6.8 A number of general settings for the alignment project can be accessed via *File > General settings*. This pops up the *General settings* dialog box (see Figure 4-75).

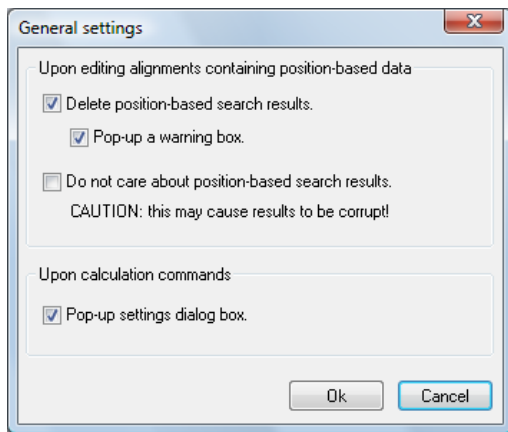


Figure 4-75. The *General settings* dialog box.

If *Delete position-based search results* is checked, sequence searches, mutation listings and bookmarks will be deleted when an alignment is changed, e.g. by recalculating a dendrogram, swapping branches of a dendrogram, or manual editing of an alignment. In case *Pop-up a warning box* is checked, the program will display a warning and will allow the user to cancel the operation prior to deleting the position-based search results. If *Do not care about position-based search results* is checked, the program will continue to display


the sequence searches, mutation listings and bookmarks, even though the actual positions may not correspond anymore.

The setting *Pop-up settings dialog box (Upon calculation commands)* is checked by default. If it is unchecked, commands to calculate alignments and/or clusterings will be executed using the last specified settings, without first displaying the corresponding settings dialog box.

In a multiple alignment with highly homologous sequences, it is possible that an ambiguous position in a certain sequence can be filled in.

4.6.6.9 To edit a sequence, place the cursor on the position to be edited (either in the *Sequence display 1* or *Sequence display 2* panel), right-click and select *Open sequence experiment card* from the floating menu.

The sequence experiment card opens, with the selected position highlighted. It is possible to edit the sequence directly in the experiment card. However, this way of working is not advised when the sequence was imported as trace files via Assembler, as this will break the link with Assembler. A better option is to edit the sequence in the Assembler program.

4.6.6.10 Press the  button in the sequence experiment card to launch Assembler with the corresponding assembly project loaded and the sequence position highlighted.

4.6.6.11 The sequence can now be edited as described in . Make sure to save the assembly project before closing Assembler.

The edited sequence is automatically updated in the *Alignment* window. Position-based search positions will be lost.


4.6.7 Adding and removing entries


Selections of entries made in the *Database entries* panel of the *InfoQuest FP main* window are also shown in the *Information fields* panel of the *Alignments* window and vice versa. The entries in a newly created alignment project are all marked with a colored selection arrow, since they were all selected from the database. You can manually select and unselect entries in the *Information fields* panel (see Figure 4-73), using the CTRL and SHIFT keys as described in .

Selected entries can be added to or removed from an existing alignment.


4.6.7.1 First unselect all entries by pressing the F4 key.



4.6.7.2 Select some entries from the alignment project.

4.6.7.3 With *Edit > Cut selected sequences* or  (shortcut CTRL+X on the keyboard), the selected entries are removed from the alignment project and copied to the clipboard.

4.6.7.4 With *Edit > Paste selected sequences* or  (shortcut CTRL+V on the keyboard), the same entries are placed back into the alignment project.


Entries can be added to an existing alignment project at any time. The entries first need to be copied to the clipboard from the *InfoQuest FP main* window or e.g. from a comparison or another alignment project.

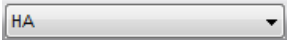
4.6.7.5 To copy entries to the clipboard, select the entries (e.g. in the *InfoQuest FP main* window) first and use the *Edit > Copy selection* command or  (shortcut CTRL+C on the keyboard).

4.6.7.6 To cut entries from one alignment project into another, use *Edit > Cut selected sequences* or  (shortcut CTRL+X on the keyboard) in one alignment project and *Edit > Paste selected sequences* or  (shortcut CTRL+V on the keyboard) in the other alignment project.

NOTE: When adding entries to or deleting entries from an alignment project, the dendrogram, similarity matrix and sequence alignment need to be calculated again.


4.6.8 Aligning sequences

In order to obtain a multiple sequence alignment, a pairwise alignment similarity matrix and a dendrogram (most often a UPGMA dendrogram) should be calculated first (see 4.5.1 for more background information on sequence analysis). The *Alignment* window of InfoQuest FP offers the possibility to calculate a multiple sequence alignment by means of a single command (*Alignment > Calculate > Multiple alignment* or ). The settings for the successive pairwise and multiple alignment steps are grouped within a single dialog box.

In case the project consists of more than one alignment, e.g. if multiple sequence types were imported for the selected entries, the active (i.e. currently displayed) sequence alignment can be selected from the drop-down list in the main toolbar: .

4.6.8.1 Select experiment type **HA** from the drop-down list.

Initially, the *Dendrogram* and *Similarities* panels are empty, since no dendrogram is calculated yet.

4.6.8.2 Select *Alignment > Calculate > Multiple alignment* or press the  button.

The *Multiple alignment settings* dialog box pops up, allowing a number of parameters to be set (see Figure 4-76). The *Algorithm* for pairwise sequence comparisons can be set to be InfoQuest FP own proprietary algorithm (*InfoQuest FP*), *Needleman-Wunsch*¹ or *Wilbur-Lipman*² (ClustalW). Depending on which algorithm is selected, the settings which can be specified, differ. If **InfoQuest FP** is selected as algorithm, the dialog box is displayed as in Figure 4-76 and the following settings can be specified.

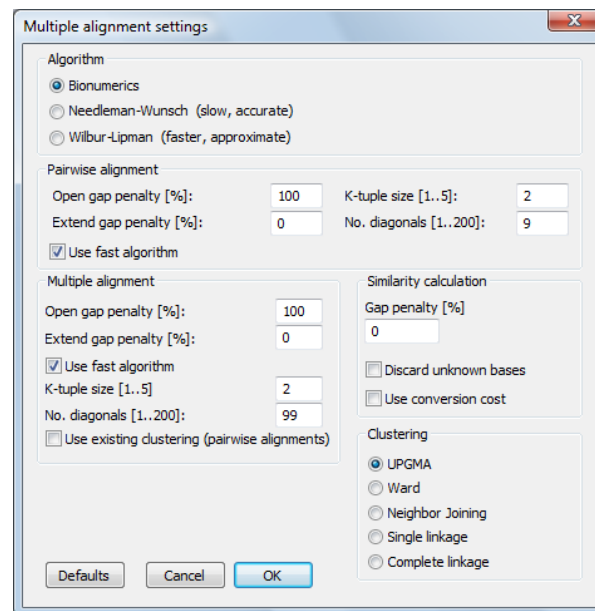


Figure 4-76. *Multiple alignment settings* dialog box, InfoQuest FP algorithm checked.

The *Pairwise alignment* settings are grouped together: The *Open gap penalty* is the cost, expressed as a percentage, to introduce a gap in a sequence. The default value is 100%, which is the same penalty as a mismatch. The *Extend gap penalty* is the cost (in percent) to increase an existing gap with one position. The default value is 0. The parameters *K-tuple size* and *No. diagonals* are only available when the option *Use fast algorithm* is checked. This algorithm creates a lookup table of groups of bases for both sequences (*words*). The *K-tuple size* is the size of such a word. The smaller the words are, the more precise the alignment will be, but the more computing time it will take to calculate the alignment. The parameter can be varied between 1 and 5, with 2 as default. The number of diagonals (*No. diagonals*) is the maximum number of relative positions

1. Needleman, S. and C. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.* 48,443-453.
2. Wilbur, W.J. and D.J. Lipman. 1983. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* 80,726-730.

between both sequences the algorithm will consider. The values can be varied between 0 and 200 with 9 as default. The larger the number, the more gaps the algorithm can create to align every two sequences, but the longer the alignment will take. Therefore, the two parameters allow you to custom-define the alignment process from very fast and relatively rough, to slow and very accurate.

For the *Similarity calculation*, a *Gap penalty* is a cost that can be specified as a percentage of a match (default value is 0). *Discard unknown bases* lets you decide whether or not ambiguous bases are taken into account. If unchecked, the program uses a predefined cost table for scoring ambiguous bases. Checking *Use conversion cost* makes the similarity calculation faster, e.g. for draft alignments, by transforming the calculated conversion cost into a similarity value.

As *Clustering* algorithm can be selected between *UPGMA*, *Ward*, *Neighbor Joining*, *Single linkage* and *Complete linkage*.

The *Multiple alignment settings* include an *Open gap penalty* (default value: 100) and an *Extend gap penalty* (default value: 0). When the option *Use fast algorithm* is checked, the additional parameters *K-tuple size* (default value: 2) and *No. diagonals* (default value: 99) become available. Checking *Use existing pairwise clustering* allows you to calculate the multiple alignment based on an existing pairwise clustering (see 4.6.8.6). This option can be used e.g. to employ a ClustalW tree as input for the InfoQuest FP multiple alignment algorithm.

If the *Needleman-Wunch* algorithm is used, the *Multiple alignment settings* dialog box is displayed as in Figure 4-77 and following settings apply.

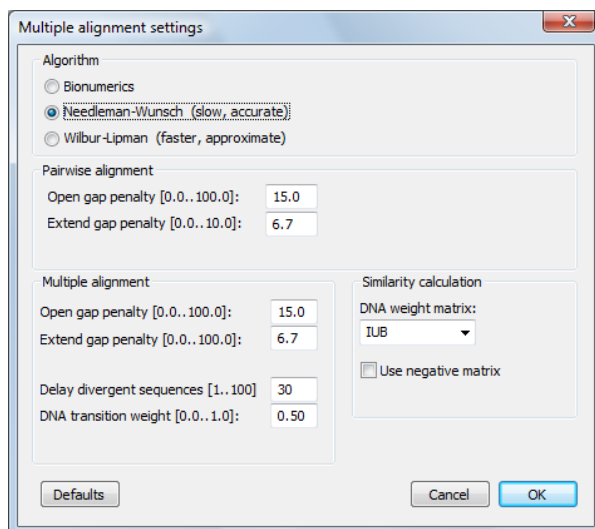


Figure 4-77. Multiple alignment settings dialog box, Needleman-Wunsch algorithm checked.

Under the *Pairwise alignment* settings, an *Open gap penalty* (default value: 15%) and an *Extend gap penalty*

(default value: 6.7%) can be specified, similar as for the InfoQuest FP algorithm.

Under the *Similarity calculation* settings, the *DNA weight matrix* can be selected. The choice is offered between the default *IUB* (International Union of Biochemistry) and the *CLUSTALW* DNA weight matrix. Check *Use negative matrix* to allow the use of negative values in the DNA weight matrix.

The *Multiple alignment settings* include an *Open gap penalty* (default value: 15) and an *Extend gap penalty* (default value: 6.7), similar to the InfoQuest FP algorithm. *Delay divergent sequences* allows you to delay the alignment of the most distantly related sequences until after the most closely related sequences have been aligned. The value displayed is the percent identity level below which the addition of a sequence is delayed; sequences that are less identical than this level to any other sequences in the alignment will be aligned later. The default value is 30. The *DNA transition weight* gives transitions (A <--> G or C <--> T, i.e. purine-purine or pyrimidine-pyrimidine substitutions) a weight between 0 and 1; a weight of zero means that the transitions are scored as mismatches, while a weight of 1 gives the transitions the match score. The default value is 0.50. For distantly related DNA sequences, the weight should be set near to zero; for closely related sequences it can be useful to assign a higher score.

If the *Wilbur-Lipman* algorithm is used, the *Multiple alignment settings* dialog box is displayed as in Figure 4-78 and following settings apply.

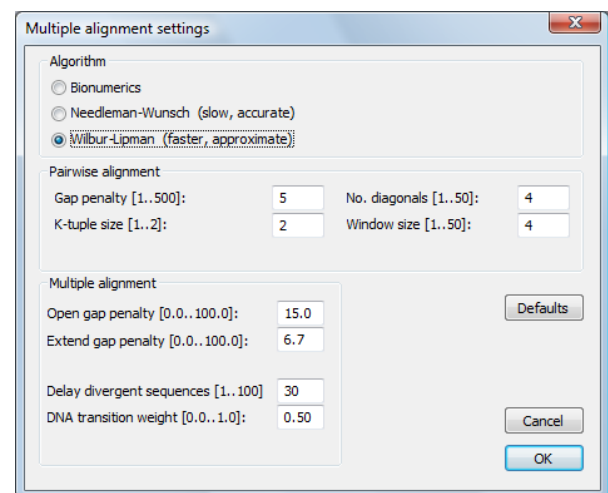


Figure 4-78. Multiple alignment settings dialog box, Wilbur-Lipman algorithm checked.

Under the *Pairwise alignment* settings, a *Gap penalty* (default value: 5), a *K-tuple size* (default value: 2), and the *No. diagonals* (default value: 4) can be specified. These parameters are similar as explained for the InfoQuest FP algorithm. The *Window size* is a parameter specific for the Wilbur-Lipman algorithm and corresponds to a region within the similarity scores matrix

where matches are considered. The higher this value is set, the more accurate the pairwise alignment will be, but the more calculation time required. The default value is 4.

The *Multiple alignment* options are identical to the options described above for the Needleman-Wunch algorithm.


4.6.8.3 Irrespective of the selected algorithm, pressing the **<Defaults>** button will restore the default settings for the *Multiple alignment settings* dialog box, i.e. the InfoQuest FP algorithm with default parameters.

4.6.8.4 Leave all parameters at their default value and press **<OK>** to calculate a multiple alignment using the InfoQuest FP algorithm.

NOTE: The dendrogram and similarity matrix that are displayed after calculating a multiple alignment are still based on pairwise similarity values. See 4.6.12 on how to calculate a global cluster analysis.

The *Alignment* window also offers the option to calculate a multiple alignment as a two-step process, i.e. calculate a pairwise alignment and dendrogram first and then calculate a multiple alignment based on the previously obtained pairwise clustering. This option can be useful, e.g. when one wants to base a multiple alignment on a ClustalW dendrogram or when applying a Jukes and Cantor correction (see 4.6.8.7). A pairwise alignment and dendrogram can be calculated as follows:

4.6.8.5 First, reset the calculated alignment by selecting **Alignment > Reset**. This step is strictly spoken not necessary, but it makes understanding of the workflow in the following paragraphs easier.

4.6.8.6 Select **Clustering > Calculate > Pairwise clustering** or press the  button.

The *Pairwise clustering settings* dialog box pops up (see Figure 4-79). Similar as for the *Multiple alignment settings* dialog box, the parameters that can be set depend on the algorithm selected. Figure 4-79 shows the *Pairwise clustering settings* dialog box with the InfoQuest FP algorithm checked (the default option).

For any of the three algorithms selected, the *Pairwise alignment*, *Clustering* and *Similarity calculation* settings are the same as explained for the *Multiple alignment settings* dialog box (see Figure 4-77).

In addition to the parameters discussed above, a parameter **Correction** is available for the InfoQuest FP algorithm in the *Pairwise clustering settings* dialog box. The options are **None** to use no correction (the default choice) or to select the **Jukes and Cantor (1969)¹** correction, a *one parameter* correction for the evolutionary

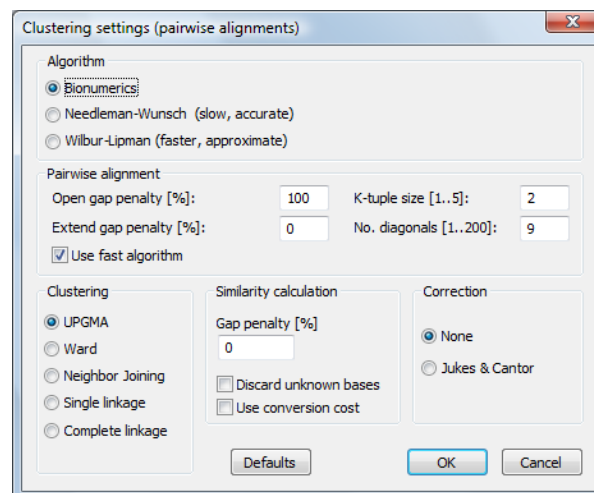



Figure 4-79. The Pairwise clustering settings dialog box, InfoQuest FP algorithm checked.

distance as calculated from the number of nucleotide substitutions.

4.6.8.7 Select for example the *Needleman-Wunsch* algorithm, check **Use ClustalW tree** and press **<OK>** to calculate the pairwise clustering.

The ClustalW dendrogram and similarity matrix as calculated by the Needleman-Wunsch algorithm appear. The sequences in the *Sequence display 1* panel are still unaligned.

The obtained pairwise clustering can now be used to calculate a multiple alignment.

4.6.8.8 Select **Alignment > Calculate > Multiple alignment** or press the  button.

4.6.8.9 In the *Multiple alignment settings* dialog box, select **InfoQuest FP** as algorithm, check **Use existing pairwise clustering** and press **<OK>**.


The multiple alignment that is now displayed is calculated using the InfoQuest FP algorithm, based on a pairwise alignment and ClustalW dendrogram calculated by the Needleman-Wunsch algorithm.

4.6.9 Calculating a consensus sequence

A consensus sequence can be used to obtain a multiple alignment of all sequences against one single sequence. Depending on the type of analysis, the user may wish to assign a single sequence as consensus or may want to calculate the consensus sequence based on several sequences. A consensus sequence is also required for mutation searches (see 4.6.20) and allows additional identity display settings for the alignment (see 4.6.10) to be chosen. To allow maximum flexibility, a consensus sequence is always calculated for the currently selected entries in an alignment.

1. Jukes, T.H. and C.R. Cantor. 1969. In "Mammalian Protein Metabolism III" (H.N. Munro, ed.), p. 21. Academic Press, New York.

4.6.9.1 Select the entries that you want to base the consensus on. For this example, select all entries in the alignment (CTRL+A on the keyboard).

4.6.9.2 Select *Alignment > Consensus > Create from selected entries* or press the  button. The *Consensus definition* dialog box pops up (see Figure 4-80).

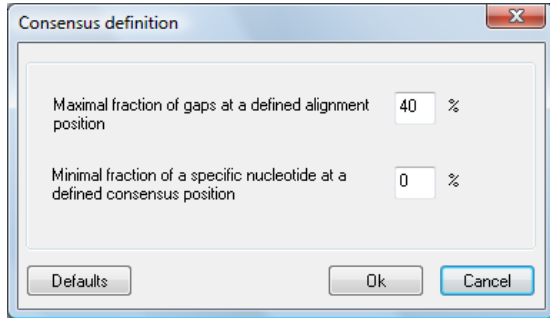


Figure 4-80. The *Consensus definition* dialog box.


The parameter *Maximal fraction of gaps at a defined alignment position* determines the maximum occurrence of a gap at a certain position in the sequences on which the consensus is calculated. If the actual occurrence is higher than the specified percentage, the position will be excluded from the consensus sequence. The default value is 40%.

The *Minimal fraction of a specific nucleotide at a defined consensus position* defines the “threshold” occurrence that a nucleotide should reach in order to contribute to the consensus. The default value of 0% will result in any mismatch leading to an ambiguous base in the consensus.


4.6.9.3 Leave the setting for gaps at the default value but enter 20% for *Minimal fraction of a specific nucleotide at a defined consensus position*. Press <OK> to calculate a consensus based on the selected sequences.

The consensus sequence is displayed in the header of the *Sequence display 1* panel. Entries that were used to calculate the consensus are preceded with a blue dot in the *Information fields* panel. Note that positions for which eight sequences have one nucleotide (e.g. “C”) and the two others have another (e.g. “T”), the 20% threshold is met and the consensus displays the IUPAC code for an ambiguous position (e.g. “Y”). Positions for which nine out of ten sequences have the same nucleotide, cause the consensus to have that same nucleotide (the second occurrence is only 10%).

When changes are made to the alignment (e.g. when recalculated using other settings or after manual editing), the consensus can be recalculated using the same sequences and calculation settings:


4.6.9.4 Select *Alignment > Consensus > Recalculate* or press the  button.

When a consensus sequence is present, it can be used to align other sequences against. This option obviously only makes sense in an alignment containing highly related sequences.

4.6.9.5 Select *Alignment > Calculate > Pairwise alignment* or press the  button. This pops up the *Pairwise alignment settings* dialog box.

The parameters that can be set in the *Pairwise alignment settings* dialog box are identical to those discussed for the *Pairwise clustering settings* dialog box (see Figure 4-79).


4.6.9.6 Press <OK> to align all sequences against the consensus sequence.

4.6.9.7 To remove a calculated consensus sequence, press F4 to unselect any entries in the *Alignment* window and select *Alignment > Consensus > Create from selected entries* or press the  button.


4.6.10 Display options for sequences and curves


The sequence alignment is displayed in the *Sequence display 1* panel by default. The *Sequence display 2* panel is configured to display the curves (chromatograms) of the individual trace files by default. Before the curves can be displayed, they need to be loaded first.

4.6.10.1 In the alignment project **MyAlignment**, select **HA** from the drop-down list in the main toolbar.

4.6.10.2 To load the curves into the alignment project, select *Alignment > Load curves* or press the  button.

The *Sequence display 2* panel now shows the curves for the experiment type **HA**.

The display options for the *Sequence display 1* panel and the *Sequence display 2* panel can be set via *Options > Sequence display 1* and *Options > Sequence display 2*, respectively. Alternatively, press the settings button () in the upper left corner of the corresponding sequence display panel.

4.6.10.3 Select *Options > Sequence display 1* or press the corresponding settings button () to set the display options for the *Sequence display 1* panel. The *Display settings* dialog box appears (see Figure 4-81).

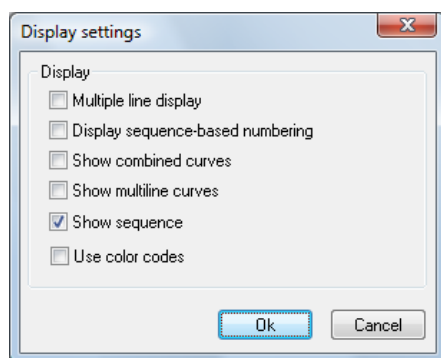


Figure 4-81. The *Display settings* dialog box, default values for the *Sequence display 1* panel shown.

For *Sequence display 1* panel, *Show sequence* is checked by default. *Multiple line display* means that the alignment will be wrapped into the width of the panel and displayed on more than one line. If a dendrogram is calculated, it will be repeated for each line. Consequently, the horizontal scroll option disappears in the sequence display panel. *Display sequence-based numbering* shows position numbers for each individual sequence. *Show combined curves* and *Show multiline curves* display the curves (chromatograms) of the sequences respectively superimposed or on different lines. *Use color codes* displays the bases and amino acids in the sequence in different colors.

NOTE: Curves can only be displayed if they are available for the sequence (i.e. if the sequences were generated from chromatogram files and assembled using the Assembler program in InfoQuest FP, and if the curves are loaded into the alignment project (see 4.6.10.2).

4.6.10.4 Color codes for nucleic acids can be changed by selecting *Options > Text color settings > Nucleic acids*. This pops up the *Color code settings* dialog box for nucleic acids (see Figure 4-82).

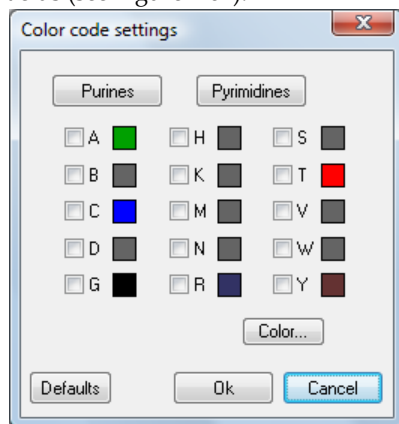


Figure 4-82. The *Color code settings* dialog box for nucleic acids.

4.6.10.5 To change the color for a specific nucleotide or a number of nucleotides, select the nucleotide(s) via the check box (e.g. select *A*) and press *<Color>*. This pops up a color picker from which standard colors can be picked and/or custom colors defined.

4.6.10.6 Select any color of your choice and press *<OK>*. The *Color code settings* dialog box now displays the new color for adenine.

4.6.10.7 The buttons *<Purines>* and *<Pyrimidines>* provide a shortcut to select specifically the purine and pyrimidine nucleotides, respectively.

4.6.10.8 Press *<Defaults>* to restore the default colors for nucleic acids and *<OK>* to close the dialog box.

4.6.10.9 Color codes for amino acids can be changed by selecting *Options > Text color settings > Amino acids*. This pops up the *Color code settings* dialog box for amino acids (see Figure 4-83).

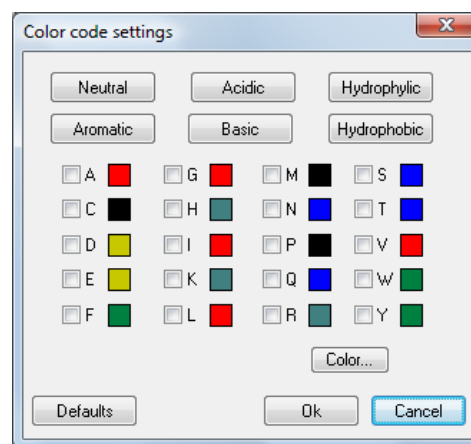


Figure 4-83. The *Color code settings* dialog box for amino acids.

Color settings for amino acids can be changed in a similar way as described for nucleic acids (see 4.6.10.5 to 4.6.10.6).

4.6.10.10 The buttons *<Neutral>*, *<Acidic>*, *<Hydrophilic>*, *<Aromatic>*, *<Basic>*, and *<Hydrophobic>* provide a shortcut to select specifically the corresponding group of amino acids.

A number of options are designed to enhance the visualization of conserved parts in the alignment. They are specific to the *Sequence display 1* panel, and are grouped in the menu item *Alignment > Identity display*.

4.6.10.11 Select *Alignment > Identity display > Conserved blocks* to display the sequence positions that are conserved throughout the alignment in grey.

4.6.10.12 Select *Alignment > Identity display > Neighbor identity blocks* to display sequence positions in grey when at least one of the neighboring sequences has the same nucleotide at the corresponding position.

The two other options (*Identity with consensus* and *Difference with consensus*) are self-explanatory but require a consensus to be calculated first (see 4.6.9).

The zoom sliders of the *Sequence display 1* panel and the *Sequence display 2* panel can be used to zoom selectively in the horizontal and vertical direction. For detailed information on the use of zoom sliders, see 1.6.7. When the *Sequence display 1* panel is zoomed vertically, the *Dendrogram*, *Information fields* and *Similarities* panel are zoomed proportionally.

4.6.11 Editing an alignment

A multiple alignment as calculated by the software (see 4.6.8) can be edited via drag-and-drop of individual positions or sequence blocks. The manually edited sequence alignment is saved along with the alignment project and is used to base the global clustering on (see 4.6.12).

The *Sequence display 1* and *Sequence display 2* panels contain a cursor which is synchronized between both panels. Similar as in a text processor, the cursor always appears between two characters (bases or amino acids).

4.6.11.1 In the alignment project **MyAlignment**, select **HA** from the drop-down list in the main toolbar.

4.6.11.2 If no multiple alignment is present, calculate one as described in 4.6.8.

4.6.11.3 Using the mouse, select one of the corner positions of the block to define and while holding down the left mouse button, drag the mouse pointer to define the desired block. Note that a block can comprise a single or several sequences. The selection is visible as a black rectangle (see Figure 4-61).

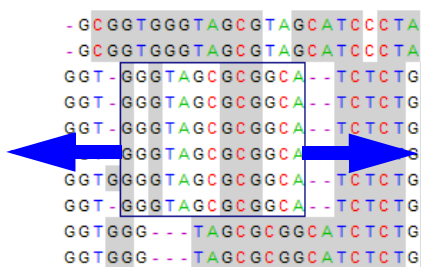


Figure 4-84. Selecting blocks of bases for drag-and-drop manual alignment.

4.6.11.4 A selection block can also be made using the keyboard, by holding down the SHIFT key while pressing the arrow keys.



4.6.11.5 Click within the selection and, while keeping the mouse button pressed, drag it to towards the left or the right to create a gap. The sequences are automatically realigned as soon as the mouse button is released.

If necessary, the block can be moved over other bases at the left or right side. This will then force a gap to be introduced in the sequences up and down from the block, in order to preserve the original alignments left and right from the block and to align the block the way the user has forced it to.

NOTES:

(1) A gap common for all sequences in the alignment project, i.e. spanning the complete alignment, will be automatically removed. As a consequence, nothing will happen if you try to realign a block of bases spanning the whole alignment.


(2) Double-clicking a position in the alignment will select a continuous stretch (without gaps) of the clicked sequence.

4.6.11.6 For manual alignment editing, as well as for other actions performed on the alignment, a multilevel undo and redo function is available. The undo function can be accessed with *Edit > Undo* or the  button (shortcut CTRL+Z on the keyboard). The redo function is accessible through *Edit > Redo* or the  button (shortcut CTRL+Y on the keyboard).

NOTE: Within the constraints of the dendrogram, the order in which sequences appear in the multiple alignment can be changed by repeatedly swapping dendrogram branches, as described in 4.6.13, until the sequences are listed in the desired order.

4.6.12 Calculating a global cluster analysis

4.6.12.1 To calculate a global clustering based on the sequences in the alignment, select *Clustering > Calculate > Clustering (multiple alignment)* or press the

 button; the *Clustering settings (multiple alignment)* dialog box pops up (see Figure 4-85).

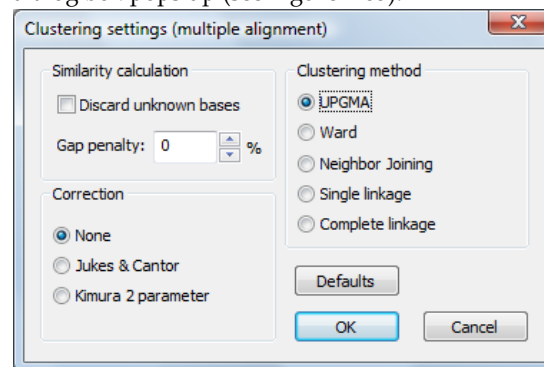


Figure 4-85. Clustering settings (multiple alignment) dialog box.

When *Discard unknown bases* is disabled, the program will use a predefined cost table for scoring uncertain or unknown bases. For example, **N** with **A** will have 75% penalty, as there is only 25% chance that **N** is **A**. **Y** and **C** will be counted 50% penalty because **Y** can be **C** or **T** with 50% probability each. If this setting is enabled, all uncertain and unknown bases will not be considered in calculating the final similarity. The *Gap penalty* is a parameter which allows you to specify the cost the program uses when one single gap is introduced. This cost is relative to the score the program uses for a base matching, which is equal to 100%. The program uses 0% as default.

Under *Correction*, one can select the *Jukes and Cantor* (1969)¹ correction, a *one parameter* correction for the evolutionary distance as calculated from the number of nucleotide substitutions. Alternatively, the *Kimura 2 parameter* correction (Kimura, 1980)² can be selected.

A selection can be made between the available clustering algorithms: *UPGMA*, *Ward*, *Neighbor Joining*, *Single linkage* and *Complete linkage*.

Pressing the **<Defaults>** button resets all parameters to their default value.

4.6.12.2 Select *Neighbor Joining* and press **<OK>** to calculate the global cluster analysis.

The *Similarities* panel now displays the global similarity matrix, on which the Neighbor Joining tree in the *Dendrogram* panel is calculated.

4.6.13 Dendrogram display functions

A dendrogram is displayed from the moment a pairwise or global clustering is performed (see 4.6.8 and 4.6.12, respectively). Similar as in the *Comparison* window (see 4.1.11), several dendrogram display functions are available in the *Alignment* window.

4.6.13.1 Press F4 to unselect any previous selection of database entries.

Entries can be selected from within the *Dendrogram* panel of the *Alignment* window:

4.6.13.2 To select an individual entry, hold the CTRL key and click on a dendrogram tip (where a branch ends in an individual entry). Alternatively, right-click on the dendrogram tip and choose *Select branch into list* from the floating menu. Repeat this action to unselect the entry.

4.6.13.3 To select a cluster on the dendrogram at once, hold the CTRL key and left-click on a branch node. Alternatively, right-click on a branch and choose *Select branch into list* from the floating menu.

When a dendrogram node or tip is clicked on, a diamond-shaped cursor appears at that position. The average similarity between the entries at the cursor's place is shown in the upper left corner of the *Dendrogram* panel.

In some cases, it may be necessary to select the root of a dendrogram, for example if you want to (un)select all the entries of the dendrogram. In case of large dendrograms, selecting the root may be difficult using the mouse.

4.6.13.4 With *Clustering > Select root*, the cursor is placed on the root of the dendrogram.


Two branches grouped at the same node can be swapped to improve the layout of a dendrogram or make its description easier:

4.6.13.5 Select the node where two branches originate and *Clustering > Swap branches*.

NOTE: When position-based search results, such as sequence searches (see 4.6.19), mutation listings (see 4.6.20) or bookmarks (see 4.6.21), are mapped on the alignment, the program will warn about this when a dendrogram is recalculated or branches are swapped. When continuing with the action, the search results will be lost. This default behaviour can be changed in the general settings for the alignment project (see 4.6.6.8).

Another function, *Clustering > Reroot tree*, only applies to neighbor joining trees in the *Alignment* window. This clustering method produces trees without any specification as to the position of the root or origin (*unrooted* trees). Since users will often want to display such trees in the familiar dendrogram representation, the tree is to be rooted artificially. "Rerooting" is usually done by adding one or more unrelated entries (so-called *outgroup*) to the clustering, and using the outgroup as root. The result is a *pseudo-rooted* tree.

To illustrate the rerooting of an unrooted tree, we will first calculate a neighbor joining tree based on sequence type **HA** in **MyAlignment**.

4.6.13.6 Select *Clustering > Calculate > Clustering (multiple alignment)* or press the  button and specify *Neighbor Joining* in the *Clustering settings* dialog box. A neighbor joining tree is calculated based on the multiple alignment of **HA** sequences.

The entry with key inflA007 (serotype H5N2) is more distantly related to the other entries (all serotype H5N1) and can therefore be used as an outgroup in this analysis.

1. Jukes, T.H. and C.R. Cantor. 1969. In "Mammalian Protein Metabolism III" (H.N. Munro, ed.), p. 21. Academic Press, New York.

2. Kimura, M. J. 1980. Mol. Evol. 16: 111.

4.6.13.7 Click somewhere in the middle of the branch connecting entry inflA007 with the other entries in the alignment. A secondary, X-shaped red cursor appears.

4.6.13.8 Select *Clustering > Reroot tree*, and the new root connects the outgroup with the rest of the entries.

4.6.14 Cluster significance tools

Similar as for comparisons, a number of cluster significance tools are available for alignment projects. For more background on these tools, see 4.1.13.

4.6.14.1 In the alignment project **MyAlignment**, select **HA** from the drop-down list in the main toolbar.

4.6.14.2 Make sure a rooted dendrogram is present by selecting *Clustering > Calculate > Clustering (pairwise alignment)*.

4.6.14.3 Select *Clustering > Calculate error flags*.

The error flags, i.e. the standard deviations of the dendrogram branches compared to the corresponding sections in the similarity matrix, are now shown for each branch (similar as shown in Figure 4-12). The average similarity and the standard deviation at the position of the cursor is shown in the left upper corner.

4.6.14.4 Select *Clustering > Calculate error flags* again to remove the error flags.

The cophenetic correlation (see also 4.1.13), i.e. the correlation between the dendrogram-derived similarities and the matrix similarities, is a parameter to express the consistency of a cluster.

4.6.14.5 Select *Clustering > Calculate cophenetic correlations*.

The cophenetic correlation is now shown at each branch (similar as shown in Figure 4-13), together with a colored dot, of which the color ranges between green, yellow, orange and red according to decreasing cophenetic correlation.

A bootstrap analysis is based on “sampling with replacement” (for more background information, see 4.1.13) and can only be calculated on a global clustering.

4.6.14.6 To calculate a global clustering based on a multiple sequence alignment, select *Clustering > Calculate > Clustering (multiple alignment)* or press the



button and press **<OK>** in the *Clustering settings (multiple alignment)* dialog box (leaving all settings at their defaults).

4.6.14.7 Select *Clustering > Bootstrap analysis*, enter the number of simulations (samplings) to perform (e.g. 100) and press **<OK>**.

The bootstrap values are displayed on the dendrogram in a similar way as the cophenetic correlation values.

4.6.15 Matrix display functions

The similarity matrix is displayed in the *Similarities* panel, in default configuration located at the right hand side of the *Alignment* window (see Figure 4-73).

Initially, the matrix is displayed as differentially shaded blocks representing the similarity values. Similar as for the *Comparison* window, the interval settings for the shadings is graphically represented in the caption of the *Similarities* panel (Figure 4-14).



Figure 4-86. Adjustable similarity shading scale.

There are two ways to change the intervals for shading:

4.6.15.1 Drag the interval bars on the scale; the matrix is updated instantly.

4.6.15.2 Right-click within the *Similarities* panel and select *Similarity shades* from the floating menu. The maximum/minimum values for each interval can be entered as numbers, from low (top) to high (bottom).

NOTE: If it is difficult to read the similarity values on the shaded background, you can remove the shades by entering 100% for each interval.


4.6.15.3 To export a tab-delimited text file of the similarity matrix, select *Clustering > Export matrix*.

4.6.15.4 Select a location to save the text file when the program prompts for a destination path. If desired, the suggested name can be modified.

The text file contains the similarity values with the entry keys as descriptors.

4.6.16 Printing and exporting a sequence alignment

When printing from the *Alignment* window, InfoQuest FP first shows a print preview. This print preview shows the dendrogram, entry keys and sequence alignment in multiline view and looks exactly as it will look on printed pages.

4.6.16.1 In the *Alignment* module, select *File > Show print preview* or press the  button to display the *Print preview* window (see Figure 4-21).

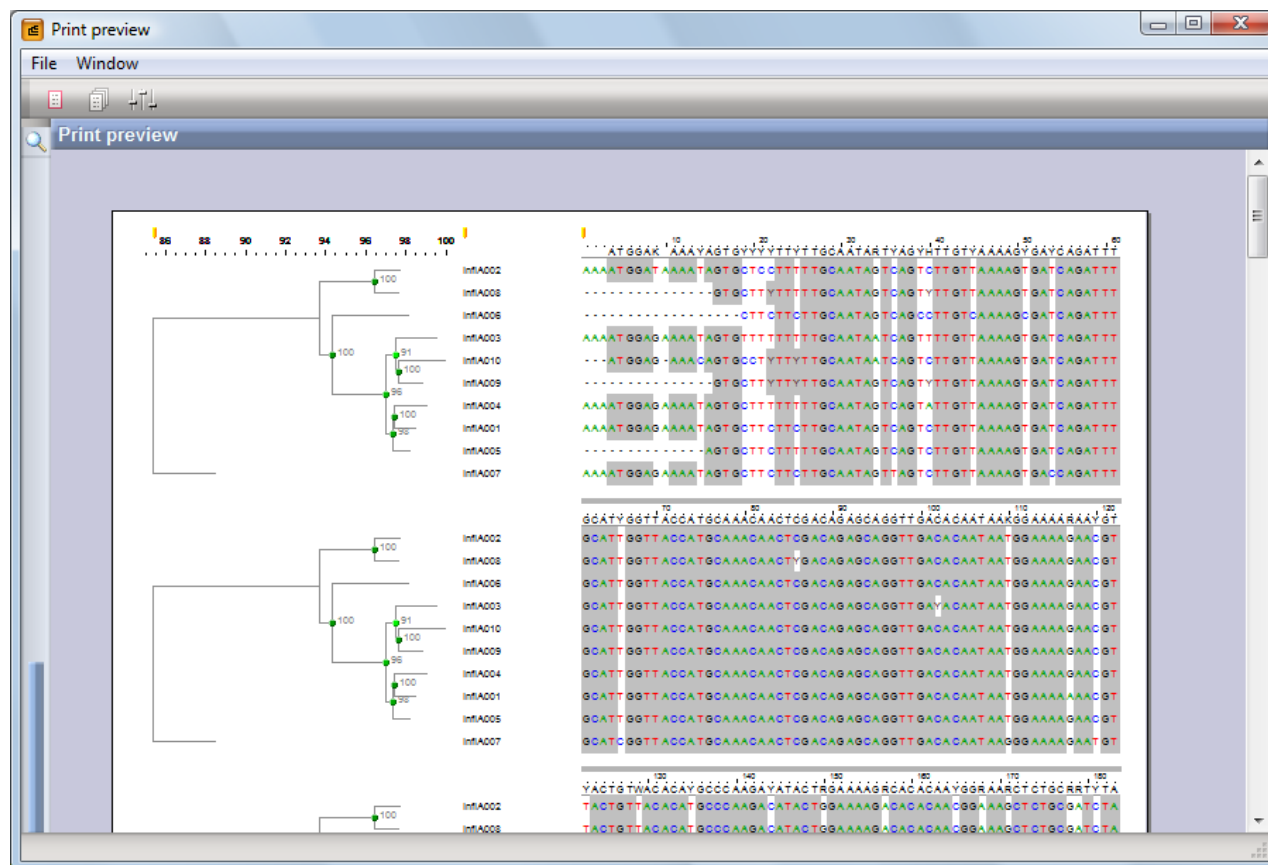


Figure 4-87. The *Print preview* window.


4.6.16.2 It is possible to zoom in and out on a page using the zoom slider, located in default configuration on the left hand side of the *Print preview* panel.


4.6.16.3 When zoomed, the horizontal and vertical scroll bars allow you to scroll through the page.


On top of the first preview page, there are three small yellow slide bars (Figure 4-21). These slide bars represent the following margins, respectively:

- Left margin of the dendrogram;
- Right margin of the dendrogram;
- Left margin of the alignment;

Each of these slide bars can be shifted individually to reserve the appropriate space for the mentioned items. The image is printed exactly as it looks on the preview.

4.6.16.4 The menu command *File > Printer setup* or  allows you to set the paper orientation, the margins, and other printer settings for the default printer.

4.6.16.5 With *File > Print selected pages* or , the selected pages are printed. Selected pages are indicated with a red border. Use CTRL+click to select multiple pages.

4.6.16.6 Use *File > Print all pages* or  to print all pages at once.

NOTE: When a part of the alignment is selected in the Alignment window using the mouse (see 4.6.11), only the selected part will be printed.

4.6.16.7 Close the *Print preview* window.

An alignment or part of an alignment can also be exported, e.g. for reporting purposes.

4.6.16.8 Select the part of the alignment that you would like to export, using the black selection rectangle as described in 4.6.11.3.

4.6.16.9 Select *File > Copy selected alignment to clipboard*.

The selected part of the alignment is now copied to the clipboard and can be pasted in other applications as windows metafile or enhanced metafile.

4.6.17 Finding sequence positions in an alignment

You can have the cursor jump automatically to a certain position on a sequence in the alignment. This can be particularly useful e.g. when examining the occurrence of a certain SNP in a set of equal-length sequences.

4.6.17.1 In the alignment project **MyAlignment**, select **HA** from the drop-down list in the main toolbar.

4.6.17.2 Select **Edit > Find > Position**. This pops up the *Position search* dialog box (see Figure 4-88).

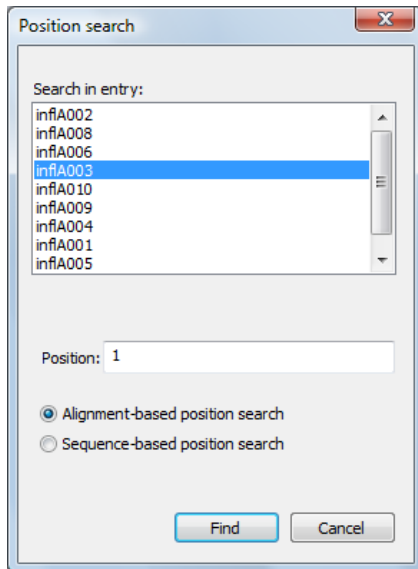


Figure 4-88. The *Position search* dialog box.

The sequence on which the cursor currently resides is highlighted in the *Search in entry* list, but any other sequence can be selected as well. A position can be entered as a number. The alignment-based numbering (*Alignment-based position search*) or the individual sequence numbering (*Sequence-based position search*) can be used. Both are equivalent in case no gaps were introduced in the sequences.

4.6.17.3 Enter a number and press **<Find>**. The cursor jumps to the corresponding position on the sequence. Note that the search position is the position *before* the cursor.

4.6.18 Sequence translation

InfoQuest FP can automatically translate an alignment of nucleotide sequences into amino acids according to a selected translation table and within a certain translation frame. The translated amino acid sequence is displayed in the sequence alignment.

4.6.18.1 To set the settings that will be used for the translation, select **Alignment > Translation > Define**. The *Translation settings* dialog box pops up (see Figure 4-89).

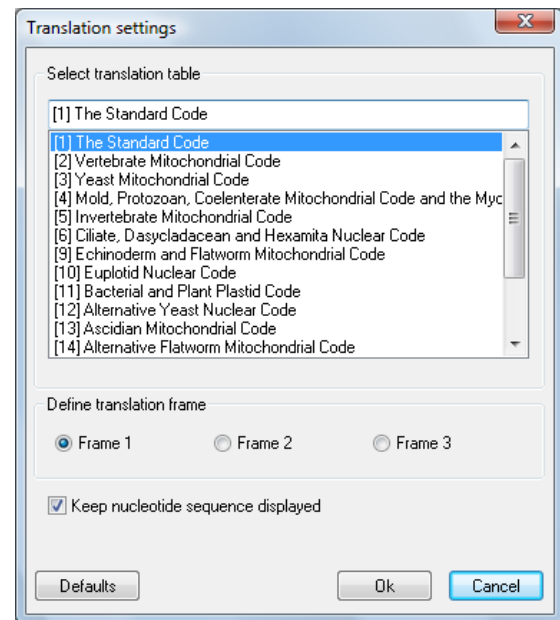



Figure 4-89. The *Translation settings* dialog box, displaying the default settings for translation of nucleotide sequences into amino acid sequences.

From the list under *Select translation table*, the different translation tables, corresponding to variants of the standard genetic code, can be selected. By default, the standard code is used. The translation frame can be set under *Define translation frame*. Uncheck *Keep nucleotide sequences displayed* if you only want the amino acid sequences to be displayed, without the nucleotide sequences from which they originated.

4.6.18.2 For the **HA** experiment type in **MyAlignment**, select **Frame 1** and press **<OK>** to accept the settings.

4.6.18.3 Select **Alignment > Translation > Show/Hide** or press the  button to display the translated amino acid sequence.


If the nucleotide sequences you are working with are never translated, for example in case of ribosomal RNA sequences, select **Alignment > Translation > None** to disable the translation into amino acids. In this case, no amino acid changes will be listed in the *Mutation listing* panel when an alignment is searched for mutations (see 4.6.20 on how to perform a mutation search).

4.6.19 Subsequence search

The complete alignment or any selection of sequences within the alignment can be searched for the occurrence of a subsequence. This subsequence can correspond to

e.g. a restriction site, primer sequence, repeat pattern, or any other specific sequence you are interested in.

4.6.19.1 In the alignment project **MyAlignment**, select **HA** from the drop-down list in the main toolbar.

4.6.19.2 Select **Edit > Find > Sequence** from the main menu or press the  button in the *Sequence search results* panel. The *Find sequence* dialog box appears (see Figure 4-90).

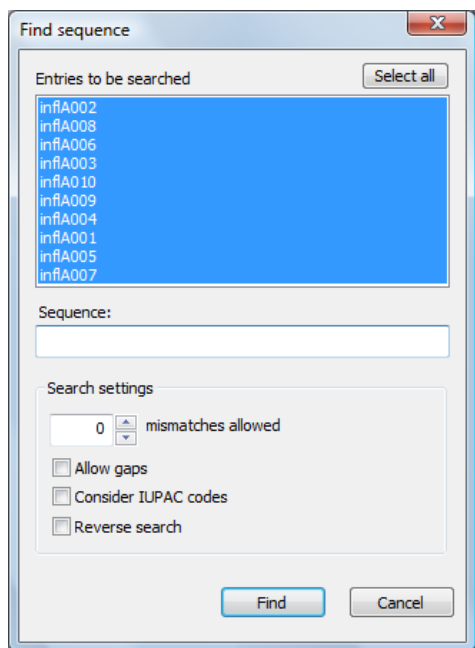


Figure 4-90. The *Find sequence* dialog box.

By default, all entries present in the alignment project are highlighted in the *Entries to be searched* list. Any other selection of entries can be made from this list.

Under *Sequence*, you can enter the subsequence to search for.

The *Search settings* are applicable to the current subsequence search:

- The number of *mismatches allowed* for a subsequence to match with a target sequence.
- *Allow gaps*: Whether or not you allow the subsequence to be interrupted by gaps.
- *Consider IUPAC codes*: Allows the search sequence to be matched with uncertain positions denoted as IUPAC unresolved positions (e.g. "R", "Y", etc., including "N") and allows IUPAC code to be used in the search sequence. When unchecked, only A, T, C or G will be matched against the target sequence(s).
- *Reverse search*: Whether or not the invert-complemented sequence will be searched as well.

4.6.19.3 Enter a sequence string e.g. `ttcttggtcmg`, allow two mismatches and check *Consider IUPAC codes*. Press **<Find>** to start the search.

The search results are displayed in the *Sequence search result* panel (see Figure 4-91). This grid panel lists every match of the entered subsequence, as defined in the search settings.

'Entry' displays the key of the sequence in which the match occurs. 'Position entry' is the position on the individual sequence where the match occurs, while 'Position alignment' uses the position in the alignment. 'Direction' is either forward (a blue arrow pointing to the right) or reverse (a red arrow pointing to the left). The column 'Match' shows the search sequence (top) matched with the target sequence (bottom) and 'Mismatch' displays the number of mismatches occurring. The latter will always be lower than or equal to the number of mismatches specified in the *Find sequence* dialog box (see Figure 4-90).

Sequence search results						
[1] ttcttggtcmg						
Entry	Position entry	Position alignment	Direction	Match	Mismatch	
infIA003	392	393-404	→	TTCTTGGTCMG TTYTTGGTCCG	2	
infIA004	393	393-404	→	TTCTTGGTCMG TTCTTGGTCAG	1	
infIA005	393	393-404	→	TTCTTGGTCMG TTYTTGGTCAG	2	
infIA006	393	393-404	→	TTCTTGGTCMG TTYTTGGTCAG	2	
infIA007	393	393-404	→	TTCTTGGTCMG TTCTTGGTCAG	1	
infIA008	393	393-404	→	TTCTTGGTCMG TTYTTGGTCAG	2	
infIA009	393	393-404	→	TTCTTGGTCMG TTYTTGGTCAG	2	
infIA010	393	393-404	→	TTCTTGGTCMG TTCTTGGTCCA	2	

Figure 4-91. The *Sequence search results* panel from the *Alignment* window.

The sequence search results are listed in the order in which they occur in the alignment. The alignment is screened from top to bottom, left to right. With other words, from the first position on the first, second, third, etc. sequence to the second position on the first, second, third sequence, etc.

4.6.19.4 Click on any match listed in the *Sequence search results* panel.

The cursor will jump to the corresponding sequence block in the alignment (in the *Sequence display 1* panel) and to the corresponding block on the curves (in the *Sequence display 2* panel; if displayed).

4.6.19.5 If more than one sequence search was performed, the results of a previous search can be displayed again by selecting this search from the drop-down list in the toolbar of the *Sequence search results* panel (see Figure 4-92).

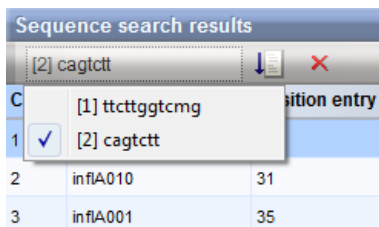



Figure 4-92. Drop-down list in the toolbar of the *Sequence search results* panel, displaying previous sequence searches.

4.6.19.6 To delete a sequence search, select it from the drop-down list and press the  button.

NOTE: Sequence searches are not saved along with the alignment project.


4.6.20 Mutation search

The mutation search tool is designed to detect mutations in individual sequences based on comparison with the consensus derived from the multiple alignment. Therefore, in order to perform a mutation search, a consensus sequence should first be calculated (see 4.6.9). The settings for defining the consensus determine the way mutations are defined. For example, if the *Minimal fraction of a specific nucleotide at a defined consensus position* (see 4.6.9) is set to 10%, all bases at a position that have more than 10% occurrence will contribute to the consensus: if 85% is T and 15% is C, the consensus will then be "Y" at that position. The mutation search algorithm will consider a sequence with a "T" or "C" at that position as NOT mutated. A sequence that has "A" at that position will be recorded as a mutation, since "A" is not contained in the consensus sequence. Using the right settings for the consensus sequence, the mutation search tool can be used for SNP discovery as well.

4.6.20.1 In the alignment project **MyAlignment**, select **HA** from the drop-down list in the main toolbar.

4.6.20.2 In case the sequences are still unaligned, calculate a multiple alignment as described in 4.6.8.

4.6.20.3 If no consensus sequence is defined yet, proceed as described in 4.6.9 to calculate a consensus sequence for the alignment. Use e.g. entry *inflA002* to base the consensus on.

4.6.20.4 In the *Alignment* window, select **Mutations > Search** from the main menu or press the  button in the *Mutation listing* panel to search for mutations. This pops up the *Find mutations* dialog box (see Figure 4-93).

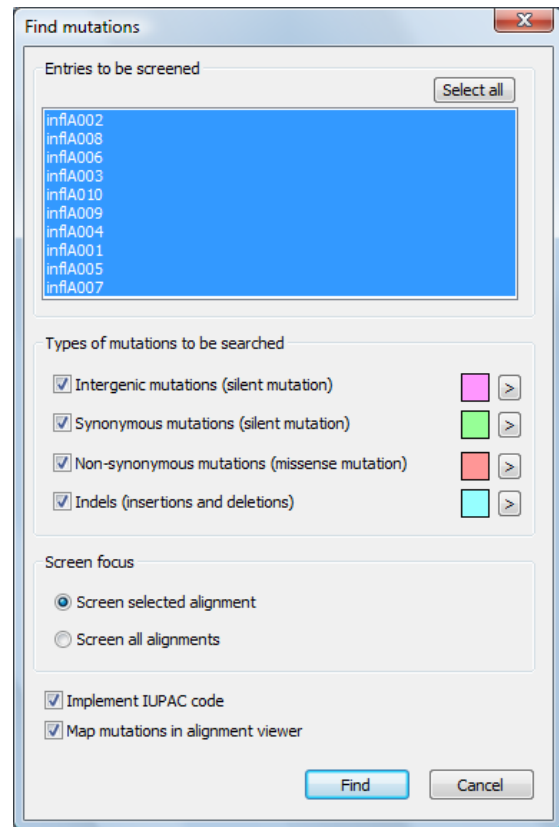


Figure 4-93. The *Find mutations* dialog box.

By default, all entries present in the alignment project are highlighted in the *Entries to be screened* list. Any other selection of entries can be made from this list.

A checklist allows the *Types of mutations to be searched* to be selected. Each type of mutation (intergenic, synonymous, non-synonymous or indel) is displayed in a different color. Clicking the arrow button next to the color box allows you to pick a different color. The *Screen focus* determines whether only the selected alignment is screened or all sequence alignments present in the project.

If **Implement IUPAC code** is unchecked, any ambiguous position will be listed as a mutation. When checked, InfoQuest FP will consider the IUPAC nomenclature and score mutations in a “conservative” way. For example, for a position denoted as “A” in the consensus, any occurrences of “R” (A or G), “M” (C or A) or “W” (T, U or A) will not be scored as a mutation.

If you check **Map mutations in alignment viewer**, the positions of the mutations are indicated in the *Sequence display 1* and *Sequence display 2* panel using blocks of the corresponding color.

4.6.20.5 Leave all settings at their default and press **<Find>** to start the mutation search. The results are displayed in the *Mutation listing* panel (see Figure 4-94).

This grid panel lists all mutations that were found in comparison with the consensus. The column ‘Entry’ shows the key of the entry where the mutation occurs. ‘Alignment’ is the name of the alignment and ‘Position’ the nucleotide position at which the mutation occurs. The type of the mutation (‘Type’) can be silent, missense or indel and the color of the small square is as defined in the *Find mutations* dialog box (see Figure 4-93). ‘NA change’ is the nucleotide change and ‘AA change’ is the change in amino acid (if any). The mutations are listed in the order in which they occur in the alignment. The alignment is screened from top to bottom, left to right. In other words, from the first position on the first, second, third, etc. sequence to the second position on the first, second, third, etc. sequence, and so on.

*NOTE: If you check **Alignment > Translation > None** in the main menu, no amino acid changes will be shown in the Mutation listing panel after a subsequent mutation search (column ‘AA change’ will be empty)*


Entry	Alignment	Position	Type	NA change	AA change
infiA003	HA	9	missense	t → g	D → E
infiA010	HA	9	indel		
infiA004	HA	9	missense	t → g	D → E
infiA001	HA	9	missense	t → g	D → E
infiA007	HA	9	missense	t → g	D → E
infiA010	HA	14	missense	t → c	I → T
infiA003	HA	19	missense	c → t	L → F
infiA010	HA	20	missense	t → c	L → P
infiA008	HA	21	silent	c → t	
infiA006	HA	21	silent	c → t	
infiA003	HA	21	missense	c → t	L → F
infiA010	HA	21	missense	c → t	L → P
infiA009	HA	21	silent	c → t	
infiA004	HA	21	silent	c → t	
infiA001	HA	21	silent	c → t	
infiA005	HA	21	silent	c → t	
infiA007	HA	21	silent	c → t	
infiA003	HA	22	missense	c → t	L → F


Figure 4-94. The *Mutation listing* panel in the *Alignment* window.

and all mutations will be marked as either silent or indel.

4.6.20.6 Click on any of the mutations listed in the *Mutation listing* panel.

The cursor will jump to the corresponding position on the alignment (in the *Sequence display 1* panel) and to the corresponding position on the curves (in the *Sequence display 2* panel; if displayed).

4.6.20.7 To scroll through the mutation list, select **Mutations > Jump to next** or press  in the *Mutation listing* panel.

4.6.20.8 To return to a previous mutation, select **Mutations > Jump to previous** or press  in the *Mutation listing* panel.

If more than one mutation search was performed, a previous listing can be displayed again by selecting the **Mutation search** from the drop-down list in the toolbar of the *Mutation listing* panel (see Figure 4-95).

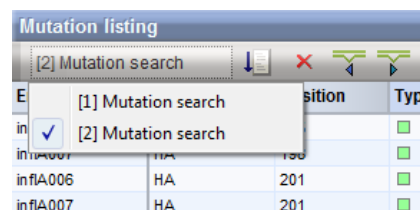



Figure 4-95. Drop-down list in the toolbar of the *Mutation listing* panel, displaying previous mutation searches.


4.6.20.9 To delete a mutation search, select it from the drop-down list and press the  button.

NOTE: Mutation listings are not saved along with the alignment project.


4.6.21 Defining bookmarks in a sequence alignment

A sequence position or sequence block, e.g. corresponding to a primer, probe, protein active site, etc., can be bookmarked in order to retrieve it easily.

To keep bookmarks organized, lists can be created to store related bookmarks in.

4.6.21.1 Select *Alignment > Bookmarks > Add new list* or press the  button in the *Bookmarks* panel. Call the list e.g. "Primer positions".

The drop-down list in the toolbar of the *Bookmarks* panel now displays "Primer positions", but the list does not contain any bookmarks yet.

4.6.21.2 To create a bookmark, select a sequence position or sequence block in the alignment (e.g. the first 20 positions of the first sequence) and select *Alignment > Bookmarks > Add bookmark* or press the  button in the *Bookmarks* panel.


4.6.21.3 The program will prompt for the bookmark name, you can call it e.g. "forward primer".

4.6.21.4 Repeat steps 4.6.21.2 to 4.6.21.3 to bookmark the last 20 positions on the first sequence and call it e.g. "reverse primer".

NOTES:

(1) *Bookmarks can also extend vertically and span multiple sequences. This can be useful, e.g., to define a region on all sequences to print or to create a subset.*

(2) *Bookmarks should not necessarily belong to a bookmark list but can be defined directly. In that event, they are stored under **All bookmarks**.*

4.6.21.5 To delete a bookmark from the list, click on it and select *Alignment > Bookmarks > Delete selected bookmarks* or press the  button.

Bookmark lists different from the one currently displayed can be selected from the drop-down list in the toolbar of the *Bookmarks* panel (see Figure 4-96).

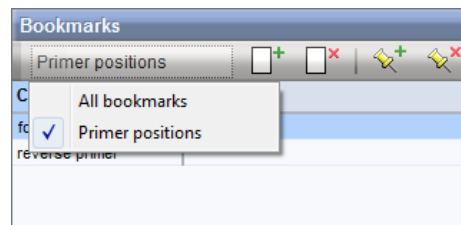



Figure 4-96. Drop-down list in the toolbar of the *Bookmarks* panel, displaying the available bookmark lists.

4.6.21.6 To delete a bookmark list, select it from the drop-down list and press the  button.

NOTE: All bookmarks are saved along with the alignment project.

4.7 Cluster analysis of trend data CL TD

4.7.1 Trend data comparison settings

The comparison settings for a trend data type include the coefficients and clustering methods used for creating dendrograms from trend data.

4.7.1.1 The settings can be changed from the *Trend data type* window, or in the *Comparison* window, by clicking on the trend data type in the *Experiments* panel and selecting *Clustering > Calculate > Cluster analysis (similarity matrix)*.

As a result, the *Comparison settings* dialog box for trend data types pops up (Figure 4-97). For calculation of a similarity coefficient, two basic options can be chosen:

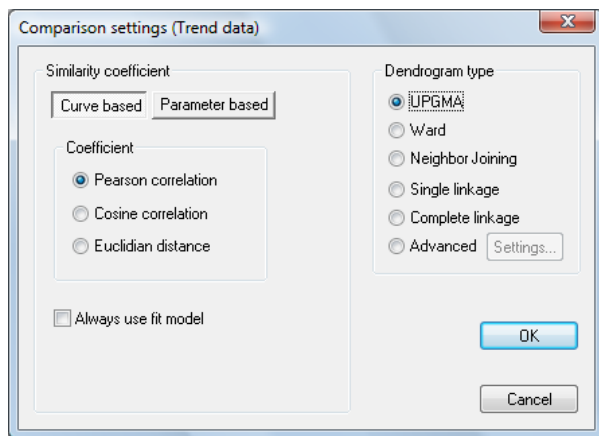


Figure 4-97. *Comparison settings* dialog box for trend data, *<Curve based>* option selected.

<Curve based>: In this option (Figure 4-97), a coefficient is calculated on the curves directly. As an additional option, the raw curve can be used, i.e. the original input data values, or the fit curve as produced from the default trend curve fit model used. For curve based comparison, one can choose *Pearson correlation*, *Cosine correlation*, and *Euclidean distance* as a coefficient.

<Parameter based>: Using this option (Figure 4-98), the similarity or distance is calculated from the parameters defined for the experiment type. The way parameter values are processed into data matrices is illustrated in : each parameter defined leads to a data matrix with the number of characters (values per entry) defined by the number of curves defined for the experiment type. In case two parameters are defined, two data matrices are generated. For each parameter, a separate coefficient can

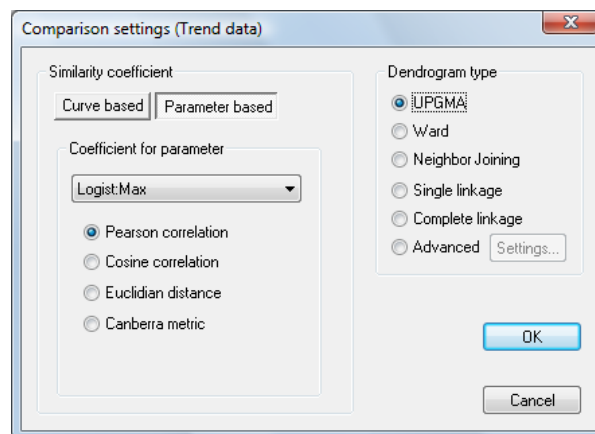


Figure 4-98. *Comparison settings* dialog box for trend data, *<Parameter based>* option selected.

be chosen to analyze the associated data matrix (Figure 4-99). The obtained similarity matrices are averaged and a dendrogram is calculated.

The pull-down listbox in the *Comparison settings* dialog box (Figure 4-98) allows you to specify a coefficient for each parameter separately. Coefficients of choice are *Pearson correlation*, *Cosine correlation*, *Euclidean distance*, and *Canberra metric*. In case Euclidean distance is chosen, a range can be specified. Since Euclidean distance has no inherent scaling, the range specified allows the similarity matrices to be weighted for scaling differences between the parameters used.

The clustering methods are the same as for other experiment types; see 4.1.9.

4.7.2 Display options for trend data

A number of options related to trend data are categorized under the menu *TrendData* in the *Comparison* window.

4.7.2.1 With *TrendData > Show parameter colors*, you can have the values of the parameters displayed as colors, defined in the *Trend data type* window.

4.7.2.2 With *TrendData > Show parameter values*, the values of the parameters are displayed as numerical values.

4.7.2.3 A combination of the two above options is obtained with *TrendData > Show parameter values & colors*.

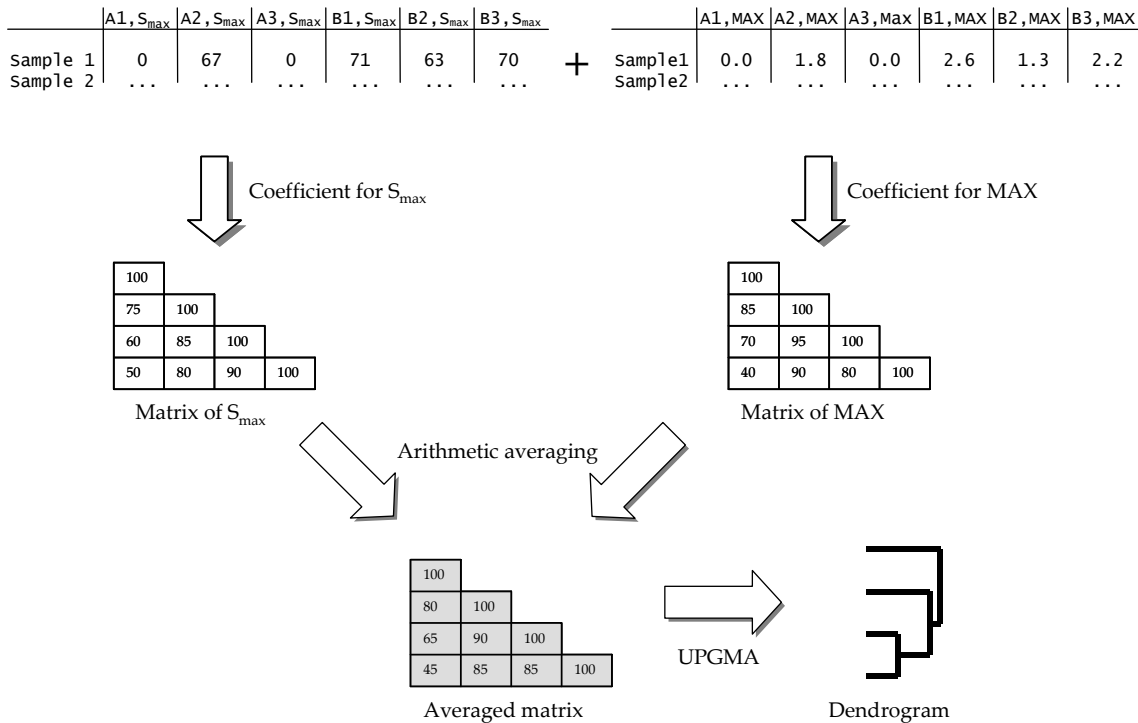


Figure 4-99. Schematic representation of parameter-based cluster analysis of trend data. This example, where two parameters were defined, is a continuation of the processing scheme presented in .

4.7.2.4 With *TrendData > Order by parameter*, the parameter list can be ordered by grouping the same parameters from the different curves together.

4.7.2.5 Alternatively, with *TrendData > Order by curve*, the parameter list is ordered by grouping the different parameters from the same curves together.

4.7.2.6 A *Trend data* window can be created from the entries contained in the comparison with *TrendData > Create trend data window*.

4.7.2.7 For a selected parameter, the entries can be sorted according to increasing value using *TrendData > Sort entries by character*.

NOTE: The separator bar between the parameter names and the values can be dragged down if the names are not completely visible.

4.7.2.8 A tab-delimited text file of the entries and trend data values contained in the current comparison can be exported with *TrendData > Export character table*.

4.8 Cluster analysis of composite data sets CL

4.8.1 Principles

A clustering based upon a similarity matrix can be performed on an individual experiment type or on a combination of experiment types. The methods that InfoQuest FP uses to arrive at dendrograms representing combined techniques are represented schematically in Figure 4-100.

- Flows 1 and 2 represent the steps to obtain dendrograms for two single experiments, experiment 1 and experiment 2, respectively. The steps involve the creation of a similarity matrix and the calculation of a dendrogram based on this matrix.
- Flow 3 is the first method to calculate a combined dendrogram from multiple experiments: the individual similarity matrices are first calculated and from these matrices, a combined matrix (A) is calculated by averaging the values. The averaging can happen in two ways: each value can be considered

equally important, or the program can assign a weight proportional to the number of tests in an experiment. In addition, the user can define an extra weight for each experiment manually.

- Flow 4 starts directly from the character tables, and merges all characters from different experiment types to obtain a *composite data set*. From this composite data set, a similarity matrix is calculated (combined matrix B), resulting in combined dendrogram B.

Both steps 3 and 4 require a *composite data set* to be generated.

4.8.2 Calculating a dendrogram from a composite data set

Calculating a dendrogram from a composite data set is almost the same as for a single experiment. The creation

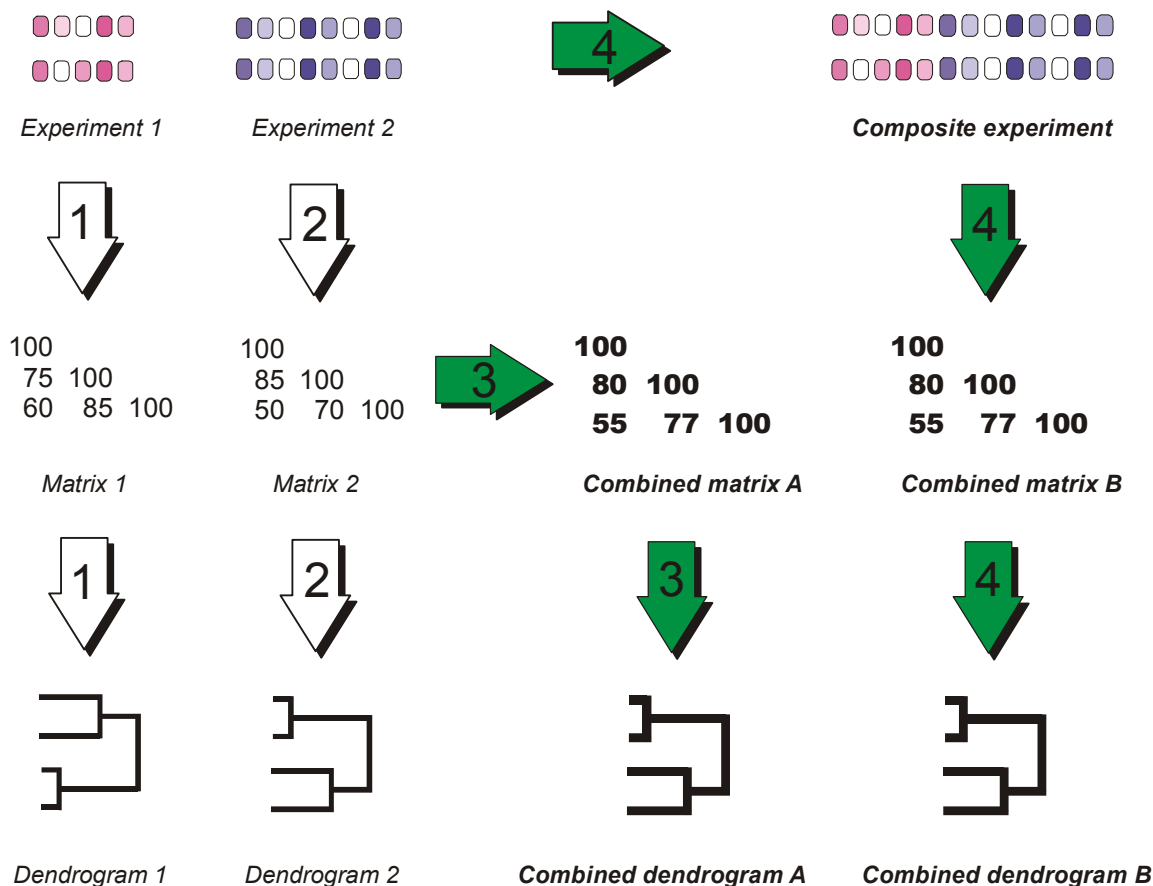



Figure 4-100. Scheme of possibilities in InfoQuest FP to obtain combined dendrograms from multiple experiments.

of a composite data set and its functions is described in , and if you have gone through that paragraph, a composite data set **All-Pheno** should be available in the **DemoBase** database, including the character types **PhenoTest** and **FAME**.

4.8.2.1 In the *InfoQuest FP* main window with **DemoBase** loaded, open comparison **All**, or create a comparison containing all entries except the STANDARDS (see 4.1.9).

4.8.2.2 Select **All-Pheno** in the *Experiments* panel and show the character image by pressing the  button of **All-Pheno**.

NOTE: The order in which experiment data are displayed in a composite data set is the same as the experiment order in the Experiments panel of the InfoQuest FP main window: re-sorting the Experiments panel will result in an updated display order of the experiment data in all composite data sets when the comparison is opened again.

4.8.2.3 Right-click on the image and select *Show quantification (colors)* from the floating menu.

4.8.2.4 Select *Clustering > Calculate > Cluster analysis (similarity matrix)*.

The *Composite data set comparison* dialog box (Figure 4-101) allows you to choose between step 3 (averaging the matrices of the experiments) and step 4 (merging the experiments to a composite experiment) of Figure 4-100. With the *Similarity* option *Average from experiments*, the matrices from the individual experiments are averaged according to the defined weights (step 3 in Figure 4-100). With one of the coefficients under *Binary coefficient*, *Numerical coefficient* (non-binary coefficients), or *Multi-state coefficient*, step 4 in Figure 4-100 will be followed using a composite character table. For a description of the coefficients, see 4.4.1.

If non-binary characters (values) are used, it may be meaningful to enable the feature *Standardized characters* in the following cases. (1) For some techniques, e.g. fatty acid methyl ester analysis, it is common that some fatty acids occur in high amounts, whereas other fatty acids occur only in very small amounts. It is likely that the major fatty acids will account for most of the discrimination between the organisms studied, whereas the minor fatty acids, which may be as valuable from a taxonomic point of view, are masked. (2) When creating composite character sets from different experiments, the ranges of the experiment may be different. When using a coefficient such as the correlation coefficient, characters with a higher range will have more influence on the similarity and the dendrogram. The feature *Standardized characters* standardizes each character by subtracting its mean value and dividing by its standard deviation. The result is that all characters have equal influences on the similarity.

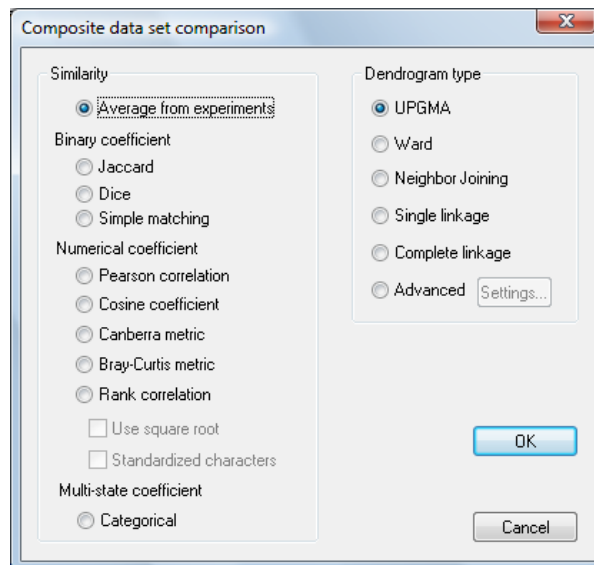


Figure 4-101. Composite data set comparison dialog box.

The feature *Use square root* is intended for character sets that yield high similarities within groups. In such cases, it may be useful to combine *Use square root* with Pearson correlation and Cosine coefficient (or Euclidean distance in case of non-composite data sets).

The *Rank correlation* coefficient first transforms an array of characters into an array of ranks according to the magnitude of the character values. The rank arrays are then compared using the Pearson product-moment correlation coefficient. The Rank correlation is known to be a very robust coefficient, but with low sensitivity.

4.8.2.5 Select *Pearson correlation* with *Standardized characters* and *UPGMA* as clustering method.

If the combined experiments are comparable in terms of biological meaning, reaction type and numerical range, it is possible to use one of the binary coefficients *Jaccard*, *Dice*, *Simple matching*, or one of the numerical coefficients *Pearson correlation*, *Cosine correlation*, or *Canberra metric*. For example, if both experiments involve substrate utilization tests and are recorded either positive or negative, the best option is to select under **Binary coefficient** (Jaccard, Dice or Simple matching). If both character tests are registered quantitatively as numerical values between 0 and 100, a suitable option is to select under **Numerical coefficient** (Pearson correlation, Cosine correlation, Canberra metric, or Rank correlation).

*NOTE: It can be proven that in case of binary data sets the option **Average from experiments** offers exactly the same results when **Correct for internal weights** is enabled in the Composite data set settings.*

In the case however, that the *ranges* of the combined experiments are different, e.g. a range between 0 and 10 for one experiment and between 0 and 100 for another experiment, the numerical coefficients Pearson correla-

tion, Cosine correlation, or Canberra are not suitable, as they would assign much more weight to the second experiment than to the first. In such cases, you should either take the similarity values from the individual experiments (*Average from experiments*) and average them into a new matrix, or specify user-defined weights for the experiments, so that their final weights are comparable.

The *Categorical* coefficient can be chosen in case all the characters of the individual experiment types are *multi-state* characters. As opposed to *binary*, where only two states are known, multistate characters are defined as characters that can take more than two states. However, as opposed to *numerical* characters, the different states represent discrete categories, which cannot be ranked somehow. Examples are phage types, Multilocus Sequence Types (MLST), colors, etc.

4.8.2.6 In the example data sets **PhenoTest** and **FAME**, the character sets have different ranges, so select *Average from experiments*.

4.8.2.7 Press <OK> to calculate the cluster analysis.

The resulting dendrogram is based upon the average matrix of both similarity matrices. In this composite data set, we have chosen the averaging to correct for internal weights, so since the experiment type **FAME** contains more characters than **PhenoTest**, it is assigned more weight proportionally. Hence, we can expect that the composite clustering will have a higher congruence with **FAME** than with **PhenoTest**. You can check this as follows:

4.8.2.8 First make sure that a matrix is present for both **FAME** and **PhenoTest**: select *Calculate cluster analysis* for both experiments.

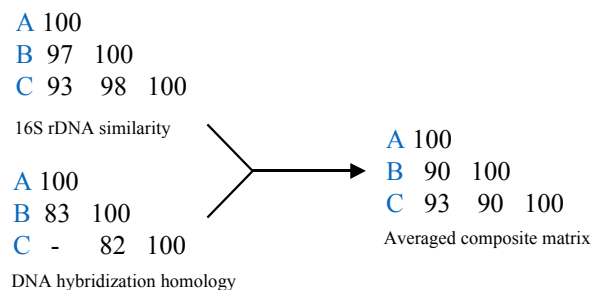
4.8.2.9 Select *Clustering > Congruence of experiments* (see 4.1.18). **All-Pheno** shows 96.4% similarity with **FAME** and only 71.6% with **PhenoTest**.

*NOTE: If you want to see the difference when **Correct for internal weights** is not enabled, save and close the Comparison window, open **All-Pheno** in the Experiments panel, and uncheck **Experiment > Correct for internal weights**. Open the comparison again, **Calculate cluster analysis** again for **All-Pheno**, and **Clustering > Congruence of experiments**. The similarity of **All-Pheno** now is 88.6 and 84.7 with **PhenoTest** and **FAME**, respectively.*

It is obvious that the possibility of approach 4 described in 4.8.2, i.e. merging two character sets into a combined character set, is only applicable to comparable character sets. It makes no sense, and is even impossible to combine a phenotypic test panel with a sequencing experiment in this way. When such experiments of different nature are to be used for consensus groupings, the only remaining approach is to combine the obtained individual similarity matrices (approach 3 in 4.8.1). However, the option to create an *average* matrix from

individual experiment matrices only works well in case two conditions are fulfilled: (i) the expected *similarity range* for both experiments is comparable, and (ii) the matrices are complete, i.e. for each experiment there is a similarity value present for each pair of entries. Suppose that two experiment types are to be combined which generate strongly different similarity levels, e.g. DNA homology values on the one hand and 16S rDNA similarity on the other hand. In many cases, DNA homology values will range from 100% to 40% or less, whereas 16S rDNA similarity will range between 100% and 90% or even higher. It is clear that the small but very significant similarity differences in 16S rDNA homology will be masked by the much larger differences (including experimental error) of DNA hybridization, and will have no contribution to the clustering based upon averaging of matrices. In such cases, other methods are needed to compose a consensus matrix, that "takes the best of it all".

The principle of averaging matrices is even worse when one or more matrices are incomplete. Suppose three entries in InfoQuest FP, A, B, and C. Consider the following matrices for these three entries, generated from 16S rDNA aligned sequence similarity and DNA hybridization. The DNA hybridization matrix is incomplete, a situation which may happen frequently.



The averaged matrix created in the composite data set from these two experiments shows averaged values for (AB) and (BC) but for (AC) it has taken the only available value, 93%. The resulting matrix provides a completely distorted view of the relationships between these three organisms, as it suggests A and C to be closest related. In reality however, one can predict, based upon the lower 16S rDNA similarity, that (AC) will be much less related than (AB) and (BC).

This is an obvious example where averaging similarity matrices is not a good approach, and therefore, another algorithm has been incorporated in InfoQuest FP, based upon linearization of the consensus matrix with respect to the individual experiment matrices. The consensus matrix is composed in such a way that it constitutes a third degree function of each individual experiment matrix, and the result is that it reflects each of the constituent matrices as closely as possible.

The *consensus matrix* can be calculated in InfoQuest FP as follows:

4.8.2.10 If not existing yet, create a new composite data set **All-Exp**, in which you add all experiments available in **DemoBase**.

4.8.2.11 Open a the comparison **All** or create a comparison containing all but the STANDARD lanes, and calculate a matrix (*Calculate cluster analysis*) for each experiment.


4.8.2.12 Select **Composite > Calculate consensus matrix**. The consensus matrix and a corresponding consensus dendrogram is calculated. The resulting groupings can be considered as the most faithful “compromise” from all available data.

NOTE: The feature to correct for internal weights () does not apply to a consensus matrix.

4.8.3 Finding discriminative characters between entries

InfoQuest FP offers the possibility to rearrange the characters in a composite data set according to their discriminatory power. As an example, we use the composite data set **All-Pheno** including **FAME** and **PhenoTest** as described in 4.8.2.

4.8.3.1 In database **DemoBase**, have a *Comparison* window open with all non-“STANDARD” entries selected (e.g. comparison **All**, see 4.1.9) and the composite data set **All-Pheno** shown.

4.8.3.2 Make sure that the image of the composite data set is shown, by pressing the  button of **All-Pheno** in the *Experiments* panel.

4.8.3.3 Minimize or reduce the *Comparison* window so that the *InfoQuest FP main* window (at least the menu and toolbar) becomes visible.

4.8.3.4 Press F4 to make sure that no entries are selected.

4.8.3.5 In the *InfoQuest FP main* window, select **Edit > Search entries** (shortcut F3 on the keyboard), enter *Veringetorix* in the ‘Genus’ field and press <Search>.

All *Veringetorix* entries are selected in the *Database entries* panel of the *InfoQuest FP main* window and in the *Information fields* panel of the *Comparison* window.

4.8.3.6 To group the selected entries, choose **Edit > Bring selected entries to top** in the *Comparison* window or press CTRL+T on the keyboard.

4.8.3.7 Select **Composite > Discriminative characters**.

The characters are reorganized in such a way that those characters positive for the selected entries and negative for the other entries occur left, and those characters negative for the selected entries and positive for the other entries occur right.

Similar as for character types, in composite data sets it is possible to list the entries according to the value of a selected character.

4.8.3.8 Show the composite data set **All-Pheno** as intensity table with **Composite > Show quantification (colors)**.

4.8.3.9 Click on a character of **All-Pheno** in the character header of the *Experiment data* panel (e.g. ‘FAME:16:0’) and select **Composite > Sort by character**.

The entries are now sorted by increasing intensity of the selected character.

4.8.4 Transversal clustering

The input for a cluster analysis in a composite data set is a *data matrix*. A data matrix of n entries having p characters looks like in Figure 4-102: the entries are presented as *rows* and the characters as *columns*. In InfoQuest FP, the data matrix should not necessarily be complete: some missing character values are allowed, for example if test results are ambiguous or not available.

	Char 1	Char 2	...	Char p
Entry 1	Val 11	Val 12	...	Val 1 p
Entry 2	Val 21	Val 22	...	Val 2 p
...
Entry n	Val $n1$	Val $n2$...	Val np

Figure 4-102. Data matrix of n entries and p characters.

A simple and efficient way to visualize associated groups of characters (columns) with groups of entries (rows) in a data matrix is to construct a two-way clustering of the data matrix, i.e. in which the entries are clustered by means of their character values (the conventional clustering as described in 4.1.8; also called *Q-clustering*), and the characters are clustered by means of their values per entry (*R-clustering*).

The result is a data matrix in which both the entries and the characters are ordered according to their relatedness (Figure 4-103), which we will call *transversal clustering*. This representation makes it easy to visually associate clusters of characters with clusters of entries. For example, the first group of entries (E1, E9, and E5) is separated from the others by a cluster of characters (C5, C16, C3, and C11) which are all more positive in the first cluster than in the other clusters. Another group of three characters (C14, C17, and C20) separates the second

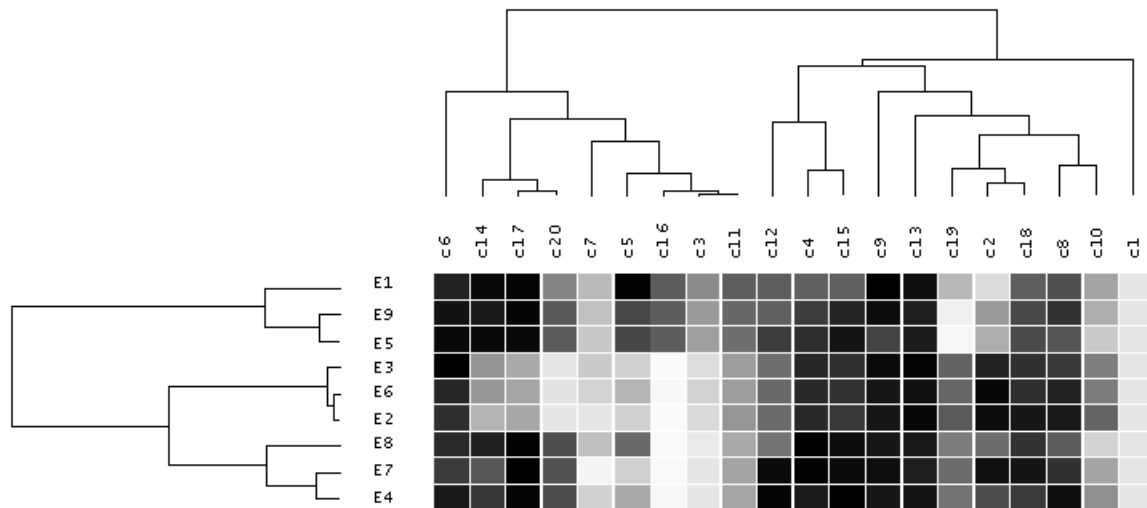




Figure 4-103. Transversal clustering of entries (horizontal) and characters (vertical).

group of entries (E3, E6, and E2) from the other clusters because they are less positive.

In InfoQuest FP, it is possible to calculate a transversal clustering from a composite data set. As an example, we use the composite data set **All-Pheno** in **DemoBase** including **FAME** and **PhenoTest** as described in 4.8.2.

4.8.4.1 Create a *Comparison* window with a selection of entries and select the composite data set in the *Experiments* panel. You can show the character image by pressing the  button of **All-Pheno**.

4.8.4.2 Calculate a cluster analysis of the entries as described in 4.1.9.

4.8.4.3 Choose *Composite* > *Calculate clustering of characters* or click the  button. A dialog box offers a choice between the *Pearson correlation* for numerical characters, the *Jaccard*, *Dice*, and *Simple matching* coefficients for binary data, and the *Categorical* coefficient for multi-state or categorical characters. For a description of the coefficients, see 4.4.1.

4.8.4.4 Select *Pearson correlation* and press <OK> to calculate a character dendrogram, which appears horizontally in the caption of the data matrix display of the composite data set.

4.8.4.5 It may be useful to drag the separator bar between the image panel and its caption down to obtain more space for the character dendrogram and the character names.

4.9 Phylogenetic clustering methods CL

4.9.1 Introduction

In addition to the *Neighbor Joining* method, which we described previously (4.1.11.10 and 4.5.11.2), InfoQuest FP offers two alternative phylogenetic clustering methods, based on the concepts of *maximum parsimony* and *maximum likelihood*, respectively. Maximum parsimony can be applied to any data set that can be presented as a *binary* or *categorical data matrix*. As such it can be applied to fingerprint type data on condition that a band matching is performed (see Section 4.9). It also can be applied to character type data with binary or categorical character data. In case of non-binary numerical data, the default *Binary conversion settings* for the character type will be used (). Likewise, the maximum parsimony method can also be used for composite data sets (see Section 4.8). The maximum likelihood clustering method can only be applied to nucleic acid sequence data, and we will describe them with the sequence type **16S rDNA** of database **DemoBase**. In case of sequence data, the maximum parsimony and maximum likelihood clustering methods only work on aligned sequences: a multiple alignment must be present (see 4.5.3).

4.9.2 Maximum parsimony of fingerprint and character type data


Since maximum parsimony requires a binary or categorical data matrix as input, it can only be applied to fingerprint type data for which a band matching is performed. For the fingerprint type you want to cluster using maximum parsimony, a band matching should be performed as described in Section 4.9. The program will use the *binary* band presence table associated with the band matching as input for maximum parsimony.

In case of character type data, the maximum parsimony can be calculated directly on the data set. If the data set is non-categorical and non-binary, the default *Binary conversion settings* for the character type will be used. You can check this setting by opening the *Character type* window and selecting *Settings > Binary conversion settings*, which pops up the *Conversion to binary data* dialog box ().

4.9.2.1 In the *InfoQuest FP main* window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

4.9.2.2 Click on a character type (e.g. **FAME**) in the *Experiments* panel.

4.9.2.3 To calculate a maximum parsimony dendrogram, select *Clustering > Calculate > Maximum parsimony tree (evolutionary modelling)*. Alternatively, you can

press the  button, in which case the floating menu as shown in Figure 4-4 pops up. Select *Calculate maximum parsimony tree* from the floating menu.

The *Maximum parsimony cluster analysis* dialog box for character data appears (Figure 4-104).

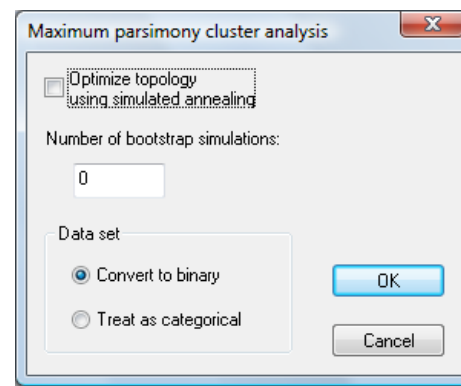


Figure 4-104. The *Maximum parsimony cluster analysis* dialog box for character type data.

4.9.2.4 Under *Data set*, you can specify how to treat the data, i.e. *Convert to binary* or *Treat as categorical*. In case of fingerprint type data and binary character sets, these options are redundant.

4.9.2.5 InfoQuest FP uses methods that are described in the literature to optimize the topology of parsimonious trees. An alternative method, which sometimes finds even more parsimonious trees, but which is considerably slower, is the mathematical principle of *Simulated annealing*.

4.9.2.6 In addition, InfoQuest FP can do a *Bootstrap* analysis on the parsimony clustering, for which you can enter the *Number of bootstrap simulations*. If zero is entered, no bootstrap values are calculated.


Caution: enabling simulated annealing and at the same time entering a number of bootstrap simulations will increase the computing time dramatically. We do not recommend to combine these options.


The resulting *Unrooted dendrogram* window (Figure 4-106) is discussed in the next paragraph.


4.9.3 Maximum parsimony clustering of sequence data

4.9.3.1 In the *InfoQuest FP* main window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

If a dendrogram and a sequence alignment are already present for the sequence type **16S rDNA**, you can proceed with 4.9.3.5. Otherwise, a sequence alignment can be created as follows:

4.9.3.2 Select **16S rDNA** in the *Experiments* panel and press the  button (or select *Layout > Show image* from the menu).


4.9.3.3 The similarity matrix is calculated with *Clustering > Calculate > Cluster analysis (similarity matrix)* or the  button. Leave all settings default and press **<OK>**.

4.9.3.4 Select *Sequence > Multiple alignment* or . Leave all settings default and press **<OK>**.

First, we will reduce the number of entries in the comparison, in order to make maximum likelihood (4.9.4) possible in a reasonable time.

4.9.3.5 Select all entries in the comparison, and then unselect a couple of entries per cluster, so that some 10 entries from all clusters are unselected in total.

4.9.3.6 Remove the selected entries from the comparison with *Edit > Cut selection*.

4.9.3.7 Select *Clustering > Calculate > Maximum parsimony tree (evolutionary modelling)*. You can also press the  button, in which case the floating menu as shown in Figure 4-4 pops up.

From the floating menu you can select *Calculate maximum parsimony tree (evolutionary modelling)*. Note that, in case of aligned sequence data, an extra option *Calculate maximum likelihood tree* becomes available, which is discussed in 4.9.4.

The *Maximum parsimony clustering* dialog box (Figure 4-105) allows you to specify a cost for each base conversion (mutation) in the *Cost table*. The default settings is 100% for each possible conversion.

Gaps can be dealt with in two ways: the program can *Ignore positions with gaps*, or can consider gaps as an *Extra state*. In the first case, when gaps are ignored, every position that contains a gap in one or more sequences of the multiple alignment, will be excluded from the analysis. For very diverse sequences, this may result in the omission of a considerable part of the sequence from the similarity calculation.

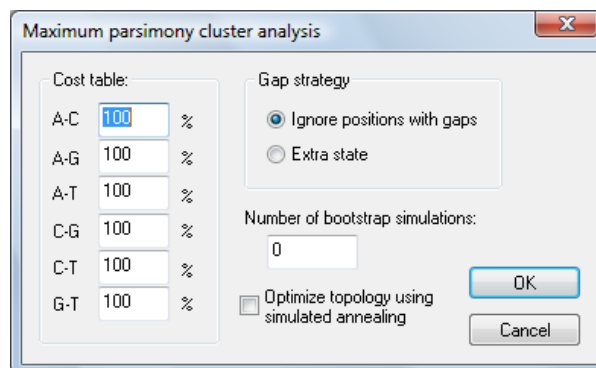






Figure 4-105. *Maximum parsimony clustering* dialog box.

4.9.3.8 Leave the *Cost table* unaltered, check *Ignore positions with gaps*, enable *Optimize topology*, and leave the *Number of bootstrap simulations* zero.


4.9.3.9 Press **<OK>** to start the calculations.


The result is an *Unrooted dendrogram* window, of which the parsimony (the total number of base conversions over the three) is given in the status bar (Figure 4-106). If groups were defined previously (see 4.1.12), the entries are represented in the group colors.

4.9.3.10 To zoom in or out on the tree, use the  and  buttons or *Layout > Zoom in* and *Layout > Zoom out*.

4.9.3.11 You can toggle between the colors and the black-and-white representation mode with *Layout > Show group colors* or . When the group colors are shown, this button is displayed as .

In black-and-white mode, the groups are represented (and printed) as symbols.

4.9.3.12 The drop-down list  allows you to select a coloring based on groups or any of the available field states.

4.9.3.13 With *Layout > Show keys or group numbers* or , the entry keys are displayed next to the dendrogram entries.

However, the entry keys may be long and uninformative for the user, so the entry keys can be replaced by a group code. The program assigns a letter to each defined group, and within a group, each entry receives a number. The group codes are shown as follows:

4.9.3.14 In the parent *Comparison* window, select *Layout > Use group numbers as key*.

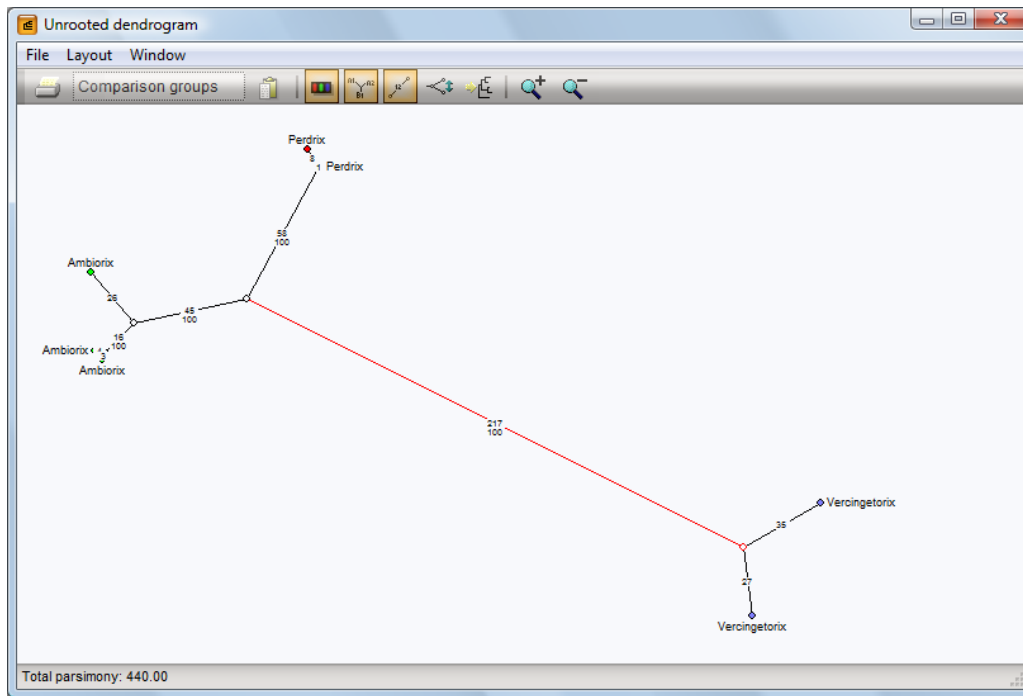



Figure 4-106. Unrooted maximum parsimony tree. Number of mutations are indicated on the branches (top) as well as the bootstrap values (bottom).

4.9.3.15 A legend to the group numbers can be obtained with *File > Export database fields* in the parent *Comparison* window.


Alternatively, a selected information field can be displayed instead of the key:

4.9.3.16 In the parent *Comparison* window, click on the information field which you would like to display as key (e.g. 'Genus').

4.9.3.17 Select *Layout > Use field as key* from the menu in the *Comparison* window. The 'Genus' field is now displayed in the maximum parsimony tree (see Figure 4-106).

4.9.3.18 With *Layout > Show branch lengths* or , the lengths of the branches, as numbers of base conversions, are shown. In case bootstrap values were calculated, this option displays the bootstrap values as well.

In more complex trees, the spread of the branches may not be optimal. The program can iteratively optimize the spread of the branches:

4.9.3.19 In the *Unrooted dendrogram* window, select *Layout > Optimize branch spread* or .

The user can rotate and swap the branches manually if the tree layout is not satisfactory.


4.9.3.20 Left-click in the proximity of a node or a branch tip.


4.9.3.21 While holding down the mouse button, rotate the branch to the desired position.

4.9.3.22 If you select entries in the parent *Comparison* window or in the *InfoQuest FP main* window, these entries are shown within a square in the *Unrooted dendrogram* window.

4.9.3.23 You can also select entries directly in the *Unrooted dendrogram* window, by holding the CTRL key while clicking in the proximity of a node. All entries branching off from this node will be selected.


4.9.3.24 Repeat this action to unselect entries.

4.9.3.25 To copy the unrooted tree to the clipboard, select *File > Copy image to clipboard* or .

4.9.3.26 The unrooted tree can be printed with *File > Print image* or .

Since interpreting unrooted trees is not always easy, especially with large numbers of entries, it is possible to create a rooted dendrogram from the unrooted tree. This process requires an artificial root to be defined as follows:

4.9.3.27 Select a branch by clicking in the proximity of one of the two nodes it connects. The selected branch is red.

4.9.3.28 In the menu, choose *Layout > Create rooted tree* or .

The dendrogram in the parent *Comparison* window now is a rooted version of the maximum parsimony or maximum likelihood tree.

NOTE: All dendrogram display functions (see 4.1.11) also apply to unrooted trees, except the incremental clustering: one cannot delete or add entries while the tree is automatically updated.

In publications and presentations, particularly in a phylogenetic context, a dendrogram is sometimes represented as a real tree with a stem and branches. Such representations can be achieved from a parsimony tree using the *rendered tree* option. This option should be used with care, as it will only produce acceptable pictures from a limited number of entries and with fairly equidistant members.


4.9.3.29 If you want to create a *rooted* rendered tree from the parsimony tree, you first have to select the branch on the tree from which the root will be constructed. Usually, the longest branch on the tree is taken as root.

4.9.3.30 Create a rendered tree from the *Unrooted parsimony tree* window using *File > Export rendered tree*.

The functions of the *Rendered tree* window are described in 4.1.17.

4.9.4 Maximum likelihood clustering

A maximum likelihood cluster will be calculated for the **16S rDNA** sequence type of the comparison as created in the previous paragraph (see 4.9.3.1 to 4.9.3.6).

4.9.4.1 In the *Comparison* window, select *Clustering > Calculate > Maximum likelihood tree (evolutionary modelling)*. You can also press the  button, in which case the floating menu as shown in Figure 4-4 pops up. From the floating menu you can select *Calculate maximum likelihood tree (evolutionary modelling)*.

The *Maximum likelihood clustering* dialog box shows up (Figure 4-107).

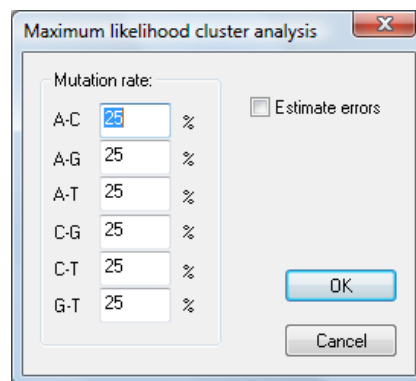


Figure 4-107. Maximum likelihood clustering dialog box.

Similar as for the maximum parsimony clustering, a *Mutation rate* can be defined for each individual base conversion. The default of the program is 25% for each possible mutation. The maximum likelihood clustering algorithm also allows a standard deviation to be calculated for each branch (*Estimate errors*). This is only an approximate error estimation, but since maximum likelihood clustering is exceptionally slow, it is absolutely impossible to perform bootstrap analysis.

4.9.4.2 Leave the *Mutation rate* unaltered, and enable the *Estimate errors* checkbox.

4.9.4.3 Press **<OK>** to start the calculations.

Maximum likelihood clustering is an extremely time-consuming process; depending on the length of the sequences, clustering 30-50 entries may take several hours on a powerful computer. The calculation time increases with the third power of the number of entries included.

When the calculations are finished, an unrooted tree is shown which has all the same functions as described for maximum parsimony (4.9.3). Rendered trees (4.1.17) can also be constructed from maximum likelihood trees.

4.9.4.4 If you want to see the estimated errors on the branch lengths, use *Layout > Show branch lengths* or



4.10 Advanced clustering and consensus trees CL

4.10.1 Introduction

Cluster analysis is one of the most popular ways of revealing and visualizing hierarchical structure in complex data sets. As explained before (4.1.8), cluster analysis is a collective noun for a variety of algorithms that have the common feature of visualizing the hierarchical relatedness between samples by grouping them in a dendrogram or tree. The most universally applied methods are pairwise clustering algorithms that use a distance or similarity matrix as input (Figure 4-108). UPGMA (Unweighted Pair Group Method using Arithmetic Averages), Complete Linkage, Single Linkage, and Ward's method are examples of such methods. The advantage of these methods is that they can be applied to any type of data, as long as there exists a suitable similarity or distance coefficient that can generate a similarity (distance) matrix from the data. As such, similarity-based clustering can be applied to incomplete data sets or data that is not presented in the form of a data matrix (e.g., electrophoresis band sizes).

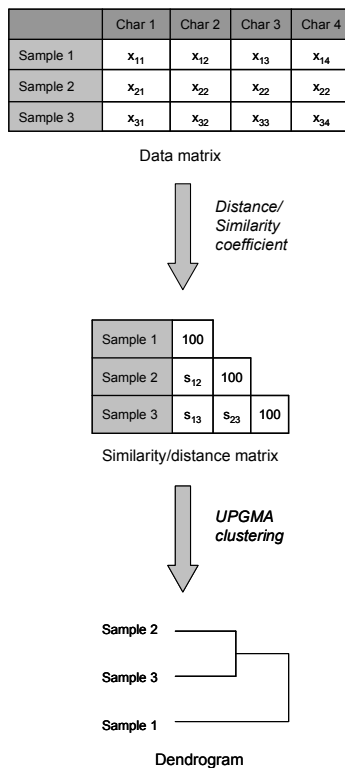


Figure 4-108. Steps in similarity based cluster analysis.

In the analysis steps outlined in Figure 4-108, one should consider the matrix of pairwise similarities (or distances) as the complete comparative information between all the samples analyzed. Obviously, for larger numbers of samples, interpreting a similarity matrix becomes hardly simpler than looking at the original data. This is why a similarity matrix is not usually calculated as a final result, but as an intermediate step for grouping algorithms such as cluster analysis or multi-dimensional scaling.

The real simplification of the data is obtained by cluster analysis. Both the power and the weakness of a dendrogram lie in its ability to present an easy to interpret, well-structured, hierarchical grouping of the samples. Indeed, simplification means loss of information, and there is no way to present the data in a simple and easily interpretable way, yet holding all the information. As a consequence, every dendrogram resulting from a non-artificial data set will contain errors, the amount of error being proportional to the complexity of the similarity matrix. A second source of error results from the fact that hierarchical clustering always imposes hierarchical structure, even if the data does not support it. The fact that even a perfectly random data set results in a dendrogram with branches, is a clear example of the danger that hierarchical clustering holds. Various statistical methods allow the error associated with dendrogram branches or their uncertainty to be estimated, e.g., standard deviation values and the cophenetic correlation (see 4.1.13). Other methods, such as bootstrap, allow the probability of dendrogram branches, as a result of the data set, to be indicated.

4.10.2 Degeneracy of dendrograms

Another problem with pairwise hierarchical clustering methods such as UPGMA is the degeneracy of the solution. Whereas UPGMA results in just one tree, in many cases there exist a number of equally good alternative solutions. Such degeneracies are very likely to occur in cases where the similarity matrix contains multiple identical values. In practice, binary and categorical data sets and banding patterns treated as absent/present states result in frequent occurrence of identical similarity values, whereas quantitative measurements registered as decimal numbers almost never yield identical similarity values. To understand how the occurrence of identical similarity values can result in multiple possible trees, we consider the example of three banding patterns (Figure 4-109). As can be seen from this simple example, $s[A,B]$ and $s[B,C]$ are both 0.75, whereas $s[A,C]$ is 0.50. The way how UPGMA constructs a dendrogram is by first searching for the highest similarity value in the

matrix, and linking the two samples from which it results. In the present example, [A,B] and [B,C] are equivalent solutions, two partial dendrograms can be constructed: one with [A,B] linked at 75% (solution 1) and the other with [B,C] linked at 75% (solution 2). In the next step of UPGMA, the remaining sample is linked at the average of its similarity with the samples already grouped. In solution 1, this leads to C being linked at 62.5% to [A,B], whereas in solution 2, A is being linked at 62.5% to [B,C]. Both dendrograms suggest a quite different hierarchical relatedness but actually none of them truly reflects the relationships suggested by the data set and the similarity matrix.

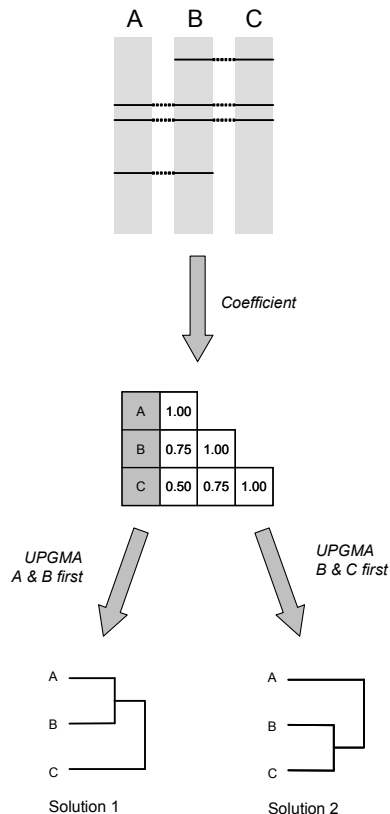


Figure 4-109. A scenario of three banding patterns resulting in two possible UPGMA solutions.

Another inconsistency in pairwise clustering results from the inability to deal with infringements upon the transitivity rule of identity. When sample A is identical to sample B, and sample B is identical to sample C, the transitivity rule predicts that A will be identical to C as well. Infringements upon this rule are particularly found in the comparison of banding patterns, where the identity of bands is judged based upon their distance, using a position tolerance value that specifies a maximum distance between bands to be considered identical. The example below (Figure 4-110) illustrates the result of a UPGMA clustering of three banding patterns for which one band is slightly shifted. With a position tolerance as indicated on the figure, the pairs of patterns [A,B] and [A,C] will have a 100% score, whereas [A,C] will have only 75% similarity as the

distance between their lower bands is greater than the position tolerance specified. Similarly as explained above, the UPGMA algorithm has two choices to perform the first linkage, and the results are displayed as solution 1 and solution 2. Neither of the two dendrograms reflects the discrepancy indicated by the similarity values, but instead, each dendrogram falsely suggests a hierarchical structure that is not supported by the data.

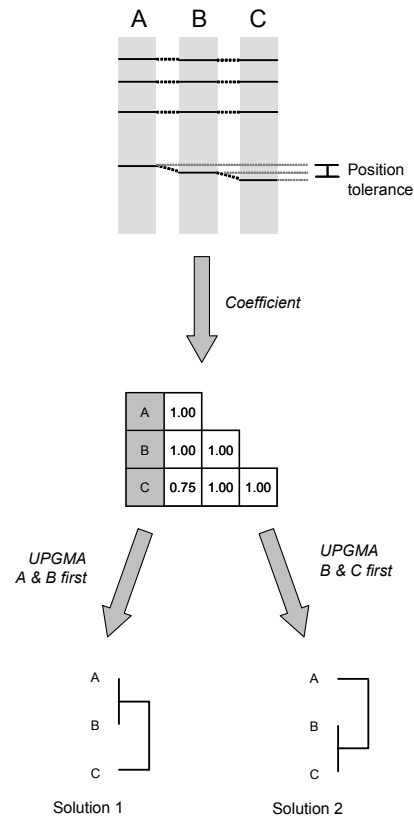


Figure 4-110. Infringement upon the transitivity rule for sample identity and resulting dendrograms.

4.10.3 Consensus trees

A more truthful representation of the relationships given in Figure 4-109 and Figure 4-110 can only be obtained by respecting the indeterminacy resulting from the identical similarity values. Using the conventional pairwise linkage dendrogram representation, this cannot be achieved, and therefore, a new dendrogram type has been introduced in InfoQuest FP, allowing more than two entries or branches to be linked together. The resulting tree can be called a *consensus tree* because it allows all entries that are part of a degeneracy to be linked at one similarity level in a single consensus branch (Figure 4-111). To obtain such a consensus representation of the different trees possible, InfoQuest FP will first calculate all possible solutions and draw a consensus tree that uses pairwise linkage as the primary

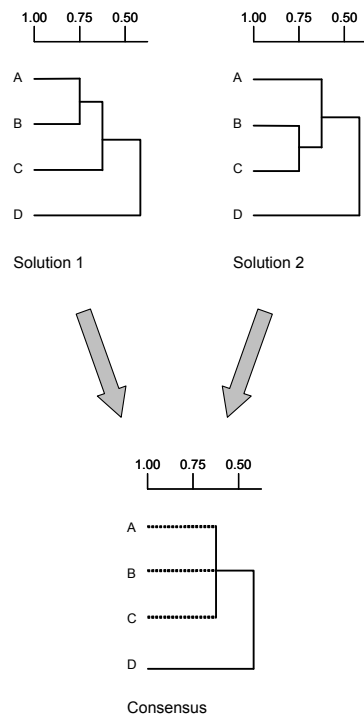


Figure 4-111. Displaying different UPGMA solutions as a consensus branch.

criterion, but applies multi linkage in those cases where branches or entries are degenerated.

Another advantage of the presentation method that supports multi linkage of entries or branches is that it can be used to calculate consensus trees from trees generated from different data sets as well. The same algorithms can be applied to compare the different and common branches on the trees, and the example shown in Figure 4-111 could as well be a case where Solution 1 and Solution 2 result from different data sets.

4.10.4 Advanced clustering tools

The advanced clustering tools in InfoQuest FP offer some additional functionality compared to the standard clustering tools in the *Comparison* window. This functionality is related to the possibility of linking more than two entries or branches together, as shown in Figure 4-111. As such it becomes possible to display multiple solutions of a cluster analysis in a consensus representation, as well as representing two trees from different data sets in one consensus tree. In addition, each tree obtained using the advanced clustering tools is automatically saved, which makes it possible to have more than one stored tree per experiment type. This feature is useful if one wants to compare trees generated using different similarity coefficients or using varying parameters such as position tolerance for banding patterns.

4.10.5 Displaying the degeneracy of a tree

In InfoQuest FP, select a data type that can potentially result in multiple tree solutions, for example, the fingerprint type **RFLP1** in **DemoBase**.

4.10.5.1 In the *InfoQuest FP main* window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

4.10.5.2 Select the fingerprint type **RFLP1** from the *Experiments* panel in the *Comparison* window and choose **Clustering > Calculate > Cluster analysis (similarity matrix)**. This pops up the *Comparison settings* dialog box (Figure 4-29), which shows five clustering options (UPGMA, Ward, Neighbor Joining, Single Linkage and Complete Linkage) and an option **Advanced**.

4.10.5.3 If the option **Advanced** is checked, a button **<Settings>** becomes available, which will open the *Advanced cluster analysis* dialog box (Figure 4-112).

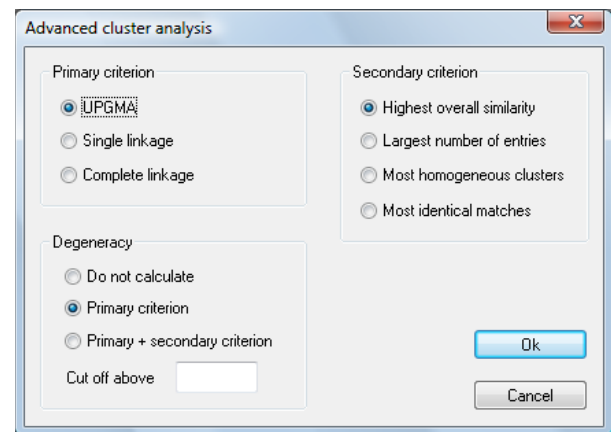


Figure 4-112. The *Advanced cluster analysis* dialog box.

Under *Primary criterion*, the criterion for clustering can be chosen, which can be **UPGMA**, **Single Linkage** or **Complete Linkage**. All three methods are pairwise clustering algorithms, i.e. which will construct dendrograms by grouping branches and/or entries pair by pair, using the highest similarity as criterion. In **UPGMA** the similarity between clusters is calculated as the average of all individual similarities between the clusters, whereas in **Single Linkage** it is the highest similarity found between the clusters. In **Complete Linkage**, it is the lowest similarity found between the clusters.

The *Secondary criterion* applies to those cases where two clusters have the same (highest) similarity with a third, in which case two different tree solutions exist. The program will then apply one of the following criteria to solve the indeterminacy left by the standard clustering algorithm (i.e., the primary criterion):

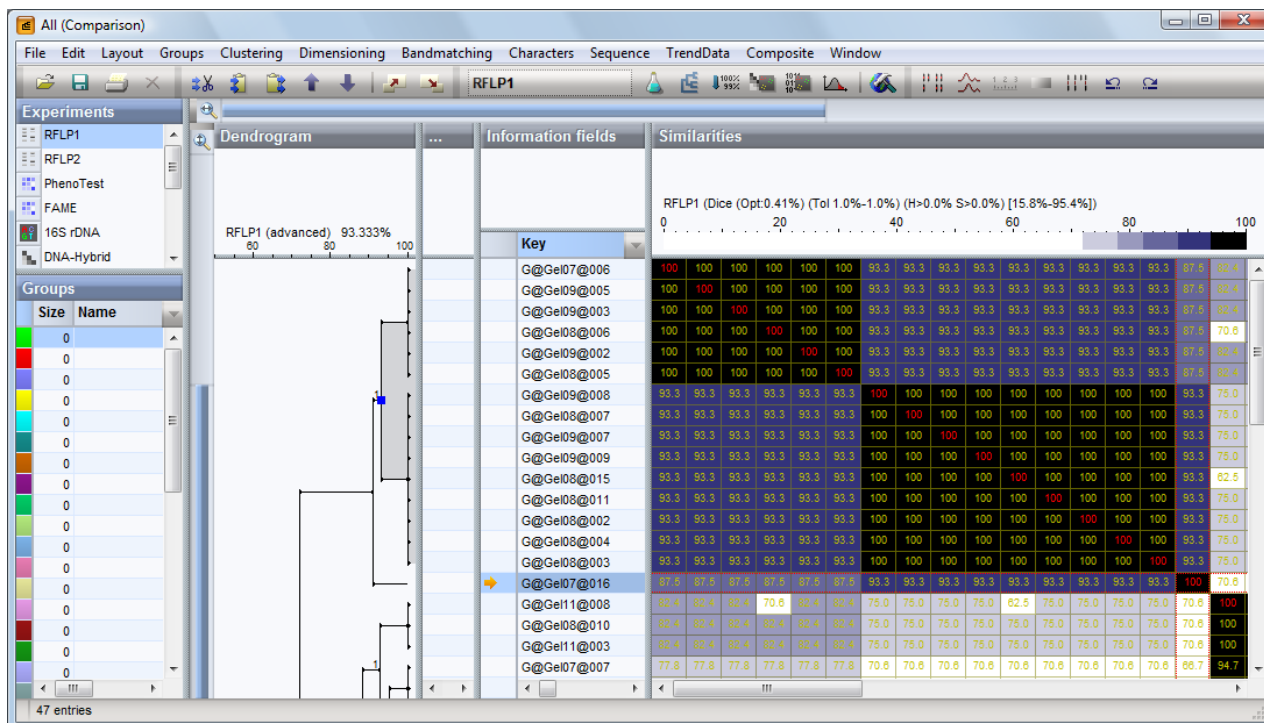


Figure 4-113. Advanced tree representation with a highlighted cluster, indication of the number of degenerated entries relative to the cluster, and the degenerated entry selected.

(1) *Highest overall similarity*: the two clusters will be joined that result in the cluster with the highest overall similarity with all other members of the comparison.

(2) *Largest number of entries*: the two clusters will be joined that result in the cluster with the largest number of entries.

(3) *Most homogeneous clusters*: the two clusters will be joined that result in a cluster that has the highest internal homogeneity.

Note that criteria (1) and (3) are complementary to each other as (1) will only consider the external similarity values of the resulting clusters whereas (3) will only consider their internal similarity values.

Under *Degeneracy*, three options allow one to deal with degenerated trees:

(1) *Do not calculate* will not look for degeneracies and will display just one solution. The differences with a conventional cluster analysis are that (i) the solution presented is the best according to the secondary criterion specified, and (ii) the resulting tree is saved automatically as an *advanced tree* and can be used together with other advanced trees to calculate a *Consensus Tree*.

(2) The option *Primary criterion* will calculate all degeneracies resulting from the primary criterion only and will not consider any secondary criterion specified.

(3) *Primary + secondary criterion* will use the specified secondary criterion to solve the degeneracies resulting

from the primary criterion and will only display the degeneracies that remain after the secondary criterion. It is very unlikely that there will remain any degeneracies with this option checked.

The *Cut off above* parameter specifies the maximum allowed number of degenerate entries relative to a cluster. A *degenerate entry* is an entry that does not belong to a given cluster in the present tree, but that does belong to the cluster in at least one alternative solution. If zero is entered as cutoff value, no degenerate entries are allowed and as a consequence, a consensus tree is generated that includes all possible solutions. If the field is left blank, the degeneracy of the tree will not be reduced at all. If a number is entered, for example 2, all clusters for which there are more than 2 degenerate entries will be displayed as consensus clusters with the degenerate entries included.

Each cluster that has degenerate entries relative to it, will have an indication of the number of degenerate entries (see Figure 4-113, which shows one degenerated entry for the selected cluster).

4.10.5.4 When a cluster is selected by clicking on its branching node, the cluster is filled in gray (Figure 4-113), which makes it easier to see which entries belong to it.

4.10.5.5 If there are degenerated entries relative to the highlighted cluster, you can find them by choosing *Clustering > Advanced trees > Select degenerate entries*. All degenerate entries relative to the cluster are now added to the selection.

The interpretation of degeneracies and tracking back their reason is sometimes difficult. The larger the tree and the deeper the branch, the more complex the degeneracies will be. The example screen in Figure 4-113 is a capture taken from experiment **RFLP1** in the **DemoBase**. The highlighted cluster has one degenerated entry, which is selected. The cluster consists of two subclusters which have an overall average similarity of 93.3%. The single degenerate entry, however, also has an average similarity of 93.3% with the second subcluster. The present solution has first linked subcluster 1 to subcluster 2 and then linked the single entry to the merged cluster. According to the criterion of UPGMA, however, an equivalent solution would be to first link the single entry to subcluster 2 and then link subcluster 1 to this new cluster. When the same clustering is done with zero as cutoff value, the cluster looks like in Figure 4-114. Note that the three subclusters are now linked together at the same level. The clusters that connect always at the displayed similarity level in the solution obtained using the secondary criterion are represented by solid lines (in the present case, the single entry), whereas subclusters that cluster at higher levels using the secondary criterion are connected by an interrupted line.

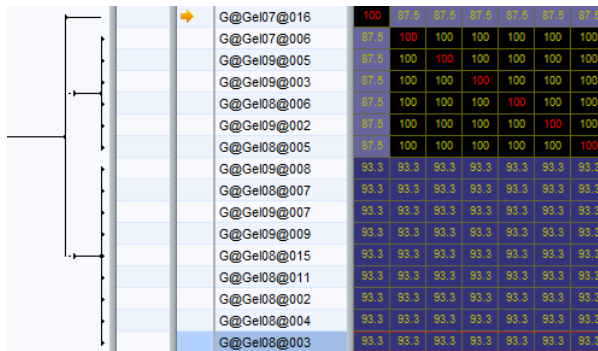


Figure 4-114. Detail of cluster highlighted in Figure 4-113, calculated with a cut off value of zero.

4.10.6 Creating consensus trees

The advanced clustering tool allows a *consensus tree* to be calculated from two or more individual dendrograms. These trees can be conventional clusterings or advanced trees, and can be generated from the same experiment type or from different experiment types. In case you want to calculate different dendrograms from the same experiment type, you should use the Advanced Clustering tools. To create a consensus tree, the program will look for all branches that hold exactly the same entries in both trees and represent them as branches in the consensus tree.

4.10.6.1 As an example, we can calculate two dendrograms in **DemoBase**: one from experiment **PhenoTest** using Pearson correlation and the other from experi-

ment **16S rDNA**. You can calculate the trees using the conventional clustering tools or using the Advanced Clustering tools.

4.10.6.2 Select **Clustering > Advanced trees > Create consensus tree**, which pops up a dialog box listing the *Stored trees* (Figure 4-115).

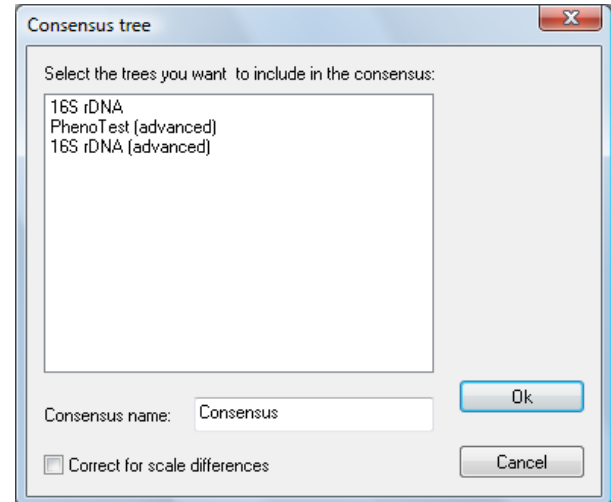


Figure 4-115. *Stored trees* dialog box to calculate a consensus tree.

4.10.6.3 Select the two calculated trees, which have the name of the experiment types they were derived from, and enter a name for the consensus tree to be generated (the default name is “Consensus”). With the option *Correct for scale differences*, the dendrograms will first be rescaled so that they have the same similarity ranges. The result is that dendrograms covering a narrow similarity range will have more impact on the consensus tree when this option is checked.

After clicking <OK>, the consensus tree is calculated, and only the clusters that contain exactly the same entries in both dendrograms are displayed.

4.10.7 Managing advanced trees

Advanced trees exist as long as a *Comparison* window is opened. Unlike conventional trees however, they are not stored along with a comparison and will disappear after the *Comparison* window is closed.

4.10.7.1 An advanced tree can be displayed by selecting it from the list that appears in **Clustering > Advanced trees**. The currently displayed tree is flagged in the menu. The currently displayed tree can be deleted with **Clustering > Advanced trees > Delete current**.

A number of dendrogram editing functions under the **Clustering** menu are not applicable to advanced trees.

4.11 Minimum spanning trees for population modelling

4.11.1 Introduction

Minimum spanning trees (MSTs) are known for a long time in the context of mathematical topology. When a set of distances is given between n samples, a minimum spanning tree is the tree that connects all samples in such a way that the summed distance of all branches of the tree is minimized.

In a biological context, the MST principle and the maximum parsimony (MP) principle share the idea that evolution should be explained with as little events as possible. There are, however, major differences between MP and MST. The MP method allows the introduction of hypothetical samples, i.e. samples that are not part of the data set. Such hypothetical samples are created to construct the internal branches of the tree, whereas the real samples from the data set occupy the branch tips. The phylogenetic interpretation of the internal branches is that they are supposed to be common ancestors of current samples, which do not exist anymore but which are likely to have existed in the past, under the criterion of parsimony.

The MST principle, in contrast, requires that all samples are present in the data set to construct the tree. Internal branches are also based upon existing samples. This means that, when a MST is calculated for evolutionary studies, there are two important conditions that have to be met: (1) the study must focus on a very short time-frame, assuming that all forms or states are still present, and (2) the sampled data set must be complete enough to enable the method to construct a valid tree, i.e. representing the full biodiversity of forms or states as closely as possible. Through these restricting conditions, the method of MST is only applicable for specific purposes, of which population modelling (micro-evolution) and epidemiology are good examples.

The trees resulting from MP on the one hand, and MST on the other hand, also have a topological difference. The MP method assumes that two (related) samples are evolved from one common ancestor through one or more mutations at either side. This normally results in a bifurcating (dichotomic) tree: the ancestor at the connecting node, and the samples at the tip. A MST chooses the sample with the highest number of related samples as the root node, and derives the other samples from this node. This may result in trees with star-like branches, and allows for a correct classification of population systems that have a strong mutational or recombinational rate, where a large number of single locus variants (SLV) may evolve from one common type¹.

An important restriction is that true MST's, e.g. according to the Prim-Jarnik algorithm can only be calculated from a true distance matrix. A criterion for a true distance matrix is that, given three samples A, B, and C, the distance from A to C should never be longer than the summed distance from A to B and B to C. This restriction implies that MSTs are not compatible with all data types. For example, a distance matrix based upon pairwise compared DNA fragment patterns does not fulfill this criterion, and hence, will not result in a true minimum spanning tree. In theory, all experiments that produce *categorical* data arrays (i.e. *multistate* character arrays) or *binary* data arrays are suitable for analysis with the MST method. The most typical applications for use with MSTs, however, are categorical Multilocus Sequence Typing (MLST) data used in population genetics and epidemiological studies.

Notwithstanding the restrictions with respect to the distance matrix, InfoQuest FP allows MSTs to be calculated from any similarity matrix. The result from similarity matrices that are known to produce untrue distance matrices (e.g. binary comparison of banding patterns) should not be regarded as true MSTs but provide interesting trees anyway.

4.11.2 Minimum spanning trees in InfoQuest FP

The MST method usually provides many equivalent solutions for the same problem, i.e. one data set can be clustered in to many MSTs with a different topology but with the same total distance. Therefore, a number of priority rules, with respect to the linkage of types in a tree, have been adopted from the BURST program (see the MLST website <http://www.mlst.net> or Feil et al., 2003²) to reduce the number of possible trees to those that have the most probable evolutionary interpretation. These rules assign priority, in decreasing order, to (1) types that have the highest number of single locus variants (SLVs) associated, (2) the highest number of double locus variants (DLVs) associated (in case of equivalent solutions), and (3) the highest number of samples belonging to the type. In InfoQuest FP, the *most frequent states* can also be used as a priority rule, and each of these rules can be assigned the first priority.

1. Maynard Smith, J., N.H. Smith, M. O'Rourke, and B.G. Spratt BG. 1993. PNAS 90: 4384-4388.

2. Feil, E.J. J.E. Cooper, H. Grundmann, D.A. Robinson, M.C. Enright, T. Berendt, S.J. Peacock, J. Maynard Smith, M. Murphy, B.G. Spratt, C.E. Moore, and N.P.J. Day. 2003. J. Bacteriol. 185:3307-3316.

As discussed in the introduction, a pure minimum spanning tree assumes that all types needed to construct a correct tree, are present in the sampled data. Conversely, algorithms like MP will introduce hypothetical nodes for every internal branch, while the samples from the data set define the branch tips.

The major problem with the MST algorithm in this view is that it requires a very complete data set to obtain a probably correct tree topology. In reality, a number of existing types may not have been included in the sampled data set. If such missing samples represent central nodes in the "true" MST, their absence may cause the resulting tree to look very different, with a much larger total spanning.

The MST algorithm in InfoQuest FP offers an elegant solution to this problem, by allowing hypothetical types to be introduced that cause the total spanning of the tree to decrease significantly. In the context of MLST, these are usually missing types for which a number of SLV (single locus variants) are present in the data set. From an evolutionary point of view, it is very likely that such types indeed exist, explaining the existence of SLVs.

4.11.3 Calculating a minimum spanning tree from character tables


The **DemoBase** does not contain a categorical data set such as MLST type data. However, the MST method can also be applied to binary data. Therefore, you can either choose to create a binary data set using the **RFLP1** fingerprint data set by calculating a global band matching table as explained in 4.3.2, or you can copy the sample MLST database which is provided on the CD-ROM. This database is a subset of 500 *Neisseria meningitidis* strains, downloaded into InfoQuest FP from the


Multi Locus Sequence Typing home page (<http://www.mlst.net>).

4.11.3.1 To generate a binary data type from **RFLP1**, follow the instructions given in 4.3.2 so as to obtain a band matching table containing the presence/absence values of the band classes for all the entries in **DemoBase**.

4.11.3.2 To install the sample MLST database of *Neisseria*, run the install program **MLST Neisseria install.exe**, that is available in the **Sample and Tutorial data\MLST sample database** directory on the CD-ROM or from the download page of the website (www.bioprad.com/softwaredownloads). This program will automatically install a new database and prompt you for the default installation directory (**C:\Program files\InfoQuest FP\data**). If this is the correct path, press Unzip to install the database. Otherwise, enter the correct path, and after installation, change the path in the InfoQuest FP Startup screen so that **MLST Neisseria.dbs** points to the correct directory.

4.11.3.3 Select all entries in the database (shortcut CTRL+A on the keyboard) and create a new comparison.

4.11.3.4 To calculate a minimum spanning tree, select **Clustering > Calculate > Minimum spanning tree (population modeling)** or press the  button and from the floating menu that appears, select

 Calculate minimum spanning tree
Population modelling

The *Minimum spanning tree* dialog box appears as depicted in Figure 4-116. This dialog box consists of four panels, about (1) the treatment of *Hypothetical types*, (2)

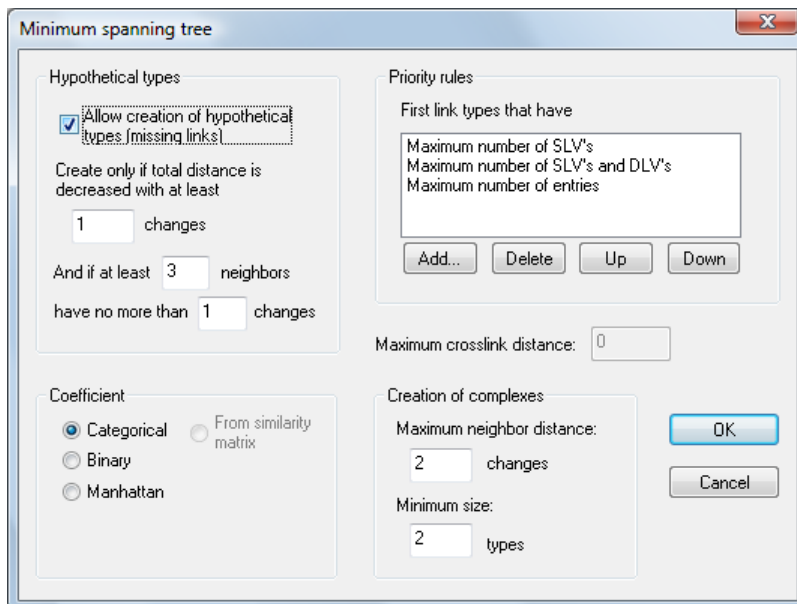


Figure 4-116. The *Minimum spanning tree* dialog box, with *Allow creation of hypothetical types* checked.

the *Coefficient* to calculate the distance matrix, (3) the *Priority rule* for linking types in the tree, and (4) the settings for the *Creation of complexes*

•Hypothetical types:

With the checkbox *Allow creation of hypothetical types (missing links)*, you can allow the algorithm to introduce hypothetical types as branches of the MST, as described in 4.11.2. When enabled, the following criteria can be specified:

- Create only if total distance is decreased with at least (default 1) changes*: Only in the case the introduction of a hypothetical type decreases the total spanning of the tree with one change, the hypothetical type will be accepted.
- And if at least (default 3) neighbors have no more than (default 1) changes*: The algorithm will only accept hypothetical types that have at least 3 neighbors (closest related types) that have no more than 1 changes (see also 4.11.2 for the interpretation of this rule).

•Coefficient:

The choice is offered between *Categorical*, for categorical data and *Binary*, for binary data. When *Manhattan* is checked, the sum of the absolute differences between the values of any two corresponding states is calculated, and the thus obtained distances are used to calculate the MST. This option can be used to cluster non-binary, non-categorical data with integer values. If non-integer (decimal) values are used, the program will round them to the closest integers.

In the *Manhattan* option, an *Offset* and a *Saturation* value can be specified. For each character compared between two types, the *offset* value determines a fixed distance that is added to the distance of these characters. If the distance is zero, however, the transformed distance remains zero. In addition, for each character compared between two types, the *saturation* determines the maximum value the distance can take. In other

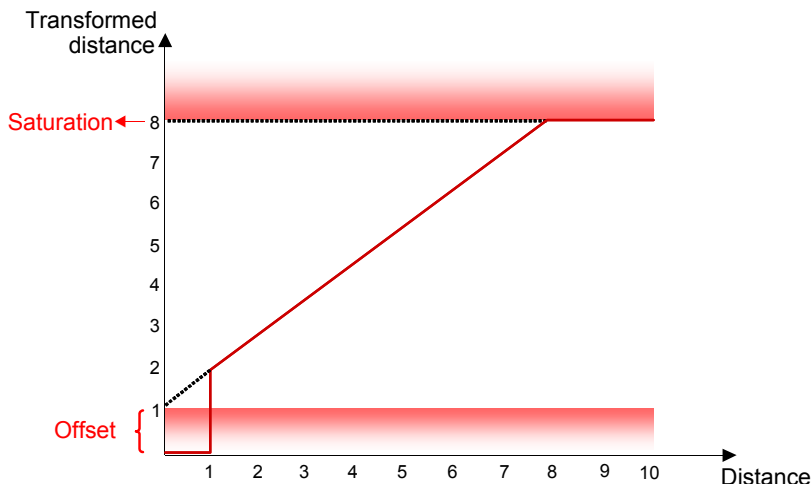


Figure 4-117. Graphical representation of the meaning of *offset* and *saturation* values.

words, above the saturation distance, different characters are all seen equally different. The relation between offset, saturation, and distance of characters is illustrated in Figure 4-117. The offset and distance can be used to tune the summed distance result between fully categorical (offset = 1 and saturation = 1) and fully numerical (offset = 0 and saturation infinite).

•Priority rules

In case of equivalent solutions in terms of calculated distance, the priority rules allow you to specify a priority based upon other criteria than distance. One or more rules can be added, with a maximum of 3. The order of appearance of the rules determines their rank.

4.11.3.5 A rule can be added by pressing the <Add> button. One of the following rules can be selected (Figure 4-118):

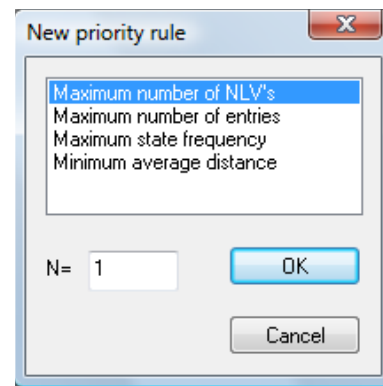


Figure 4-118. Priority rule selection box.

- Maximum number of NLV's* (Ntuple Locus Variants). N has to be chosen by the user. For example, if N is 1, the rule becomes "*Maximum number of SLV's*". This means that, in case two types having an equal distance to a linkage position in the tree, the type that has the highest number of *single locus variants* (i.e. other types that differ only in one state or character) will be linked

first. If $N=2$, the rule becomes “*Maximum number of SLV’s and DLV’s*”.

- **Maximum number of entries:** The program counts how many entries each unique type contains, and the type that has the highest number of entries will be assigned priority, in case of equivalent linkage possibilities.
- **Maximum state frequency:** The program calculates a frequency table for each state of each character. Types are thus ranked based upon the product of frequencies of their characters. In case of equivalent possibilities, types that have the highest state frequency rank are linked first.
- **Minimum average distance:** the type that has the lowest average distance with the other types will be linked first, in case of equivalent solutions.

4.11.3.6 Since the order of appearance in the list of defined priority rules is determinative for the order of execution, it is possible to move rules up or down using the **<Up>** and **<Down>** buttons.

• Maximum crosslink distance

With this option, you can allow the program to display alternative equivalent solutions under the clustering criterion used, which are then displayed as crosslinks. Suppose that the program has linked group B to group A because they differ in one state. If group B has also one state difference with another group, C, it will be shown as a crosslink between B and C. Crosslinks are indicated as dark red lines.

4.11.3.7 You can specify the maximum number of states difference before a crosslink will be displayed. For example, if 2 is entered, only crosslinks between groups that have 1 or two states difference will be indicated. If 0 is entered, crosslinks will not be shown.

• Creation of complexes

In epidemiological population genetics based upon MLST, a *clonal complex* can be defined as a single group of isolates sharing identical alleles at all investigated loci, plus single locus variants that differ from this group at only one locus¹. In another, more relaxed definition^{2,3}, a clonal complex includes all types that differ in x loci or less from at least one other type of the complex (x is usually taken as 1 or 2). Under this definition, not all types of a complex are necessarily SLVs or DLVs from one another. The latter definition is used in InfoQuest FP.

1. Feil, E.J., J. Maynard Smith, M.C. Enright, and B.G. Spratt. 2000. *Genetics* 154: 1439-1450.

2. Feil, E.J. J.E. Cooper, H. Grundmann, D.A. Robinson, M.C. Enright, T. Berendt, S.J. Peacock, J. Maynard Smith, M. Murphy, B.G. Spratt, C.E. Moore, and N.P.J. Day. 2003. *J. Bacteriol.* 185:3307-3316.

3. BURST (Based Upon Related Sequence Types) program description, see the MLST website <http://www.mlst.net>.

The maximum number of changes allowed to form complexes can be specified; the default value is 2. In addition, one can also specify a minimum number of types that should be included before the groups is defined as a complex. The default value is 2.

4.11.4 Interpreting and editing a minimum spanning tree

After pressing **<OK>** in the *Minimum spanning tree* dialog box, the *Minimum spanning tree* window will pop up. In the example shown in Figure 4-119 and Figure 4-120, a band matching table of **RFLP1** in the **DemoBase** database was created and analyzed as *Binary*, while the other parameters were left to the defaults.


The window is divided in four panels, of which the *Minimum Spanning Tree* panel displays the actual MST, the *Node content* panel lists the entries belonging to the selected node or nodes, the *Node properties* panel shows the type for the selected node or nodes and the *Rooted Complexes* panel (in default configuration tabbed view with the *Minimum Spanning Tree* panel) displays the composition of the complexes. All four panels in the *Minimum spanning tree* window are dockable.

• Display options

In the *Minimum Spanning Tree* panel, each type is represented by one node or branch tip, displayed as circles that are connected by branches. In the default settings, but with **Letter code** selected under **Type labeling**, the following information can be derived from the tree view:

- When sufficiently zoomed (using the zoom buttons



and  , the zoom slider (see 1.6.7) or the keyboard shortcuts CTRL+PgUp and CTRL+PgDn), a letter code will appear within each circle, uniquely identifying each type. In case of more than 26 types in total, a two-letter code is used, of which the second can be a digit 1-9 as well. The codes are assigned alphabetically according to the *Priority rule* specified (see 4.11.3).

- The length of the branches is proportional to the distance between the types, and the thickness, dotting, and graying of the branch lines also indicate the distance between the nodes.
- The number of entries contained in a type (node) is indicated using a color ranging from white over three blue shades to brown and red.

In the *Rooted complexes* panel, the complexes are displayed as defined under the specified calculation settings (see 4.11.3).

- Each complex is shown as a rooted tree, with the type having the highest priority, as defined by the *Priority rule* (4.11.3) defining the root. On top of the *Rooted*

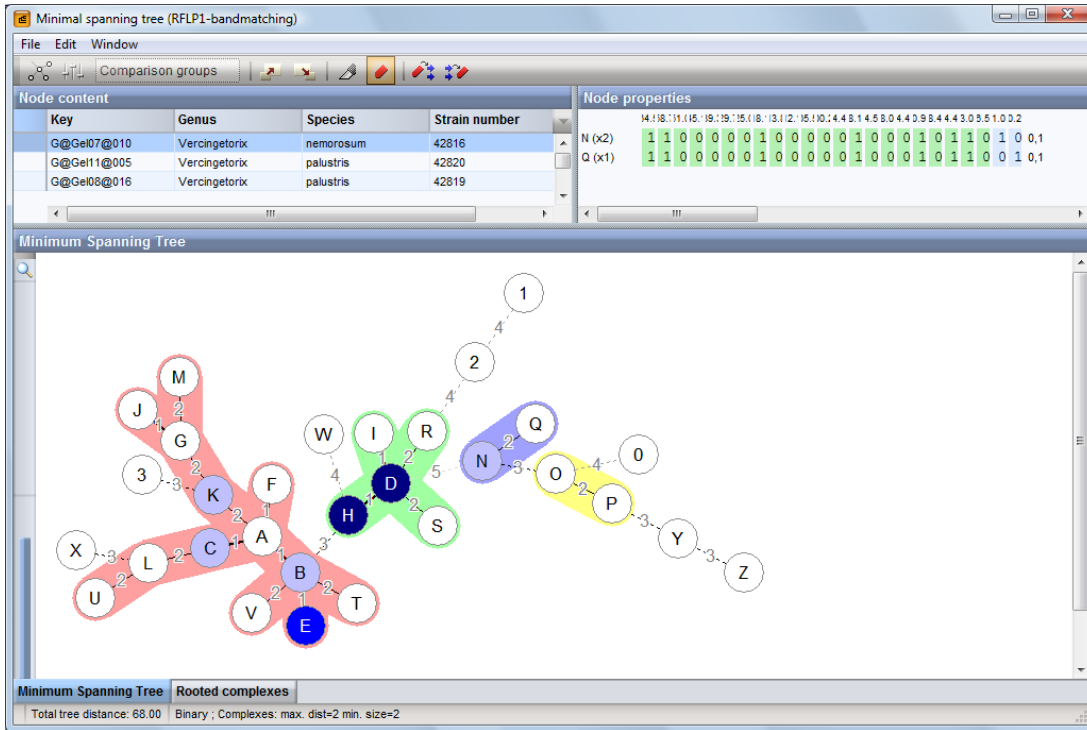


Figure 4-119. The *Minimum spanning tree* window, *Minimum Spanning Tree* panel displayed.

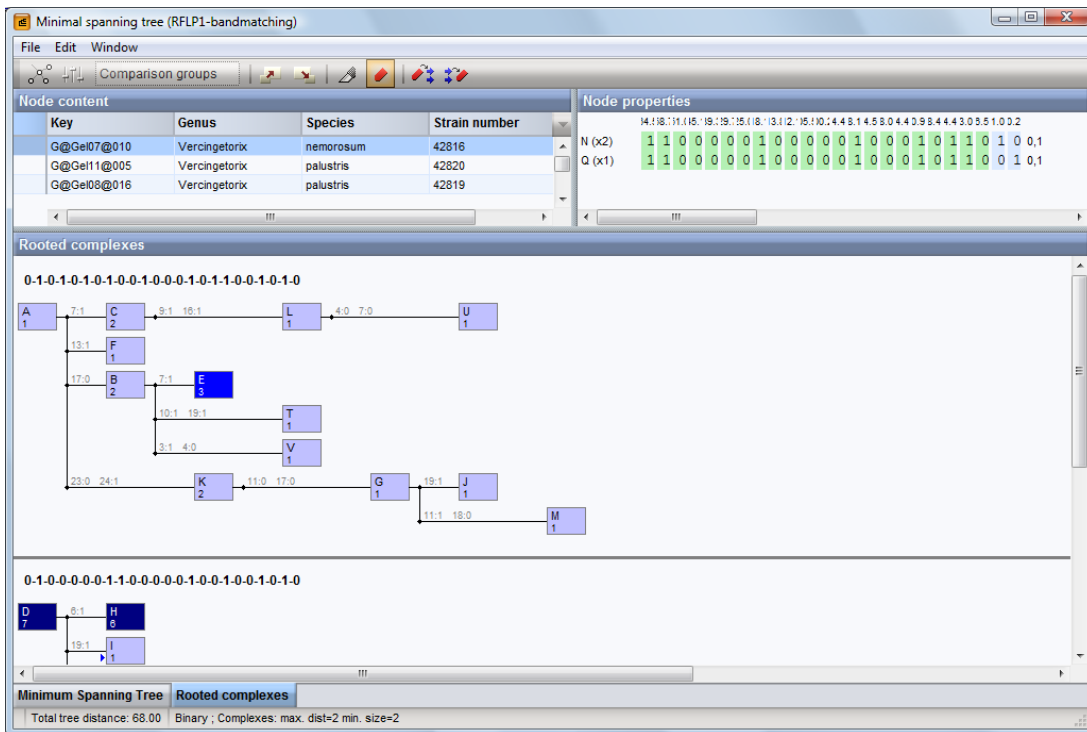


Figure 4-120. The *Minimum spanning tree* window, *Rooted complexes* panel displayed.

complexes panel, the character values of the root type are indicated. The branch lengths of the derived types (i.e., the types branching from the root) are in proportion to the distances of these types.

- For each type branching off from the root type, the change(s) is (are) indicated as two numbers separated by a colon. The first number is the character number, and the second number is the value towards the character has changed. For example, 6:003 means that character 6 has changed into 003 for this type. If more

than one change has led to a derived type, the changes are indicated next to each other.


- Similarly as on the tree, the types are indicated with a color reflecting the number of entries contained in the type. In addition, the number of entries is written just below the type code.

In the *Node content* panel, the entries contained in the selected node(s) are shown in a grid or tabular format). If the entries are selected in the *Comparison* window, this is indicated here as well, with the same colored arrows. Selections can be made in this list using the CTRL and SHIFT keys, and the entry card can be popped up by double-clicking on an entry.

The *Node properties* panel displays the details of the highlighted type(s) in the *Minimum Spanning Tree* panel or the *Rooted complexes* panel. If a type is selected in the *Minimum Spanning Tree* panel, it becomes highlighted by a red circle, and marked with a red flag. The same type becomes highlighted in the *Rooted complexes* panel, by a red rectangle. For the highlighted type, detailed information is shown in the *Node properties* panel.

- On top of the panel, the character names and character values (on green background) are shown for the highlighted type. The frequencies of the character values are indicated in gray.
- Left from the character list is the name of the type with the number of entries it contains between brackets.
- Right from the character list is the number of SLVs (single locus variants; types differing only in one character) and DLVs (double locus variants; types differing in two characters).
- In case more than one type is highlighted in the *Minimum Spanning Tree* panel or the *Rooted complexes* panel, the highlighted types are displayed under each other in the *Node properties* panel. Characters that are the same for more than 50% of the types are shown on a green background. Characters for which there is less than 50% consensus are shown on a white background. A character that is different from the majority in a type is indicated in red.

• Edit options

4.11.4.1 With *Edit > Display settings* or , the display options can be customized in the *Display settings* dialog box (Figure 4-121).

4.11.4.2 Under *Cell color*, you can use a color to display the number of entries, any groups or field states defined, or the groups pie charts. The colors are displayed both in the *Minimum Spanning Tree* panel and the *Rooted complexes* panel.

4.11.4.3 With *Number of entries* selected, a differential color will be assigned to the nodes according to the

number of entries they contain. The intervals can be specified under *Number of entries coding*.

4.11.4.4 With *Groups* selected, the colors assigned to the groups (see 4.1.11) in the comparison or to field states (if defined, see), will be given to the nodes. When a type consists of more than one group, it will become black. *Groups (pie chart)* is similar, except that, in case a type (node) consists of more than one group, the different groups will be represented in a pie chart. This option also works in combination with the *Compact complexes* option (4.11.4.9). In the *Rooted complexes* panel, the different group colors are also displayed in the type boxes, in a proportional way.

4.11.4.5 *Number of entries coding* is only enabled when *Number of entries* is selected under *Cell color*.

NOTE: By default, the first color (white) is set as ≤ 0 . This means that only empty nodes are white. This is useful to visualize hypothetical nodes (see 4.11.3) when this option is enabled. When no hypothetical nodes are allowed, it is more useful to enter a positive value, for example ≤ 1 , as has been done to create Figure 4-119.

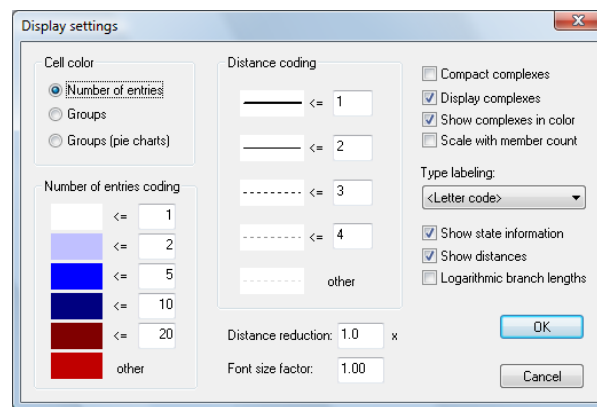


Figure 4-121. The *Display settings* dialog box in the *Minimum spanning tree* window.

4.11.4.6 Under *Distance coding*, you can specify the distance that corresponds with the different line types offered by the program.

4.11.4.7 With *Distance reduction*, you can change the length of the branches. In the *Minimum Spanning Tree* panel, this only changes the zoom, but in the *Rooted Complexes* panel, this value will determine the horizontal distance between the types displayed. With a distance reduction of e.g. 1.5x, the distance unit is decreased with a factor 1.5.

4.11.4.8 The option *Display complexes* allows you to choose whether the complexes are displayed or not. Note that this option only applies to the *Minimum Spanning Tree* panel; the complexes remain displayed in the *Rooted Complexes* panel.

4.11.4.9 Using *Compact complexes*, you can choose to display a full complex as one node on the tree. The diameter of the circle is (slightly) proportional to the number of types the complex contains, and the compacted complexes are encircled.

4.11.4.10 With *Use color*, you can display the image in color or grayscale mode.


4.11.4.11 *Scale with member count* is an option that lets the diameter of the circles depend on their size.


4.11.4.12 Under *Type labeling*, it is possible to select the *Letter code* which is automatically assigned to the types by the program, or any of the information fields the database contains. In the latter case, types (nodes) that do not all have the same string will be marked with ???. Note also that you may have to zoom in sufficiently to visualize longer labels than the letter codes. If the labels do not fit within the circle, they are represented by

4.11.4.13 The option *Show state information* relates to the *Node properties* panel, where the states of any selected types can be displayed. When this option is unchecked, the states of the characters for the selected types are not displayed.

4.11.4.14 With *Show distances*, you can have the distances indicated on the branches of the tree.

As indicated earlier, it is possible to highlight types on the tree or in the *Rooted Complexes* panel. You can use the SHIFT or CTRL keys to highlight multiple types, or drag a rectangle with the mouse in the tree or the complex panel. For a single highlighted type, you can select individual entries directly in the *Node content* panel.

4.11.4.15 For one or more highlighted types, it is also possible to select all the entries directly from the tree panel, by pressing the  button or choosing *Edit > Select all entries in selected nodes* from the menu.


4.11.4.16 Likewise, it is possible for any selected entry to highlight all the types where this entry occurs, using the  button or *Edit > Select nodes that contain selected entries*.


4.11.4.17 With *Edit > Select related nodes*, you can highlight all the types that have no more than a specified number of changes from the highlighted type(s). When choosing this menu command, the program asks to enter the maximum distance from the highlighted type(s).

4.11.4.18 On the tree, the highlighted nodes are, by default, marked with a red label, and with a red circle as well. You can choose to hide or show this label using



button or with *Edit > Label selected nodes*.

4.11.4.19 The *Cut branch tool* (*Edit > Cut branch tool* or ) is a cursor tool that allows a branch of the tree to be "cut off" and displayed as one simple end node. A branch can be cut off by selecting the branch cut tool, moving the cursor towards one end of a branch and left-clicking. When cut off, the branch is displayed as a green node which always has the same size, regardless of the zoom. To disclose the branch again, simply double-click on the green node.

4.11.4.20 The drop-down list  allows you to select a coloring based on groups or any of the available field states.

4.11.4.21 The complexes present on the minimum spanning tree can be converted into Groups using *File > Convert complexes to groups*.

4.11.4.22 With *Edit > Show crosslinks*, you can toggle between displaying and hiding the crosslinks. CTRL+C is a shortcut for this operation. This feature cannot be combined with the *Compact complexes* option (4.11.4.9).

4.11.5 Calculating a minimum spanning tree from a similarity matrix

As explained in the introduction (see 4.11.1), InfoQuest FP allows MST's to be calculated from similarity matrices obtained from any type of data. The condition to use a true distance matrix as input is thereby not necessarily met. In particular when banding patterns are compared using a binary band matching coefficient, where infringements upon the transitivity rule happen frequently (see also Section 4.10), a MST cannot perfectly depict the "odd" relationships given in the similarity matrix. Since there is no tree algorithm that can deal with intransitivities, it is equally justified to apply the MST algorithm as, for example, UPGMA to such matrices.

To convert a similarity matrix into an integer distance matrix, the software uses bins of certain similarity intervals that will be converted into distance units. For example, with a similarity bin size of 1%, two entries that have a similarity of 99.6% will have a distance of zero. Two entries that have a similarity of 98.7% will have a distance of 1.

As an example, we will analyze the fingerprint type **RFLP1** in **DemoBase**.

4.11.5.1 In the *InfoQuest FP main* window with **DemoBase** loaded, open comparison **A11**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

4.11.5.2 Select **RFLP1** in the *Experiments* panel and choose *Clustering > Calculate > Cluster analysis (similarity matrix)*.

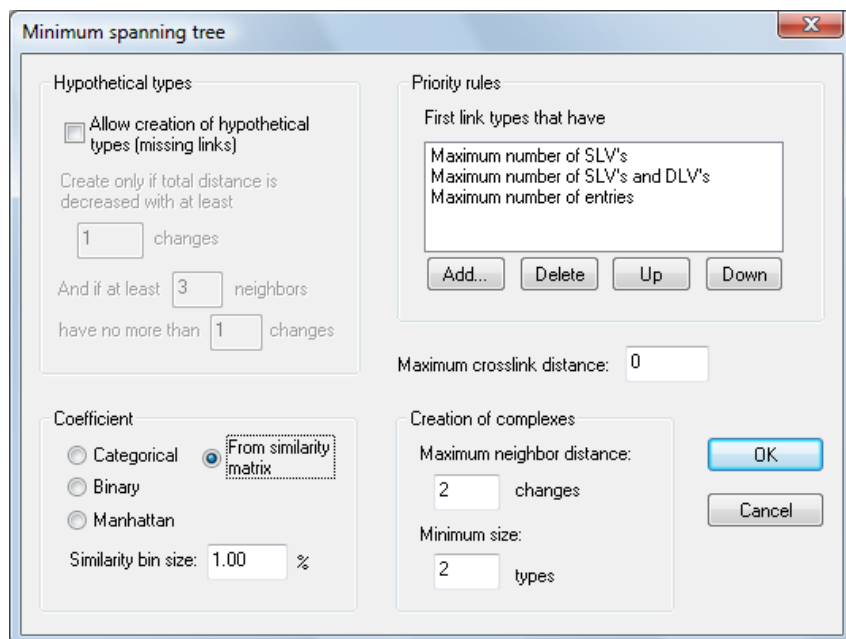


Figure 4-122. *Minimum spanning tree* dialog box, starting from similarity matrix.

Under **Similarity coefficient**, select *Different bands*. This will generate a similarity matrix with discrete integer distance values, which will result into an easily interpretable MST.

4.11.5.3 It does not matter what tree algorithm is used; only the similarity matrix is needed. Press **<OK>** to calculate the matrix.

4.11.5.4 Select *Clustering > Calculate > Minimum spanning tree (population modelling)*.

In the *Minimum spanning tree* dialog box that pops up, the option *From similarity matrix* is selected and the other options relating to character data are disabled (Figure 4-122). The *Similarity bin size* allows the bin

size for the conversion to a discrete distance matrix to be specified.

The other options in this dialog box are as explained in 4.11.3.

4.11.5.5 Leave the bin size to 1% and press **<OK>** to calculate the MST.

In the resulting MST, connecting lines between nodes can be directly translated into numbers of different bands: short thick line = 1 band different; thin full lines = 2 bands different; black dashed lines = 3 bands different etc. Of course, all entries that have no bands different fall in the same node.

4.12 Dimensioning techniques

4.12.1 Introduction

Principal Components Analysis (PCA) and Multi-Dimensional Scaling (MDS) are two alternative grouping techniques that can both be classified as *dimensioning techniques*. In contrast to dendrogram inferring methods, they do not produce hierarchical structures like dendrograms. Instead, these techniques produce two-dimensional or three-dimensional plots in which the entries are spread according to their relatedness. Unlike a dendrogram, a PCA or MDS plot does not provide "clusters". The interpretation of the obtained comparison is, more than in cluster analysis, left to the user.

PCA assumes a data set with a known number of characters and analyzes the characters directly. PCA is applicable to all kinds of character data, but not directly to fingerprint data. Fingerprints can only be analyzed when converted into a band matching table (see 4.3.2).

MDS does not analyze the original character set, but the matrix of similarities obtained using a similarity coefficient. Rather than being a separate grouping technique, MDS just replaces the clustering step in the sequence *characters > similarity matrix > cluster analysis*. However, it is a valuable alternative to the dendrogram methods, which often oversimplify the data available in a similarity matrix, and tend to produce overestimated hierarchies.


4.12.2 Calculating an MDS

4.12.2.1 Any experiment type for which a complete similarity matrix is available can be analyzed by MDS. Matrix types are not suitable for MDS clustering if the matrices are incomplete.

4.12.2.2 In the *InfoQuest FP main* window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

4.12.2.3 Select **FAME** in the *Experiments* panel and check whether a matrix is available for this experiment type by looking in the *Layout* menu if the menu command *Show matrix* is enabled (not grayed).

4.12.2.4 If *Show matrix* is grayed, first calculate a dendrogram with *Clustering > Calculate > Cluster analysis (similarity matrix)*.

4.12.2.5 Select *Dimensioning > Multi-dimensional scaling* or .

The program now asks "*Optimize positions*". InfoQuest FP iteratively recalculates the MDS, each time again optimizing the positions of the entries in the space to resemble the similarity matrix as closely as possible. If you allow the optimization to happen, the calculations take slightly longer.

4.12.2.6 Press **<Yes>** to optimize the positions.

The MDS is calculated and the *Coordinate space* window is shown (see Figure 4-123).

4.12.3 Editing an MDS

The *Coordinate space* window (Figure 4-123) shows the entries as dots in a cubic coordinate system.

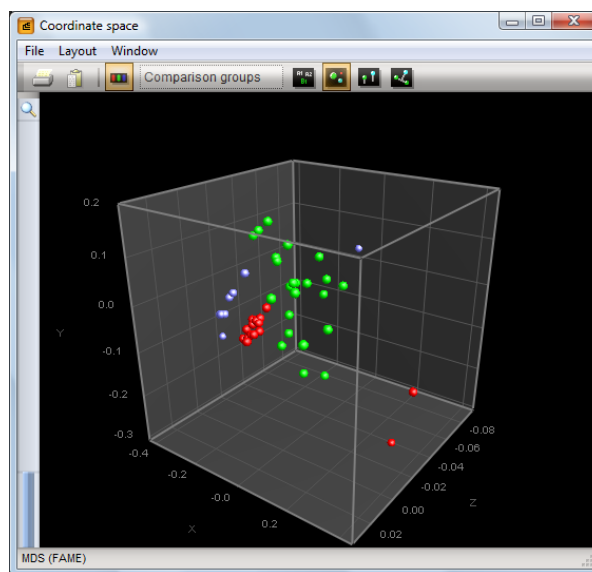



Figure 4-123. *Coordinate space* window, resulting from a PCA or MDS analysis.

4.12.3.1 To zoom in and zoom out on the image, press the PgDn and PgUp keys, respectively. Alternatively, the zoom slider can be used (see 1.6.7 for the zoom slider functions).

4.12.3.2 The image can be rotated in real time by clicking on the image and dragging in the desired direction with the mouse.

By default, the entries are represented as 3D spheres in a realistic perspective. They appear in the colors as defined for the groups on the dendrogram (4.1.11).

4.12.3.3 With *Layout > Show keys* or , you can display the database keys of the entries instead of the dots.

However, the entry keys may be long and uninformative for the user, so the entry keys can be replaced by a group code. The program assigns a letter to each defined group, and within a group, each entry receives a number. The group codes are shown as follows:


4.12.3.4 In the parent *Comparison* window, select *Layout > Use group numbers as key*.

4.12.3.5 A legend to the group numbers can be obtained with *File > Export database fields* in the parent *Comparison* window.

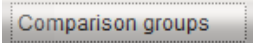
Alternatively, a selected information field can be displayed instead of the key:

4.12.3.6 In the parent *Comparison* window, click on the information field which you would like to see displayed (e.g. 'Strain number').

4.12.3.7 Select *Layout > Use field as key* from the menu in the *Comparison* window. The strain number is now displayed in the MDS plot.

4.12.3.8 With *Layout > Show group colors* or , you can toggle between the color representation and the non-color representation, in which the entry groups are represented (and printed) as symbols instead of colored dots.


On the screen, it is generally easier to evaluate the groups using colors.

4.12.3.9 If field states with corresponding color coding were defined, the drop-down list  allows you to select a coloring based on groups or any of the available field states.


4.12.3.10 Select an entry coordinate system using CTRL+left mouse click. Selected entries are contained in a blue cube.

4.12.3.11 To select several entries at a time, hold down the SHIFT key while dragging the mouse in the coordinate system. All entries included in the rectangle will become selected.


4.12.3.12 By double-clicking on an entry, its *Entry edit* window is popped up.

4.12.3.13 With *Layout > Show construction lines* or , the entries are displayed on vertical lines starting


from the bottom of the cube. This may facilitate the three-dimensional perception. Disable this option to view the next features.


4.12.3.14 With *Layout > Show rendered image* or , you can toggle between the realistic three-dimensional perspective with entries represented by spheres, and a simple mode where entries are represented as dots.

4.12.3.15 With *Layout > Preserve aspect ratio* enabled, the relative contributions of the three components are respected, which means that the coordinate system is no longer shown as a cube.

4.12.3.16 Another very interesting display option is *Layout > Show dendrogram* or .

When this option is enabled, the entries in the coordinate system are connected by the dendrogram branches from the parent *Comparison* window. This is an ideal combination to co-evaluate a dendrogram and a coordinate system (PCA or MDS).

4.12.3.17 To copy the coordinate space image to the clipboard, select *File > Copy image to clipboard* or .

4.12.3.18 The image can be printed with *File > Print image* or . The image will print in color if the colors are shown on the screen.

4.12.4 Calculating a PCA

PCA is typically executed on complete character data. It does not work on sequence types. Fingerprints can only be analyzed by PCA if a band matching table is first generated (see 4.3.2).

4.12.4.1 In the *InfoQuest FP main* window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as STANDARD (see 4.1.3.2 to 4.1.3.4).

4.12.4.2 Select **FAME** in the *Experiments* panel and *Dimensioning > Principal Components Analysis* or



The *Principal Components Analysis* dialog box (Figure 4-124) allows a number of more advanced choices to be made.

The simplest choice is "*Use quantitative values*". By default, this choice is checked, and if the technique provides quantitative information (not just absent/present), one will normally want to use this information for the PCA calculation. If this option is unchecked, the character values will be converted to binary as specified in the *Conversion to binary* settings.

	CHAR 1	CHAR 2	CHAR 3
ENTRY 1	VAL 11	VAL 12	VAL 13
ENTRY 2	VAL 21	VAL 22	VAL 23
ENTRY 3	VAL 31	VAL 32	VAL 33

Diagram annotations: A green box highlights the CHAR 2 column, with a green arrow pointing to it from the word "CHARACTER" above and another green arrow pointing down to the text "AVERAGE, VARIANCE" below. A red box highlights the ENTRY 2 row, with a red arrow pointing to it from the word "ENTRIES" on the left and another red arrow pointing to the text "AVERAGE, VARIANCE" on the right.

Figure 4-125. Character table showing the meaning of *Average* and *Variance* correction at the *Entries* and *Characters* level.

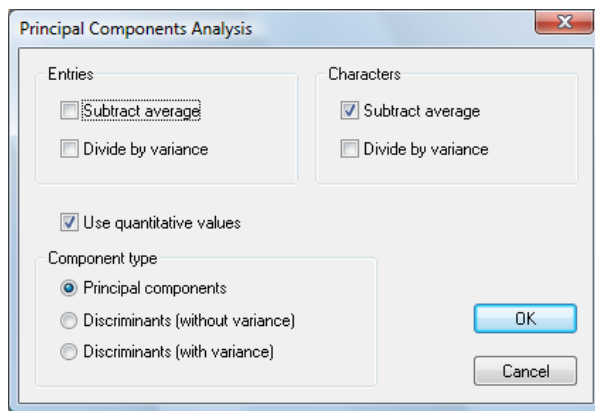


Figure 4-124. *Principal Components Analysis* dialog box.

More sophisticated options are the possibilities to *Subtract average* character value over the *Entries*, and to *Subtract average* character value over the *Characters*. Figure 4-125 explains how the averaging works.

- Subtraction of the *averages* over the *characters* (green in the figure) results in a PCA plot arranged around the origin, and therefore, it is recommended for general purposes.
- Division by the *variances* over the *characters* (green in the figure) results in an analysis in which each character is equally important. Enabling this option can be interesting in a study containing characters of unequal occurrence. For example, if fatty acid extractions are analyzed for a set of bacteria, some fatty acids may be present in abundant amounts, whereas others may occur only in very small amounts. It is well possible that the “minor” fatty acids are as informative or even more informative than the abundant ones, taxonomically seen. If no correction is applied, those minor fatty acids will be completely masked by differences in the abundant fatty acids. Dividing by the variance for each fatty acid normalizes for such range differences, making each character equally contributing to the total separation of the system.

- Subtraction of the *averages* over the *entries* (red in the figure) results in character sets of which the sum of characters equals zero for each entry. This feature has little meaning for general purposes.

- Division by the *variances* over the *entries* (red in the figure) results in character sets for which the intensity is normalized for all entries. For example, suppose that you have scanned phenotypic test panels for a number of bacterial strains and want to calculate a PCA. If some strains are less grown than others, the overall reaction in the wells will be less developed. Without correction, well developed and less developed panels will fall apart in the study. Dividing by the variances normalizes the character sets for such irrelevant differences, making character sets with different overall character developments fall together as long as the relative reactions of the characters are the same.

*NOTE: The two latter features are exactly what is done by the Pearson product-moment correlation coefficient. This coefficient subtracts each character set by its average, and divides the characters by the variance of the character set. The feature **Divide by variance** under **Entries** should not be used in character sets where the characters are already expressed as percentages (for example, fatty acid methyl esters).*

The lower panel of the dialog box (Figure 4-124) displays the *Component type*. This can be *Principal components*, *Discriminants (without variance)*, or *Discriminants (with variance)*. The first option is to calculate a principal components analysis, whereas the *Discriminants* options are to perform discriminant analysis. These options are described in paragraph 4.12.7.

4.12.4.3 In the *Entries* and *Characters* panels, check *Subtract average* under *Characters*, and leave the other options unchecked.

4.12.4.4 In the *Component type* panel, select *Principal components*, and press <OK>. Calculation of the PCA is started.

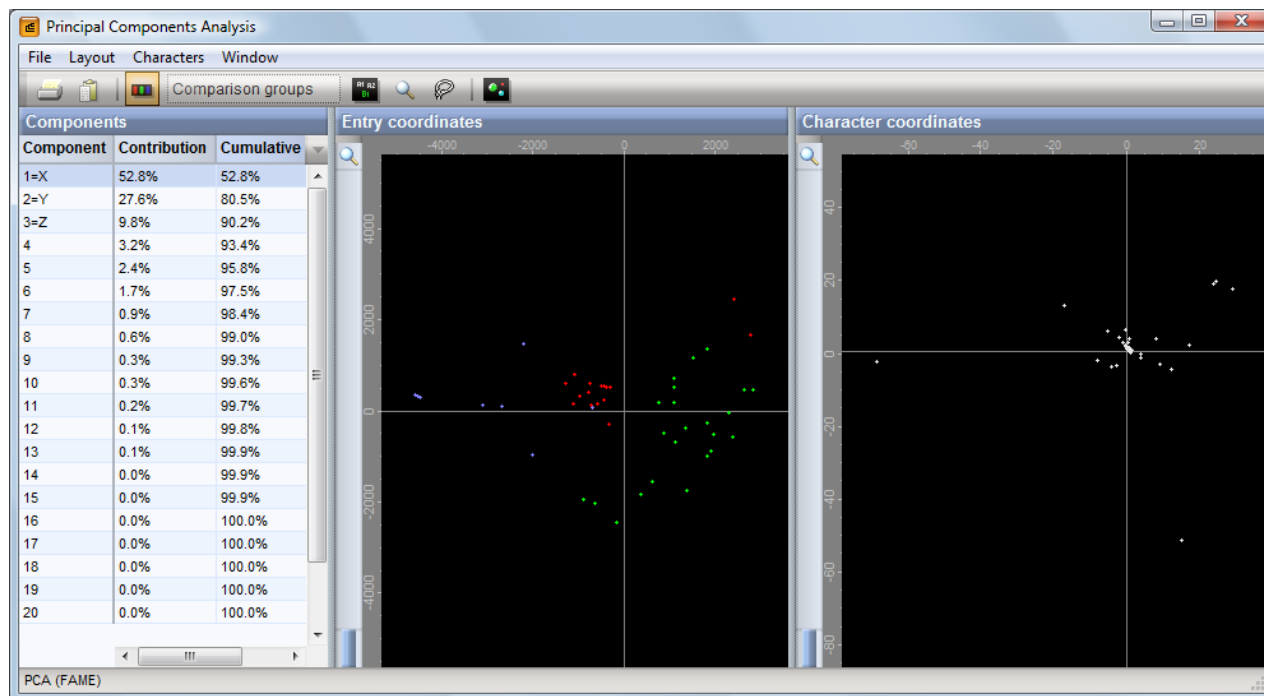


Figure 4-126. *Principal components analysis window.*

The resulting window, the *Principal components analysis* window, is shown in Figure 4-126.

The *Principal components analysis* window is divided in three dockable panels. In the *Components* panel (the left panel in default configuration), the first 20 components are shown, with their relative contribution and the cumulative contribution displayed. Also, the components used as X-, Y- and Z-axes are indicated. The *Entry coordinates* panel shows the *entries* plotted in an X-Y diagram corresponding to the first two components. The *Character coordinates* panel shows the characters plotted in the same X-Y diagram. From the *Character coordinates* panel, one can see the contribution each character has to the two displayed components, and hence, what contribution it has to the separation of the groups along the same components. For example, if a group of entries appears left along the X-axis whereas the other entries appear right, those characters occurring left on the X-axis are positive for the left entries and negative for the right entries, and *vice versa*.

By default, the first component is used for the X-axis, the second component is used for the Y-axis, and the third component is used for the Z-axis. The Z-axis is not shown here, but can be shown in the three-dimensional representation with *Layout > Show 3D plot* (see further).

4.12.4.5 If you want to assign another component as one of the axes, select the component in the *Components* panel, and *Layout > Use component as X axis*, *Layout > Use component as Y axis*, or *Layout > Use component as Z axis* (or right-click on the component).

• Layout tools:

4.12.4.6 Switching from color indication for the groups to symbol indication with *Layout > Show group colors*



4.12.4.7 Select a coloring based on groups or any of the available field states from the drop-down list


Comparison groups

4.12.4.8 Showing the keys or a unique label based upon the groups for the entries with *Layout > Show keys* or



*NOTE: In case keys are assigned automatically by the program, they are not very informative, so one should select **Layout > Use group numbers as key** in the underlying Comparison window. A list of the group codes and the corresponding entry names can be generated in the underlying Comparison window with **File > Export database fields**. Alternatively, click on an information field and select **Layout > Use field as key** in the underlying Comparison window.*

4.12.4.9 The option *Layout > Preserve aspect ratio* allows you to either preserve the aspect ratio of the components, i.e. the relative discrimination of the component on the Y axis with respect to the component on the X axis, or to stretch the components on the axes so that they fill the image optimally.



4.12.4.10 With *Layout > Zoom in / zoom out* or , you can zoom in on any part of the *Entry coordinates* or

Character coordinates panel of the PCA plot: drag the mouse pointer to create a rectangle; the area within the rectangle will be zoomed to cover the whole panel. In order to restore the original size of the image, simply left-click within the panel. Disable the zoom-mode afterwards. Alternatively, the zoom sliders of the *Entry coordinates* and *Character coordinates* panel can be used to zoom in or out on the plots (see 1.6.7 for a description of the zoom slider functions).

4.12.4.11 If you move the mouse pointer over the *Character coordinates* panel (characters), the name of the pointed character is shown.


• Editing tools:


4.12.4.12 Entries can be selected in a *Principal components analysis* window by holding the SHIFT key down and selecting the entries in a rectangle using the left mouse button. Selected entries are encircled in blue. You can also hold down the CTRL key while clicking on an entry.


4.12.4.13 An even more flexible way of selecting entries is using the lasso selection tool. To activate the lasso selection tool, choose *Layout > Lasso selection tool* or press the  button. With the lasso selection tool enabled, selections of any shape can be drawn on the plot. The lasso selection tool menu item is flagged and the button shown as  when the tool is enabled. To stop using the lasso selection tool, you have to click the button a second time, or disable it from the menu.

A PCA is automatically saved along with its parent *Comparison* window. It is possible to add entries to an existing PCA or remove entries from it. The feature to add entries to an existing PCA is an interesting alternative way of identifying new entries. They can be placed in a frame of known database entries, and in this way, identifying is just looking at the groups they are closest to. **Since the components are not recalculated when entries are added to an existing PCA, the PCA does not reflect the full data matrix anymore!**

4.12.4.14 If you want to add entries to an existing PCA, you can select new entries in the *InfoQuest FP main* window and copy them to the clipboard using *Edit >*

Copy selection or .

4.12.4.15 In the *Comparison* window, select *Edit > Paste selection* or . The new entries are placed in the *Comparison* window and in the *PCA* window.


4.12.4.16 To delete entries from a PCA, select the entries as in 4.12.4.12 and in the *Comparison* window, select *Edit > Cut selection* or .


If you started the PCA from a composite data set, you can order the characters according to the selected

component in the underlying *Comparison* window. This is an interesting feature to locate characters that separate groups you are interested in. The feature works as follows (only for composite data sets):

4.12.4.17 In the *Principal components analysis* window, first determine the component that best separates the groups.

4.12.4.18 Select that component in the *Components* panel and select *Characters > Order characters by component*. The characters are now ordered by the selected component in the underlying *Comparison* window.

4.12.4.19 The entry plot can be printed with *File > Print image (entries)* or  and the character plot can be printed with *File > Print image (characters)*.

4.12.4.20 Alternatively, the entry plot can be copied to the clipboard with *File > Copy image to clipboard (entries)* or  and the character plot can be copied to the clipboard with *File > Copy image to clipboard (characters)*.

If you want to reconstruct or analyze the PCA system in another software package, it is possible to export the coordinates of the entries along a selected component (for example the X-axis):

4.12.4.21 Select a component and *File > Export selected entry coordinates*.

If you want to reconstruct the PCA with the first two components, you should also export the second component (Y-axis), by selecting that component and *File > Export selected entry coordinates*.


4.12.4.22 It is also possible to export all entry coordinates at once in a tab-delimited format using *File > Export all entry coordinates*.

Similarly, one can export the coordinates for the characters for a certain component:

4.12.4.23 Select a component and *File > Export selected character coordinates*.

4.12.4.24 To export all character coordinates at once, use *File > Export all character coordinates*.

InfoQuest FP allows you to display three components at the same time, by plotting the entries in a 3-dimensional space.

4.12.4.25 To create a three-dimensional plot from the PCA, select *Layout > Show 3D plot* or .

The *Coordinate space* window is shown. See 4.12.3 to edit a PCA in 3-D representation mode.

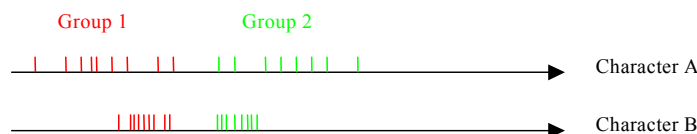


Figure 4-127. The influence of character spread on discriminant analysis.

4.12.4.26 Close the *Coordinate space* window with **File > Exit**.


4.12.4.27 Close the *PCA* window with **File > Exit**.


4.12.5 Calculating a discriminant analysis

Discriminant analysis is very similar to PCA. The major difference is that PCA calculates the best discriminating components for the character table as a whole, without foreknowledge about groups, whereas discriminant analysis calculates the best discriminating components for groups that are defined by the user. In case of discriminant analysis, these principal components are then called *discriminants*. Like PCA, discriminant analysis is executed on complete character data. It does not work on sequence types. Fingerprints can only be analyzed by discriminant analysis if a band matching table is first generated (see 4.3.2). Discriminant analysis also forms the basis for multivariate analysis of variance (MANOVA), which is explained in paragraph 4.12.7.

4.12.5.1 In the *InfoQuest FP main* window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

Since discriminant analysis works on user-delineated groups, the comparison should contain groups (see 4.1.11).

4.12.5.2 Select **PhenoTest** in the *Experiments* panel and press the  button of **PhenoTest** to display the image in the *Experiment data* panel.

4.12.5.3 Select **Dimensioning > Principal Components Analysis** or .

The *Principal Components Analysis* dialog box (Figure 4-124) allows a number of choices to be made under **Entries** and **Characters**, which are described under PCA (4.12.4). These choices also apply for discriminant analysis. However, the **Divide by variance** option under **Characters** makes no difference whether it is enabled or disabled for discriminant analysis.

The following two options are available for discriminant analysis: **Discriminants (without variance)**, and **Discriminants (with variance)**. If you select “with vari-

ance”, each character is divided by its variance. In order to understand what this implies, consider Figure 4-127.

This example shows two groups, 1 (red) and 2 (green), that are separated by two characters, A and B. On the average, group 1 is less positive both for characters A and B. Character A seems to be better discriminating between the two groups than character B, because the centers of the groups are lying further from each other in case of character A. However, if the internal spread of groups are considered, then the groups are found much more coherent for character B, which may render this character at least as much value for discriminating as character A. In a non-corrected discriminant analysis, character A will account for most of the discrimination, just by the fact that the centers of the groups are more distant. This is the case in option **Discriminants (without variance)**. When the characters are divided by the variances of the groups, the internal spread is compensated for, and character B will become at least as important as character A. This is achieved with option **Discriminants (with variance)**.

4.12.5.4 Select **Discriminants (with variance)** and **<OK>**.

The resulting window is identical to the *Principal components analysis* window described before (Figure 4-124), and the same features apply.

Similar as for a PCA (see 4.12.4.17), if you started the discriminant analysis from a composite data set, you can order the characters according to the selected discriminant in the underlying *Comparison* window.

4.12.5.5 In the *Principal components analysis* window, first determine the discriminant that best separates two groups you have in mind. You can examine the discriminants by selecting them in the *Components* panel and **Layout > Use component as Y-axis** (or X-axis).

4.12.5.6 Select that discriminant in the *Components* panel and select **Characters > Order characters by component**. The characters are now ordered by the selected discriminant in the *Comparison* window.

4.12.6 Self-organizing maps

A self-organizing map (SOM, also called Kohonen map) is a neural network that classifies entries in a two-dimensional space (map) according to their likeness. The technique which is used for grouping, i.e. the

training of a neural network, is completely different from all previously described methods. SOMs therefore provide an interesting addition to conventional grouping methods such as cluster analysis, principal component analysis and related techniques. Also, similar as in PCA, a SOM can start from the characters as input, thus avoiding the choice of one or another similarity coefficient. Unlike PCA, the distance between entries on the map is not in proportion to the taxonomic distance between the entries. Rather, a SOM contains areas of high distance and areas of high similarity. Such areas can be visualized by different shading, for example when a darker shading is used in proportion to the distance in the SOM.

When the similarity values with all of the other entries of a comparison are considered as the character set, a SOM can also be applied on similarity matrices, which makes the technique also suitable for grouping of electrophoresis patterns that are compared pair by pair using a band matching coefficient such as Dice.

To calculate a self-organizing map based on character data, use for example the character set **FAME** in **DemoBase**.

4.12.6.1 In the *InfoQuest FP main* window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

4.12.6.2 Select **FAME** in the *Experiments* panel and *Dimensioning > Self-organizing map* (the command *Dimensioning > Self-organizing map (similarities)* is to calculate the SOM from the similarity matrix).

An input box asks to enter the map size. This is the number of nodes of the neural network in each direction. For the default size 10, a neural network containing 10x6 nodes is generated. The larger the map is taken, the longer the training takes. Note that the optimal size of the map depends on the number of entries compared. For a small number of entries, a small map size will usually provide better results.

4.12.6.3 Enter 6 as map size and press **<OK>**.

4.12.6.4 The SOM is calculated and shown (Figure 4-128). Areas of high similarity are black. Selected entries in the parent *Comparison* window are also selected on the map. Note that the SOM as shown in Figure 4-128 will not necessarily correspond to the one you have calculated.

4.12.6.5 To show the information of a particular entry in the SOM, right-click on the entry and select **Edit data-base fields**.

4.12.6.6 You can (un)select entries on the SOM by left-clicking on an entry while pressing the CTRL key, or groups of entries by left-clicking and moving the mouse while pressing the SHIFT key.

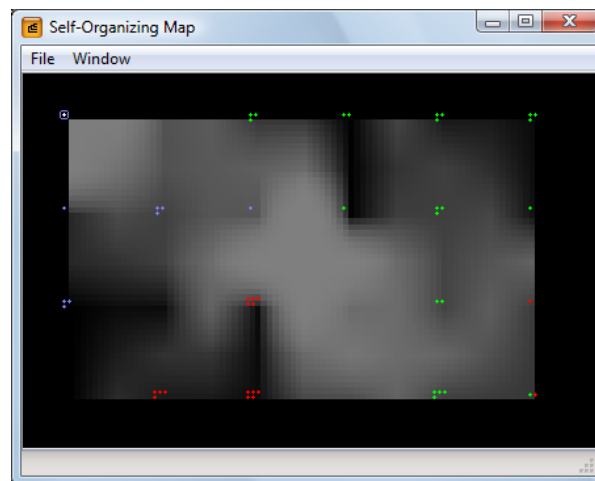




Figure 4-128. Self-organizing map calculated from the data set **FAME** in the **DemoBase** of **InfoQuest FP**.

NOTE: When a SOM is calculated on fingerprint type data, the densitometric curves are used as character data sets for training of the SOM.

A SOM is automatically saved along with its parent *Comparison* window. It is possible to add entries to an existing SOM or remove entries from it. The feature to add entries to an existing SOM is an interesting alternative way of identifying new entries. Added entries are placed in a frame of known database entries in the SOM, and in this way, identifying is just looking at the groups they are joining.

4.12.6.7 If you want to add entries to an existing SOM, you can select new entries in the *InfoQuest FP main* window and copy them to the clipboard using **Edit >**

Copy selection or .

4.12.6.8 In the *Comparison* window, select **Edit > Paste selection** or press . The new entries are placed in the *Comparison* window and in the *Self-Organizing Map* window.

NOTES:

(1) An identification based upon a self-organizing map is only reliable if the new entries belong to one of the groups the SOM is based upon. A SOM will always produce a "positive" identification: an unknown profile will **always** find a place in the SOM, i.e. the cell having the highest similarity with the new entry. If, after adding a new entry to a SOM, the entry falls next to a known entry of that SOM, this means only that the new entry has the highest similarity with that particular cell compared to the other cells; it does **not** mean that it is highly related to that entry. Hence, identification based upon a SOM is only recommended

if you are sure the unknown entries belong to one of the groups composing the SOM.

(2) Since no new cells can be created in a SOM, one should never add new entries which are known to constitute a group that is not represented in the SOM.

4.12.6.9 To delete entries from a SOM, select the entries and in the underlying *Comparison* window, select **Edit > Cut selection**.

4.12.6.10 Close the *Self-Organizing Map* window with **File > Exit**.

4.12.6.11 To create a self-organizing map from a similarity matrix obtained after cluster analysis, select **Dimensioning > Self-organizing map (similarities)**.

In this case, the result of the SOM is based on similarity values of the entries with each other and hence is dependent on the similarity coefficient used, and the tolerance and optimization settings in case of fingerprint types. Obviously, this method only works if a cluster analysis of the selected experiment is available. If not, first create a cluster analysis with **Clustering > Calculate > Cluster analysis (similarity matrix)**.

4.12.6.12 The SOM can be printed with the **File > Print** command, or exported via the clipboard as enhanced metafile using **File > Copy to clipboard**.

In these cases, the map colors are inverted, i.e. white corresponds with areas of high similarity, whereas darker shading corresponds with areas of low similarity.

A SOM is saved along with a comparison. In order to display a previously calculated SOM in a comparison, click on the experiment type in the *Experiments* panel and select **Dimensioning > Show map**.

4.12.7 Multivariate analysis of variance (MANOVA) and discriminant analysis


Multivariate Analysis Of Variance (MANOVA) is a statistical technique which allows the significance of user-delineated groups to be calculated. Since it is extremely difficult to prove that delineated groups are significant, statistical methods usually are based on the reverse approach, i.e. to prove that the chance (likelihood) to obtain equally good separations with randomly generated groups approaches zero. In addition, a statistical technique related to PCA, discriminant analysis, allows the determination of characters that are responsible for the separation of the delineated groups.

The MANOVA technique only applies to character types and composite data sets in InfoQuest FP. If you want to find the discriminating characters for a different experiment type, you should first create a composite data set containing that experiment.

MANOVA cannot be applied to incomplete data sets. In other words, all characters must be filled in for each entry. In case of “open” character types (in which the character set may grow dynamically), absent values should be considered as zero.

4.12.7.1 In the *InfoQuest FP* main window with **DemoBase** loaded, open comparison **All**, or create a comparison with all entries except those defined as **STANDARD** (see 4.1.3.2 to 4.1.3.4).

Since MANOVA and discriminant analysis work on user-delineated groups, the comparison should contain groups (see 4.1.11).

4.12.7.2 Select **PhenoTest** in the *Experiments* panel and press the  button of **PhenoTest** to display the image in the *Experiment data* panel.

4.12.7.3 Select **Groups > Multivariate Analysis of Variance**.

4.12.7.4 The program now pops up a MANOVA dialog box (Figure 4-129). The meaning of the variances (diagonal elements) is similar to the variances explained for Discriminant Analysis (4.12.5). The *Covariances* relate to the possibility of the discriminant analysis to explore correlations between characters in order to achieve a better discrimination. Most statistical approaches assume that the covariance is accounted for, however, its use becomes dangerous in case the number of characters is close to, or larger than the number of entries studied. In such cases, the result of the discriminant analysis could be that the delineated groups are perfectly separated. To avoid such unrealistic separations, you should only allow the program to account for the covariance when the number of entries is significantly larger than the number of characters.

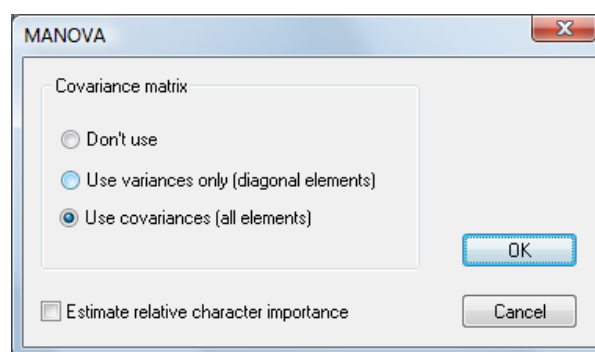


Figure 4-129. The MANOVA dialog box.

4.12.7.5 A second option is *Estimate relative character importance*. When this feature is applied, the program will repeat the discriminant analysis, each time leaving out one character. The quality of the separation when a character is left out is then compared to the quality when the character is not left out, and this is a direct measure for the importance of that character. Obviously,

the calculations take much longer when the discriminant analysis is to be calculated p times, p being the number of characters.

4.12.7.6 Select *Don't use* under *Covariance matrix*, enable the option to calculate the relative character importance and press <OK>.

If one or more characters are identical for all the entries, this will be reported in a message box and such characters will be left out from the discriminant analysis. The resulting *MANOVA & discriminant analysis* window is shown in Figure 4-130. The window is divided in three dockable panels:

The *Discriminants* panel (top panel in default configuration) shows the relative discriminatory value of the characters for each of the discriminants. A character can have a contribution to the discrimination in the positive sense (green) or in the negative sense (purple). The larger the bar, the greater the contribution, irrespective

of the sense. If character contributions have a different sense, it means that the one character will be positive in the groups where the other character will be negative and *vice versa*.

The relative importance for each character is shown as a red line, right from the character name.

Note that the total number of discriminants will always be the number of groups less one. For two groups, there is only one discriminant; for three groups, there are two discriminants, etc.

4.12.7.7 If there are more than two discriminants, you can scroll through the list of discriminants.

The first discriminant is always the most important, i.e. it accounts for most of the discrimination; the second discriminant is the second most important, etc. The percentage discrimination of a discriminant is shown in bold (left). The sum of the percentages equals 100.

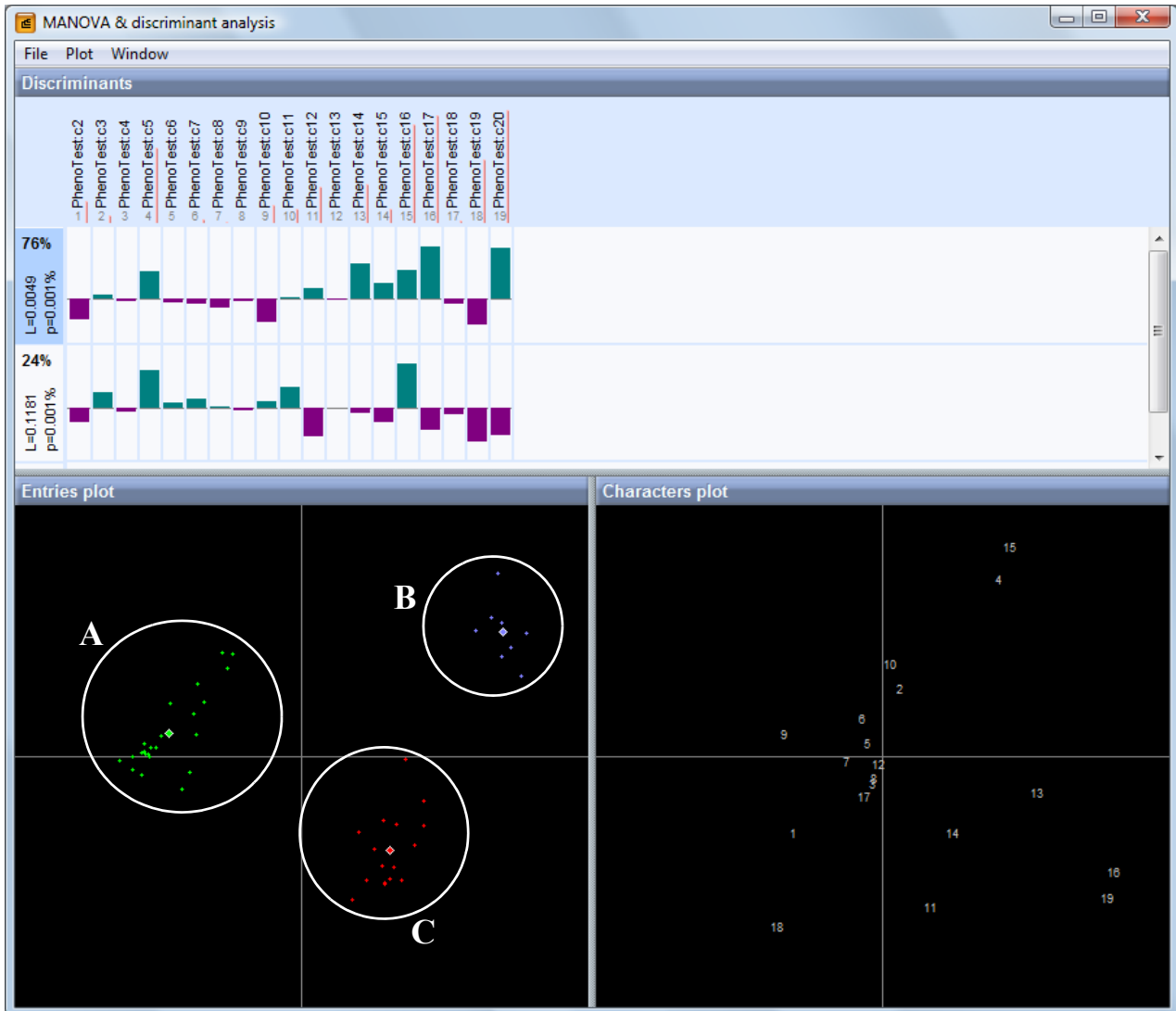


Figure 4-130. *MANOVA & discriminant analysis* window. The circles delineating groups A, B, and C are added to this figure to illustrate the interpretation of discriminant analysis.

In addition, the parameter L (Wilk's Lambda likelihood ratio test) predicts the likelihood of the obtained discrimination in the assumption that the groups are drawn from the same population. If L is low, the entries of the different groups are likely to be drawn from different populations, in other words, the existence of the groups is justified. The parameter p is the probability that a random subdivision in groups would yield the same degree of discrimination.

The *Entries plot* panel (left lower panel in default configuration) maps the entries on the first two discriminants (first = X axis, second = Y axis). On this image you can see that the X axis accounts for most of the discrimination (Figure 4-130).

The *Characters plot* panel (right lower panel in default configuration) maps the characters on the entry groups. To interpret this very informative panel, you should inspect it together with the *Entries plot* panel. The more distant a group occurs from the center along a discriminant axis, the better it may be characterized by one or more characters. These characters can be found in the *Characters plot* panel, shown by their number. Characters are very positive for the group if they fall in the same direction of the axis as the group; if they occur in the opposite direction, they are very negative for the group. The further a group and one or more characters occur in either direction of an axis, the more pronounced these characters are either positive or negative for that group. For example, group A occurs in the negative half of the X axis (first discriminant), whereas group B is the most positive group on this axis. Characters 16 and 19 (pronounced positive position) discriminate group C from group A in that they are much more positive for group C members than for group A members. Another example: group B is positive on both the X and Y discriminant, whereas groups A and C are negative on either the X or Y axis. From this, one can conclude that characters 4 and 15 discriminate group B from both groups A and C. The rhomb in the center of each group in the left panel is the average position of the group.

4.12.7.8 The third discriminant (if available) can be plotted on the images by selecting it in the *Discriminants* panel and selecting *Plot > Use discriminant as X axis* or *Plot > Use discriminant as Y axis*.

4.12.7.9 The menu *Plot > Order characters by magnitude* allows the characters to be ordered by their contribution on the selected discriminant.

4.12.7.10 The menu *Plot > Order characters by importance* allows the characters to be ordered by their relative importance factors (4.12.7.5).

4.12.7.11 It is possible to select an entry from the left panel with CTRL+left mouse click. Selected entries are encircled in blue.

4.12.7.12 To select several entries at a time, hold down the SHIFT key while dragging the mouse in the left panel. All entries included in the rectangle will become selected.

4.12.7.13 By double-clicking on an entry, its *Entry edit* window is popped up.

4.12.7.14 With *Plot > Show groups using colors*, you can toggle between the default color mode and the non-color mode where groups are represented using symbols. In the non-color mode, non-selected entries are shown in yellow, whereas selected entries are shown in blue.

The various results of a MANOVA analysis can be printed or exported as enhanced metafile to the clipboard for further processing in other packages:

4.12.7.15 Use *File > Print report* to print a detailed numerical report of all characters and their contribution along the discriminants. Similarly, *File > Export report* is used to export this report to the clipboard, tab-delimited or space-delineated.

4.12.7.16 Use *File > Print discriminants* to print the upper graphical panel, representing the selected discriminants and the relative importance of the characters shown as bar graphs. *File > Copy discriminants to clipboard* is to export this report to the clipboard as enhanced metafile.

4.12.7.17 Use *File > Print correspondence plot* to print the two two-dimensional plots, representing the entries (left) and characters (right) plotted along two discriminants. *File > Copy correspondence plot to clipboard* is to export this report to the clipboard as enhanced metafile.

4.13 Chart and statistics tools

4.13.1 Introduction

A number of simple chart tools are available in InfoQuest FP to apply to the database information fields or to character data for the entries in a comparison. InfoQuest FP also offers the possibility to perform some basic statistic analysis on the entries and variables used in a chart. Given the large variety of information and character types InfoQuest FP can contain, there are many different types of charts that can be displayed, depending on the type of the variable(s) to present. For each chart one or more standard statistical tests are implemented. The next paragraphs are intended to provide some information on the terminology (4.13.2) and the mathematical background (4.13.3) of these tests.

The use of the chart and statistics tools is described in paragraphs 4.13.4 to 4.13.11.

4.13.2 Basic terminology

4.13.2.1 Literature

This manual is not aimed to be an introduction to basic statistics. For more detailed literature, we refer to the following handbooks:

- Press W., Teukolsky S.A., Vetterling W.T., Flannery B.P., 'Numerical recipes in C', Cambridge University Press, Cambridge.
- Sheskin D.J., 'Handbook of parametric and nonparametric statistical procedures', CRC Press, Boca Raton.
- Zwillinger D., Kokoska S., 'Standard probability and statistics tables and formulae', Chapman & Hall/CRC, Boca Raton.

4.13.2.2 Application of statistic tests

In general terms, the application of a statistic test can be outlined as follows:

- Make a proposition that will be referred to as the **null-hypothesis**. *Statistical tests cannot be employed for proving that a certain hypothesis is true, but only for proving that all alternative hypotheses can be rejected.* Therefore, the null-hypothesis is what one wants to reject.

- Determine what **statistic** will be used. A statistic is a value calculated from the data set by means of some formula and which is sensitive to the null-hypothesis that will be tested for.

- If the null-hypothesis is true, the probability function of the statistic is known.

- If the statistic is located on an unfavorable position in the probability function, i.e. if its probability is very small, the null-hypothesis can be rejected. The opposite is not true: the null-hypothesis cannot be accepted as fulfilled if the statistic has a favorable location in the probability distribution.

Note that not all tests are applicable in all situations. There may be restrictions to e.g. the amount of data in the sample, or to some basic properties of the data set. These restrictions are mentioned where the tests are described.

4.13.2.3 Parametric or non-parametric tests

Parametric tests basically suppose that the data are distributed normally; they generally make use of the values for the mean and the standard deviation.

Non-parametric tests are commonly based on a ranking of the data. These ranks are distributed uniformly, hence these tests are independent of any underlying distribution. The price to pay is that an estimate of the significance is more complicated and often relies on approximations. These methods also generally lose some strength because they lose some information about the data. In comparison with parametric tests they require more data to come to an equally significant result.

For these tests the values of the data points are usually replaced by their rank among the sample. The data points are ordered, the lowest in order is assigned rank one and the highest in order is assigned the rank that equals the total sample size.

If some of the data points originally have the same values, they can be assigned the mean of the ranks (called 'tie rank') they would have had if they were different. The sum of the assigned ranks is always equal to the total sample size.

4.13.2.4 Categorical or quantitative data

Within the chart tool, a distinction between three types of variables is made.

• **Categorical variable:** this type of variable divides a sample into separate categories or classes. Examples are database fields like e.g. genus, species, etc. Also intervals of quantitative variables can be treated as categorical data.

• **Quantitative variable:** this type of variable can take either continuous numerical values or binary values. Character data are a typical example for this type of variable. Continuous numerical values can be converted into interval data if necessary. If this option is chosen, an interval size can be specified.

• **Date variable:** a variable containing a date. This variable can be converted into interval data, which means that it can be interpreted as either a categorical variable or a quantitative variable. When converting into interval data, you can choose to group the dates by day, week, month, quarter or year.

With combinations of these variables several types of plots can be created, based upon:

One variable:

- *Bar graph:* for a single categorical variable
- *1-D numerical distribution:* for one quantitative variable

Two variables:

- *Contingency table:* for two categorical variables
- *2-D scatterplot:* for two quantitative variables
- *2-D ANOVA plot:* for one categorical and one quantitative variable.

Three variables:

- *3-D scatterplot:* for three quantitative variables

For an overview of graph types and associated tests for one and two variables, see Table 4-1.

Some types of plots can be extended in the sense that they can display information from an additional categorical variable by means of a color code. These plots are the 2-D scatterplot, the 3-D scatterplot, the 2-D ANOVA plot and the 1-D numerical distribution.

	Categorical	Quantitative
---	Bar graph (4.13.3.1) <i>Chi square test for equal category sizes</i>	1-D numerical distribution (4.13.3.4) <i>Kolmogorov-Smirnov test for normality</i>
Categorical	Contingency table (4.13.3.3) <i>Chi square test for contingency tables</i>	2-D ANOVA plot (4.13.3.6) <i>See Table 4-3</i>
Quantitative	2-D ANOVA plot (4.13.3.6) <i>See Table 4-3</i>	2-D scatterplot (4.13.3.5) <i>See Table 4-2</i>

Table 4-1. Schematic representation of variable types and corresponding graphs and tests for one and two variables.

	Parametric	Non-parametric
Means	<i>T test (4.13.3.5.1)</i>	<i>Wilcoxon signed-rank test (4.13.3.5.2)</i>
Correlations	<i>Pearson correlation test (4.13.3.5.3)</i>	<i>Spearman rank-order correlation test (4.13.3.5.4)</i>

Table 4-2. Overview of tests associated with 2-D scatterplots.

	Parametric	Non-parametric
2 categories	<i>T test (4.13.3.6.1)</i>	<i>Mann-Whitney test (4.13.3.6.2)</i>
>2 categories	<i>F test (4.13.3.6.4)</i>	<i>Kruskal-Wallis (4.13.3.6.4)</i>

Table 4-3. Overview of tests associated with 2-D ANOVA plots.

4.13.3 Charts and statistics

4.13.3.1 Bar graph: Chi square test for equal category sizes

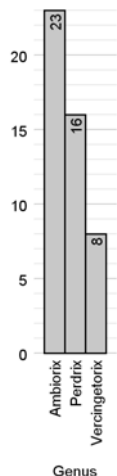


Figure 4-131. Example of a bar graph.

For a bar graph displaying the number of entries for a categorical variable, one typically likes to know if there are significant differences in the number of entries per category. Hence, the null-hypothesis is that all categories have an equal number of entries.

If this null-hypothesis holds, the *expected average count per category* (N_e) can be calculated as the total number of entries divided by the number of categories,

$N_e = N/n$, with N the total number of entries and n the number of categories. The *chi square* statistic is calculated from the values for the expected average count (N_e) and the observed entries per category (N_{oi}),

$\chi^2 = \sum_{i=1}^n \left(\frac{[N_{oi} - N_e]^2}{N_e} \right)$, with n the number of categories.

If the null-hypothesis is true and under certain conditions (see the note below) this statistic approximately follows a chi square distribution with $n-1$ *degrees of freedom*. The *p-value* that is returned gives the probability that the statistic is at least as high as the observed one. If the *p-value* is low, the null-hypothesis can be rejected. The *significance s* of the test is calculated as the complement of the *p-value*,

$$s = 100 \times (1 - p).$$

The values for these parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.5.

Chi square: 7.191 (2 degrees of freedom)
P value= 0.027440
Significance= 97.2560%

Expected average count per category: 15.67

Figure 4-132. Example of a test report for the chi square test for equal categorical sizes applied on a bar graph as shown in Figure 4-131.

NOTE: This test should not be used if the expected average count per category is less than 5. If this is the case, consider combining categories in order to increase the expected average count.

4.13.3.2 Bar graph: Simpson and Shannon Weiner indices of diversity

A commonly asked question about a number of entries occurring in different categories, is how they are distributed. Two widely used coefficients to measure the diversity are the *Shannon-Weiner index of diversity* and *Simpson's index of diversity*. Both coefficients take into account the *diversity*, i.e. the number of categories present in the sampled population, as well as the *equitability*, i.e. the evenness of the distribution of entries over the different categories.

Simpson's index of diversity is defined as the probability that two consecutive entries will belong to different categories. Given K categories present in a sampled population, the probability of sampling category i twice consecutively is as follows (n_i is the number of entries in category i):

$$P_i = \frac{n_i(n_i - 1)}{\sum_{j=1}^K n_j(n_j - 1)}$$

The probability of sampling any two samples of the same category is given by $P = \sum_{i=1}^K P_i$. Hence, the

probability D of sampling two different categories is $D = 1 - P$, which is Simpson's index of diversity.

For a sampled population of N entries belonging to K categories, the Shannon-Weiner index of diversity is calculated as follows (n_i is the number of entries in category i):

$$H = - \sum_{i=1}^K \frac{n_i}{N} \ln \left(\frac{n_i}{N} \right)$$

4.13.3.3 Contingency table: chi square test for contingency tables

A contingency table contains information on the association between two categorical variables. Each cell contains the number of entries for a specific combination of row and column categories. For this kind of representation of the data, the obvious question is usually if the information contained in the rows and columns is correlated or not. The null-hypothesis is that there is no association between the rows and columns.

Cell counts

	0	0	0	1	1.250
	0	4	4	3	1.750
	1	0	6	10	2.250
	0	3	7	8	2.750
1.250		1.750	2.250	2.750	

c4

Figure 4-133. Example of a contingency table where intervals of a numerical variable are used to create categories.

If the null-hypothesis is true, the expected count per cell can be calculated. Therefore, we need to know the total number of cells n in the table, $n = n_i n_j$ with n_i the number of rows and n_j the number of columns. The summed numbers of counts in each row and column are called the *marginal row counts* (e.g. N_{rowi} stands for the marginal row count of row i) and *marginal column counts* (N_{colj}). If there is no association between rows and columns, the expected cell count n_{ij} for a cell on row i and column j can be calculated as $n_{ij} = N_{rowi} N_{colj} / N$, with N the total number of entries.

Using these expected cell counts (n_{ij}) and the observed counts per cell (N_{oij}), a *chi square* statistic is calculated,

$$\chi^2 = \sum_{i=1, j=1}^{n_i, n_j} \left(\frac{[N_{oij} - n_{ij}]^2}{n_{ij}} \right),$$

with n_i the number of rows and n_j the number of columns.

If the null-hypothesis is true and under certain conditions (see note below), this statistic approximately follows a chi square distribution with $N - n_i - n_j + 1$ degrees of freedom. The *p-value* that is returned gives the probability that the statistic is at least as high as the observed one. If the *p-value* is low, the null-hypothesis can be rejected. The *significance s* of the test can be calculated as the complement of the *p-value*, $s = 100 \times (1 - p)$.

In case there is a significant association, its strength can be expressed using *Cramer's V*. The formula is

$V = \sqrt{\chi^2 / [N \min(n_i - 1, n_j - 1)]}$, with χ^2 the value for the statistic, N the total number of entries, n_i the number of rows and n_j the number of columns. This gives a value between 0%, in case there is no association, and 100%, in case there is a perfect association. Cramer's *V* can be used to compare the strengths of different associations.

Values for the various parameters can be found in the test report. The marginal column and row counts are expressed in absolute counts and relative to the total number of counts in the table. How such a chart and report can be created is explained in section 4.13.6.

Chi square: 10.337 (9 degrees of freedom)
P value= 0.323868
Significance= 67.6132%

Cramer's V: 27.08%

Total count: 47
Average cell count: 2.94

Marginal column counts:

1.250	1	2.13%
1.750	7	14.89%
2.250	17	36.17%
2.750	22	46.81%

Marginal row counts:

1.250	1	2.13%
1.750	11	23.40%
2.250	17	36.17%
2.750	18	38.30%

Figure 4-134. Example of a test report for the chi square test for contingency tables like shown in Figure 4-133.

The contingency table can be displayed showing the residuals for the cells. The residual is a measure for the deviation from the expected number of counts in that cell and is calculated as $[N_{oij} - n_{ij}] / \sqrt{n_{ij}}$, with N_{oij} the observed cell count and n_{ij} the expected cell count.

NOTE: This test should not be used if the expected average count per category is less than 5. If this is the case, consider combining categories in order to increase the expected average count. In practice, this also means that there should be no empty rows or columns in the contingency table.

4.13.3.4 1-D numerical distribution function: Kolmogorov-Smirnov test for normality

For a sample containing a single quantitative variable, an often recurring question is if it is normally distrib-

uted or not. In this case the null-hypothesis is that the sample is drawn from a normal distribution. The *mean value* $\langle x \rangle$ and *corrected standard deviation*

$$\sqrt{\sum_{i=1}^n (x_i - \langle x \rangle)^2 / (n - 1)}$$

(with x_i the observations

and n the sample size) are calculated from the sample and are used to determine a normal distribution that can be used as a model (further referred to as model normal distribution) for the underlying distribution of the sample if the null-hypothesis holds.

The Kolmogorov-Smirnov test for normality is applied to test how different the cumulative distribution of the sample is from the cumulative distribution of the model normal distribution. For a sample where each observation is associated with a single number of events, the cumulative distribution $F(x_j)$ gives for each observation (x_j) the total number of events associated to all observations in the sample that are smaller or equal to the observation (x_j). Hence, the cumulative distribution gives at each observation the probability of obtaining that observation or a lower one.

The test statistic is the *maximum difference* in absolute value between the cumulative distribution of the sample and the cumulative distribution of the model normal distribution. In case the null-hypothesis is true and under certain conditions (see note below), the distribution function for this statistic can be calculated approximately. The *p-value* gives the probability that the statistic obtains a higher value than the observed one. If the p-value is low, the null hypothesis can be rejected. The *significance* of the test can be calculated as the complement of the p-value, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.10.

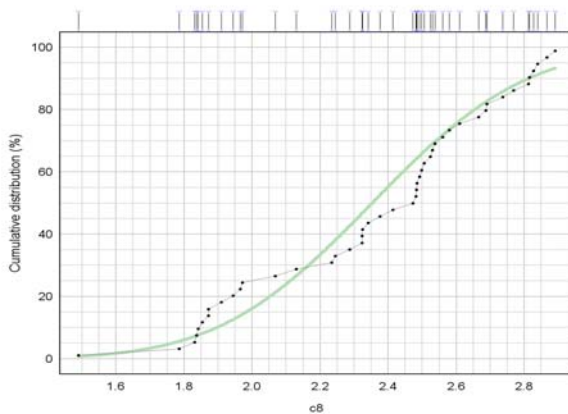


Figure 4-135. Example of a 1-D numerical distribution and model normal distribution.

Mean: 2.352766
 Corrected standard deviation: 0.357346
 Maximum difference: 0.1413
 P value= 0.282993
 Significance= 71.7007%

Figure 4-136. Example of a test report for the Kolmogorov-Smirnov test for normality applied to a 1-D numerical distribution as shown in Figure 4-135.

NOTES:

(1) The Kolmogorov-Smirnov test for normality should not be used if the number of data points is smaller than 4. The test becomes more accurate if more data points are used.

(2) This test cannot be used to prove that a sample follows a normal distribution, since its aim is only to reject the null-hypothesis with a certain level of significance.

4.13.3.5 2-D scatterplot

Scatterplots contain information on two quantitative variables that are obtained for a set of entries. The position of each dot on the plot is determined by the observations. A scatterplot is dealing with **paired** data since a specific pair of observations characterizes each entry that is represented in the plot.

For this kind of plot one could ask (1) if the means are significantly different or (2) if there is any correlation between the two variables. For both questions, there is a parametric and a non-parametric test available.

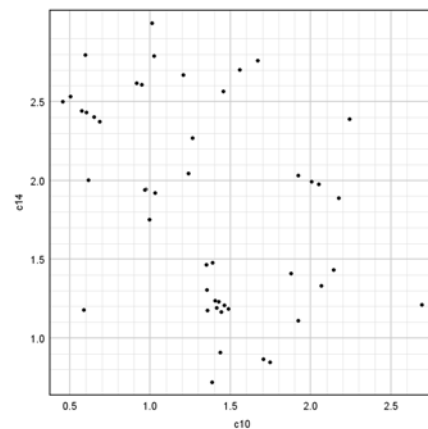


Figure 4-137. Example of a 2-D scatterplot.

4.13.3.5.1 Parametric test for means: T test

The null-hypothesis is that the two samples have the same mean values. Assume the sample observations are

x_i and y_i ($i=1, \dots, n$), with $\langle x \rangle$ and $\langle y \rangle$ the respective mean values and $s_x = \sqrt{\sum_{i=1}^n (x_i - \langle x \rangle)^2 / (n-1)}$ and $s_y = \sqrt{\sum_{i=1}^n (y_i - \langle y \rangle)^2 / (n-1)}$ the corrected variances.

For paired data, it is generally not guaranteed that all entries have a completely independent pair of observations. The test statistic should be corrected for the influences this may have on the variance of the observations. Therefore, the *corrected covariance* Cov of the sample,

$$Cov(x, y) = \left(\sum_{i=1}^n (x_i - \langle x \rangle)(y_i - \langle y \rangle) \right) / (n-1),$$

is taken into account. The sample variance can be expressed by means of the *pooled corrected standard deviation* s_d . In this case, s_d can be calculated as

$$s_d = \sqrt{(s_x^2 + s_y^2 - 2Cov(x, y)) / n}.$$

A statistic is defined as $T = (\langle x \rangle - \langle y \rangle) / s_d$. If the null-hypothesis holds and under certain conditions (see note below) this statistic follows a t distribution with $n-1$ degrees of freedom. The *p-value* gives the probability that the statistic indeed has the observed value or higher. If the *p-value* is small, the null-hypothesis can be rejected. The *significance* of the test is calculated as the complement of the *p-value*, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.7.

```

Mean values:
c10    1.3396
c14    1.8512
Corrected variances:
c10    0.2870
c14    0.4246
Corrected covariance = -0.1476
Pooled corrected standard deviation =
0.1464

T = -3.495 (46 degrees of freedom)
P value= 0.001060
Significance= 99.8940%

```

Figure 4-138. Example of a test report for the T test applied to a 2-D scatterplot like in Figure 4-137.

NOTES:

(1) This test should not be used if the data points are not normally distributed. In this case the Wilcoxon signed-rank test can be used.

(2) This test should not be used if the variances of the two samples are not the same.

4.13.3.5.2 Non-parametric test for means: Wilcoxon signed-rank test

The null-hypothesis is that the two samples have the same mean values. Assume the sample observations are x_i and y_i ($i=1, \dots, n$). The absolute values of the differences of these observations $|d_i| = |x_i - y_i|$ are ranked (zero values are eliminated from the analysis). As a first step, these ranks are assigned to rank variables R_i . Afterwards, these R_i get the sign of corresponding d_i . These two steps turn the R_i into ranks of positive or negative differences. The *sum of ranks of positive differences* (sum of all positive R_i) and the *sum of ranks of negative differences* (absolute value of the sum of all negative R_i) are determined and the smallest of these sums is called the Wilcoxon T test statistic.

If the null-hypothesis holds, the expected value for T is $n(n-1)/4$ (with n the number of pairs of observations), while the expected standard deviation on T is $\sqrt{n(n+1)(2n+1)/24}$. Hence, if the null-hypothesis holds and under certain conditions (see note below) the statistic defined as $(T - [n(n-1)/4]) / \sqrt{n(n+1)(2n+1)/24}$ approximately follows a normal distribution. The *p-value* gives the probability that the statistic is at least as high as the observed one. If the *p-value* is low, the null hypothesis can be rejected. The *significance* of the test is calculated as the complement of the *p-value*, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.7.

```

Sum of ranks of positive differences= 303.0
Sum of ranks of negative differences= 825.0

P value= 0.005746 (Normal approximation)
Significance= 99.4254%

```

Figure 4-139. Example of a test report for the Wilcoxon signed-rank test applied to a 2-D scatterplot like in Figure 4-137.

NOTES:

(1) This test should not be used if the population distribution is not symmetric.

(2) The approximation by using a normal distribution is only valid if the sample contains more than 20 observations.

4.13.3.5.3 Parametric test for correlations: Pearson correlation test

The null hypothesis is that there is no linear relationship between the sample variables. Assume the observations in the sample are x_i and y_i ($i=1, \dots, n$), with $\langle x \rangle$ and $\langle y \rangle$

the mean values, $s_x = \sum_{i=1}^n (x_i - \langle x \rangle)^2 / n$ and

$s_y = \sum_{i=1}^n (y_i - \langle y \rangle)^2 / n$ the variances and

$Cov(x, y) = \left(\sum_{i=1}^n (x_i - \langle x \rangle)(y_i - \langle y \rangle) \right) / n$ the

covariance of the sample. Pearson's correlation r is calculated as $r = Cov(x, y) / \sqrt{s_x s_y}$.

If the null-hypothesis holds and under certain conditions (see note below) the statistic defined as

$|r| \sqrt{n-2} / \sqrt{1-r^2}$ approximately follows a t distribution with $n-2$ degrees of freedom. Since $|r|$ is used to

calculate the statistic, the p -value can be calculated using a single tail of the t distribution. The p -value gives the probability that the statistic obtains a value at least as high as the observed one. The *significance* of the test is calculated as the complement of the p -value, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.7.

Mean values:
 c10 1.3396
 c14 1.8512
 Variances:
 c10 0.2809
 c14 0.4156
 Covariance= -0.1445
 Pearson correlation= -42.288%
 P value (single tail)= 0.001531 (T test

approximation)
 Significance= 99.8469%

Figure 4-140. Example of a test report for the Pearson correlation test applied to a 2-D scatterplot like in Figure 4-137.

In case there is a significant linear correlation, Pearson's r can be used to indicate its strength. A positive value for Pearson's r is associated with a positive correlation and would result in a regression line with positive slope. A negative value for Pearson's r is associated with a negative correlation and would result in a regression line with negative slope.

If the samples contain less than 30 observations, an alternative way for testing the null-hypothesis is offered by Monte-Carlo simulations. To do this, 10.000 samples with n pairs of randomly distributed observations are created. For each of these samples, a value for the statistic is obtained and is compared to the observed value. The p -value from the simulations is determined by the number of times the simulations give a larger value for the statistic than the value observed in the real sample. Also here, the *significance* is calculated as $s = 100 \times (1 - p)$. The results for the simulated p -value and significance also appear in the test report.

NOTE: This test should not be used if the distributions of x_i or y_i have strong wings or if they are not normally distributed. However, this test is acceptable for sufficiently large samples.

4.13.3.5.4 Non-parametric test for correlations: Spearman rank-order correlation test

The null-hypothesis is that there is no linear correlation between the sample rank variables, or equivalently that there is no monotonic relation between the sample variables. The sample observations x_i and y_i ($i=1, \dots, n$) are replaced by their rank after ordering them from smallest to largest. This results in a sample of ranks R_i and S_i ($i=1, \dots, n$). The Spearman rank-order correlation coefficient is defined as $r_s = Cov(R, S) / \sqrt{s_R s_S}$, with

$s_R = \sum_{i=1}^n (R_i - \langle R \rangle)^2 / n$ and

$s_S = \sum_{i=1}^n (S_i - \langle S \rangle)^2 / n$ the rank variances,

$Cov(R, S) = \sum_{i=1}^n (R_i - \langle R \rangle)(S_i - \langle S \rangle) / n$ the rank

covariance and $\langle R \rangle$ and $\langle S \rangle$ the rank mean values of the rank variables R_i and S_i respectively.

The null-hypothesis can be tested using the statistic $|r_s| \sqrt{n-2} / \sqrt{1-r_s^2}$. If the null-hypothesis holds, this statistic approximately follows a t distribution with $n-2$ degrees of freedom. Since $|r_s|$ is used to calculate the statistic, the *p-value* can be calculated using a single tail of the t distribution. The p-value gives the probability that the statistic obtains a value at least as high as the observed one. In this case, a single tail test is performed. The *significance* of the test is calculated as the complement of the p-value, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.7.

```

Rank mean values:
  c10  24.0000
  c14  24.0000
Rank variances:
  c10  184.0000
  c14  184.0000
Rank covariance= -77.1489

Spearman rank-order correlation= -41.929%
P value (single tail)= 0.001675 (T test
approximation)
Significance= 99.8325%

```

Figure 4-141. Example of a test report for the Spearman rank-order correlation test applied to a 2-D scatterplot like in Figure 4-137.

If the samples contain less than 30 observations, an alternative way for testing the null-hypothesis is offered by Monte-Carlo simulations. To do this, 10.000 samples with n pairs randomly distributed observations are created. For each of these samples, a value for the statistic is obtained and is compared to the observed value. The *p-value* from the simulations is determined by the number of times the simulations give a larger value for the statistic than the value observed in the real sample. Also here, the *significance* is calculated as $s = 100 \times (1 - p)$. The results for the simulated p-value and significance also appear in the test report.

4.13.3.6 ANOVA plot

This kind of plot presents a categorical and quantitative variable. The categorical variable splits the sample in a number of groups while the quantitative variable describes a distribution within each group. This kind of data is called **unpaired**.

A typical question is whether the groups have the same average for the quantitative variable. In case there are only two groups for the categorical variable, the parametric T test or the non-parametric Mann-Whitney test can be applied. If there are three or more groups for the categorical variable, the parametric F test or the non-parametric Kruskal-Wallis test can be applied.

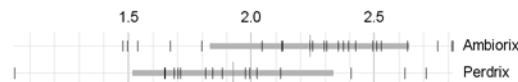


Figure 4-142. Example of an ANOVA plot with two categorical variables.

4.13.3.6.1 Parametric test for two groups: T test

The null-hypothesis is that the two groups have the same mean values. Assume the sample group observations are x_i ($i=1, \dots, n$) and y_j ($j=1, \dots, m$), with $\langle x \rangle$ and $\langle y \rangle$ the respective *mean values* for the groups. The *pooled corrected standard deviation* is defined as

$$s_d = \sqrt{\left(\sum_{i=1}^n (x_i - \langle x \rangle)^2 + \sum_{j=1}^m (y_j - \langle y \rangle)^2 \right) \left(\frac{1}{n} + \frac{1}{m} \right) / (n + m - 2)}$$

A statistic is defined as $T = (\langle x \rangle - \langle y \rangle) / s_d$.

If the null-hypothesis is true and under certain conditions (see note below) this statistic follows a t distribution with $n+m-2$ degrees of freedom. The *p-value* gives the probability that the statistic indeed has the observed value or higher. If the p-value is small, the null-hypothesis can be rejected. The *significance* of the test is calculated as the complement of the p-value, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.9.

```

Mean values:
  Ambiorix  2.552
  Perdrix   2.391
Pooled corrected standard deviation =

```

```
0.08509

T = 1.891 (37 degrees of freedom)
P value= 0.066419
Significance= 93.3581%
```

Figure 4-143. Example of a test report for a T test applied on an ANOVA plot with two categorical variables like in Figure 4-142.

NOTES:

(1) This test should not be used if the data points are not normally distributed.

(2) This test should not be used if the variances of the two samples are not the same.

4.13.3.6.2 Non-parametric test for two groups: Mann-Whitney test

The null-hypothesis is that the two groups have the same median values. Assume the observations in the sample groups are x_i ($i=1, \dots, n$) and y_j ($j=1, \dots, m$). All observations are combined into one sample and are ranked. For each group, the *sum of ranks* is determined and the smallest of those sums is taken as the U statistic. If the null-hypothesis holds and under certain conditions (see note below) this statistic approximately follows a normal distribution with mean $nm/2$ and variance $nm(m+n+1)/12$. The *p-value* gives the probability that the statistic indeed has the observed value or higher. If the p-value is small, the null-hypothesis can be rejected. The *significance* of the test is calculated as the complement of the p-value, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.9.

```
Sum of ranks:
Ambiorix 509.5
Perdrix 270.5

P value= 0.157560 (Normal approximation)
Significance= 84.2440%
```

Figure 4-144. Example of a test report for a Mann-Whitney test applied on an ANOVA plot with two categorical variables like in Figure 4-142.

NOTE: This test should not be used if one of the groups contains less than 8 members.

If the sample contains less than 30 observations, an alternative way for testing the null-hypothesis is offered by Monte-Carlo simulations. To do this, 10.000 samples with two groups of n and m randomly distributed observations are created. For each of these samples, a value for the statistic is obtained and is compared to the observed value. The *p-value* from the simulations is determined by the number of times the simulations give a larger value for the statistic than the value observed in the real sample. Also here, the *significance* is calculated as $s = 100 \times (1 - p)$. The results for the simulated p-value and significance also appear in the test report.

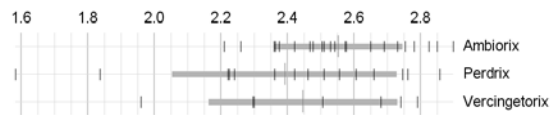


Figure 4-145. Example of an ANOVA plot with two categorical variables

4.13.3.6.3 Parametric test for more than two groups: F test

Assume that the sample contains g groups. The null-hypothesis is that all groups have the same mean. The group sizes are given by n_1, n_2, \dots, n_g , in total n observations for the complete sample. The j th observation in the i th group is denoted as x_{ij} . The sample *group means* are

$$\langle x \rangle_{groupi} = \sum_{j=1}^{n_i} x_{ij} / n_i, \text{ with } x_{ij} \text{ all observations within group } i.$$

The mean of all observations is

$$\langle x \rangle = \sum_{i=1}^g \langle x \rangle_{groupi} / g,$$

The *total sum of squares*, $SST = \sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \langle x \rangle)^2$, is

a measure for the variation in the sample around the mean of all observations. The *sum of squares among groups*

$$SSA = \sum_{i=1}^g n_i (\langle x \rangle_{groupi} - \langle x \rangle)^2$$

measures the variation among the group means. The *total within-group sum of squares*

$$SSW = \sum_{i=1}^g \sum_{j=1}^{n_i} (x_{ij} - \langle x \rangle_{groupi})^2$$

gives the variation in the sample within the groups. From the definitions it is clear that $SST=SSA+SSW$.

If the null-hypothesis holds and under certain conditions (see note below) the statistic

$F = SSA(n - g) / SSW(g - 1)$ approximately follows an F-distribution with $g-1$ and $n-g$ degrees of freedom.

The *p-value* gives the probability that the statistic obtains a value at least as high as the observed one. The *significance* of the test is calculated as the complement of the *p-value*, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.9.

```
SST= 3.347
SSA= 0.255
SSW= 3.092
```

```
F= 1.814 (2;44 degrees of freedom)
P value= 0.174938 (F approximation)
Significance= 82.5062%
P value= 0.175700 (Simulated)
Significance= 82.4300%
```

```
Group means:
Ambiorix 2.552
Perdrix 2.391
Vercingetorix 2.446
```

Figure 4-146. Example of a test report for an F test applied to an ANOVA plot with more than two categorical variables like in Figure 4-145.

In case the sample contains less than 30 observations, an alternative way for testing the null-hypothesis is offered by Monte-Carlo simulations. To do this, 10.000 samples with g groups and n_1, n_2, \dots, n_g randomly distributed observations in the groups are created. For each of these samples, a value for the F statistic is obtained and is compared to the observed value. The *p-value* from the simulations is determined by the number of times the simulations give a larger value for the F statistic than the value observed in the real sample. Also here, the *significance* is calculated as $s = 100 \times (1 - p)$. The results for the simulated *p-value* and significance also appear in the test report.

4.13.3.6.4 Non-parametric test for more than two groups: Kruskal-Wallis test

Assume the sample contains g groups. The null-hypothesis is that all groups have the same median. The number of observations in the groups are given by n_1, n_2, \dots, n_g , with n the total number of observations. All

observations are ranked, the rank for the j th observation in the i th group is denoted by R_{ij} and R_i stands for the *group rank sum* of group i .

A statistic is defined as:

$$H = \left[\frac{12}{n(n+1)} \sum_{i=1}^g \frac{R_i}{n_i} \right] - 3(n+1).$$

If the null-hypothesis holds and under certain conditions (see below) the statistic approximately follows a chi-square distribution with $g-1$ degrees of freedom.

The *p-value* gives the probability that the statistic obtains a value at least as high as the observed one. The *significance* of the test is calculated as the complement of the *p-value*, $s = 100 \times (1 - p)$.

The values for the parameters can be found in the test report. How such a chart and report can be created is explained in section 4.13.9.

```
H= 2.377 (2 degrees of freedom)
P value= 0.304666 (Chi square approximation)
Significance= 69.5334%
P value= 0.312600 (simulated)
Significance= 68.7400%
```

```
Group rank sums:
Ambiorix 623.5
Perdrix 328.5
Vercingetorix 176.0
```

Figure 4-147. Example of a test report for the Kruskal-Wallis test applied to an ANOVA plot with more than two categorical variables like in Figure 4-145.

NOTES:


- (1) In case there are only 3 groups, this test should not be used if one of the groups contains less than 6 observations.
- (2) In case there are more than 3 groups, this test should not be used if one of the groups contains less than 5 observations.

If the sample contains less than 30 observations, an alternative way for testing the null-hypothesis is offered by Monte-Carlo simulations. To do this, 10.000 samples with g groups and n_1, n_2, \dots, n_g randomly distributed observations in the groups are created. For each of these samples, a value for the H statistic is obtained and is compared to the observed value. The *p-value* from the simulations is determined by the number of times the simulations give a larger value for the H statistic than the one observed in the real sample. Also here, the *signif-*

ificance is calculated as $s = 100 \times (1 - p)$. The results for the simulated p-value and significance also appear in the test report.

4.13.4 Using the plot tool

The plot and statistics tools are available directly from the *InfoQuest FP main window* or from the *Comparison window*. In the *InfoQuest FP main window*, it can be started using *Comparison > Chart / Statistics*. When launched from the *InfoQuest FP main window*, it works on the current selection made in the database. If launched in the *Comparison window*, it works on all entries contained in the comparison.

4.13.4.1 In the *Comparison window*, click the  button or select *File > Chart / statistics*. This pops up a dialog box (see Figure 4-148) that is used to select the plot components. *All components* that can be included in a chart are listed on the left.

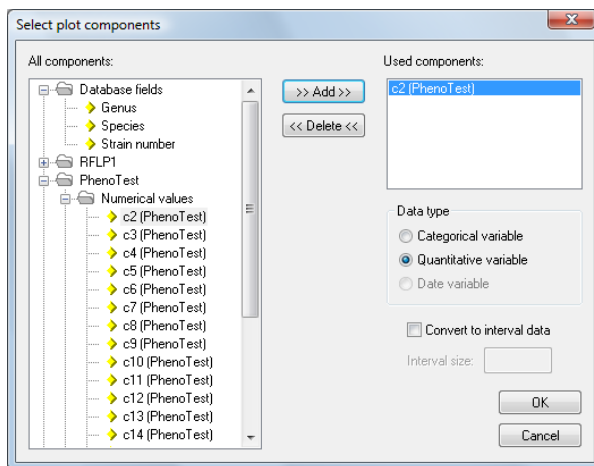


Figure 4-148. The *Select plot components* dialog box, that appears when the chart tool is started, is used to select the plot components for the chart.

4.13.4.2 To add a component to the chart, select a component from this list by clicking on it and add it to the list of *Used components* (displayed at the right) with the button *<Add>*. Also in this list, components can be clicked for selecting them. The selected component can be removed from the *Used components* list with the button *<Delete>*. For the selected component, the panel beneath the *Used components* list displays what data type it is.


4.13.4.3 Within this *Select plot components* dialog box you can convert a quantitative variable into an interval variable by checking the *Convert to interval data* checkbox. When this option is checked, the *Interval size* has to be specified. See lower right part of the panel displayed in Figure 4-148. The same procedure has to be followed if a *Date variable* has to be converted to an interval vari-


able: four choices appear in a drop-down box: *Group by day*, *Group by month*, *Group by quarter*, and *Group by year*.


4.13.4.4 For this example, select one numerical variable. After clicking the *<OK>* button, the chart appears, as in Figure 4-149. In this section, the general features and appearances of the *Chart and Statistics* window are discussed. The content of the plot will be discussed in sections 4.13.5 - 4.13.11.


4.13.4.5 To copy the plot of this window select either *File > Copy to clipboard (metafile)* or *File > Copy to clipboard (bitmap)*. A paper copy can be obtained by selecting *File > Print*.


4.13.4.6 For some type of charts, you can export the data by selecting *File > export data (formatted)* or *File > export data (TAB delimited)*. These menu items appear in grey instead of black if they cannot be applied for the current type of chart.

4.13.4.7 Selecting *Plot > Edit components* or clicking the  button pops up the *Select plot components* dialog box (see Figure 4-148). This can be used to change the *Used components*. If the list of *Used components* is modified, it is possible that the plot changes into another type of chart because the chart functionality selects the optimal representation for a given set of variables. Of course it is possible to select another type of chart (see 4.13.4.8).


4.13.4.8 From the *Plot* menu item, another type of chart can be selected. The same options are also available from the toolbar, which is displayed vertically on the left side of the window in default configuration (see Figure 4-149). Its position can be modified as described in . The toolbar has following buttons: the *Display bar graph* button ,


the *Display 2D contingency table* button ,

the *Display 2D scatterplot* button ,

the *Display 3D scatterplot* button ,



the *Display ANOVA plot* button ,

the *Display 1D distribution function* button ,

the *Display 3D bar graph* button ,

and the *Display colored bar graph* button . The button for the plot type that is presently shown is highlighted: e.g. .

If the chart type chosen is not compatible with the data type, the message "Invalid type of source data" appears.

4.13.4.9 Zooming in or zooming out on the plot can be done with *View > Zoom in* () or *View > Zoom out* (). Alternatively, the zoom slider can be used (see for a description of zoom slider functions).

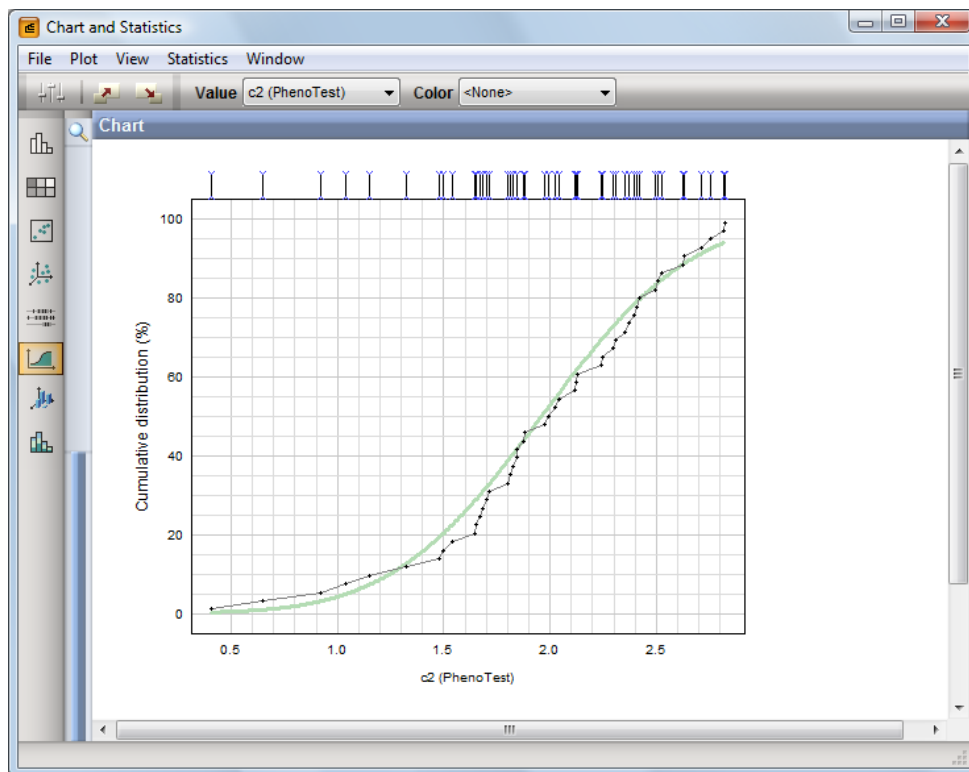


Figure 4-149. The *Chart and Statistics* window, displaying a 1-D numerical distribution function for a single quantitative variable.


The *View* menu item is divided into two parts, separated by a horizontal line. The part below the horizontal line contains menu items that change the view of the plot and that generally depend on the kind of chart that is displayed in the window. These commands will be discussed when the various charts are presented.

4.13.4.10 The last menu item is *Statistics*. Under this item, a list of statistic tests that can be applied to the selected type of plot is given. These tests will be discussed when the various charts are presented.

4.13.4.11 Selections of entries can be made within the chart, except for 3-D bar graphs. These selections are also shown e.g. in the *Comparison* window and the *InfoQuest FP main* window. If the selection is changed in the comparison, the chart is updated automatically. If another chart type is selected, the entries keep their selected/unselected state. Selections from the *Comparison* window and the *InfoQuest FP main* window are also visualized in a *Chart and Statistics* window.

In the following sections the various types of charts and their statistics are described.

4.13.5 Bar graph

4.13.5.1 Open the chart tool by clicking  in the *Comparison* window.

4.13.5.2 Select a categorical variable, e.g. an information field and add it to the list of *Used components*, then press <OK>.

4.13.5.3 This creates a *Chart and Statistics* window like shown in Figure 4-150. The component that is displayed is indicated in the toolbar. In case you selected more than one categorical variable in the *Used components* list (4.13.5.2) a drop-down list can be used to display another variable.

4.13.5.4 The entries corresponding to the bars in the chart can be selected (or unselected) by pressing the CTRL key while clicking or dragging the mouse.

4.13.5.5 Select *Statistics > Chi square test for equal category size*. This creates a *Statistics report*, as shown in Figure 4-151. A description of this test can be found in 4.13.3.1. The report can be exported by pressing <Copy to clipboard> and pasting it in another application.

4.13.5.6 With *Statistics > Index of diversity* a report window is generated which displays Simpson's index of diversity and the Shannon-Weiner index of diversity for the selected entries and categories. The report can also be copied to the clipboard.

4.13.6 Contingency table

4.13.6.1 Create a *Chart and Statistics* window with two categorical variables. This can be done from the *Compar-*

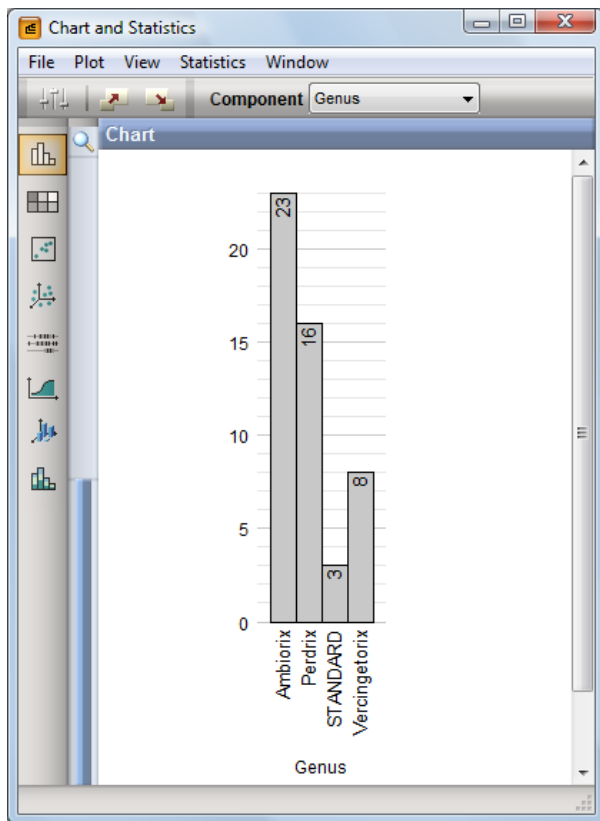




Figure 4-150. A bar graph for one categorical variable.

ison window by clicking  or within the *Chart and Statistics* window by editing the plot components after clicking the  button. Select two categorical variables into the *Used components* list. After clicking

<OK>, a contingency table like in Figure 4-152 is created.

4.13.6.2 The contents of the *X component* and *Y component* are indicated in the window. A drop-down list makes it possible to assign another categorical variable from the used components list to the *X component* and *Y component*.

4.13.6.3 Cells can be selected (or unselected) in the table by pressing CTRL while left-clicking the mouse (CTRL+Click).

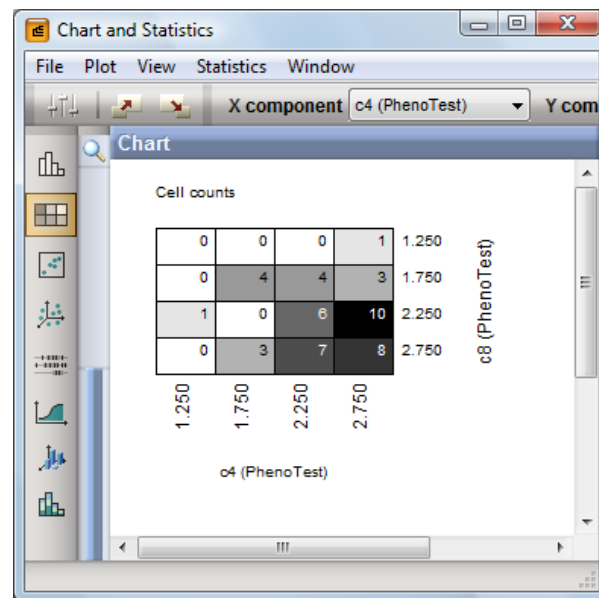


Figure 4-152. A contingency table for two categorical variables.

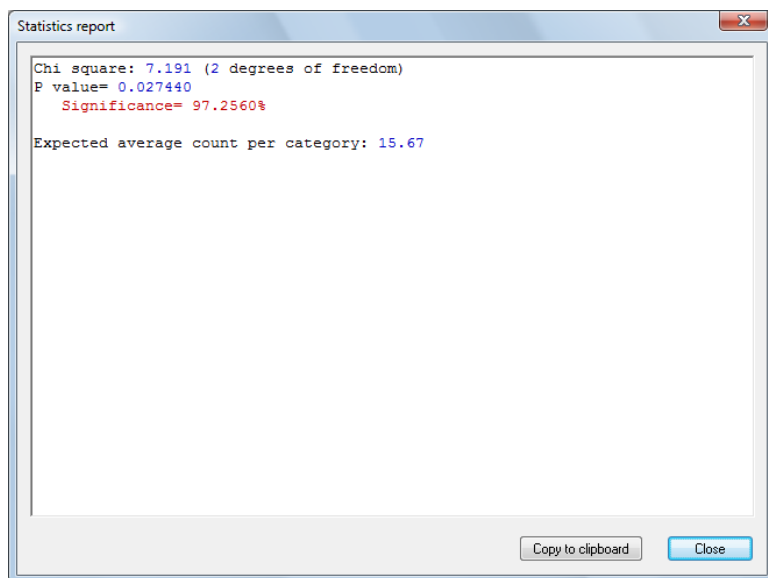



Figure 4-151. Statistics report for Chi square test for equal category sizes.



4.13.6.4 The contingency table can be displayed showing row respectively column percentages by selecting *View > Display row percentages* respectively *View > Display column percentages*.

4.13.6.5 The contingency table can be displayed in the *Chart and Statistics* window showing residuals in the cells, with *View > Display residuals*. The residual for a cell is a measure for the deviation from the expected number of counts in that cell and is calculated as $\left[N_{oij} - n_{ij} \right] / \sqrt{n_{ij}}$, with N_{oij} the observed cell count and n_{ij} the expected cell count. This view is closely related to the statistic test that can be applied to this chart (see 4.13.3.3).

4.13.6.6 Select *Statistics > Chi square test for contingency tables* to apply the statistical test that is available for this kind of plot. This creates a *Statistics report*, as shown in Figure 4-153. A description of this test can be found in 4.13.3.3.

4.13.6.7 In the *Chart and Statistics* window, you can create bar graphs for each of the two selected categorical variables by clicking the bar graph button .

4.13.7 2-D scatterplot

4.13.7.1 Create a *Chart and Statistics* window with two quantitative variables. This can be done from the *Comparison* window by clicking  or within the *Chart and Statistics* window by editing the plot components after clicking the  button. Select two quantitative variables into the *Used components* list. After

clicking **<OK>**, a 2-D scatter plot like in Figure 4-154 is created.

4.13.7.2 The contents of the *X axis* and *Y axis* are indicated in the toolbar. A drop-down list makes it possible to change the variables displayed on the axes.

4.13.7.3 Single dots can be selected in the chart by pressing CTRL + mouse click. Multiple dots are selected at once by holding the SHIFT key and drawing a rectangle around the dots with the mouse.

4.13.7.4 With the menu command *View > Regression line*, a regression line can be added to the plot. A *Regression selection* dialog box pops up (Figure 4-155), offering a choice between several types of regression lines. After selecting a regression type and clicking **<OK>** in the dialog box, a small statistics report is generated. If a regression line is fitted, it is shown as a thick green line. The 1-sigma uncertainty levels are plotted as a thin green line.

4.13.7.5 Under the menu item *Statistics*, a number of statistic test can be found: *T test for mean value (paired*

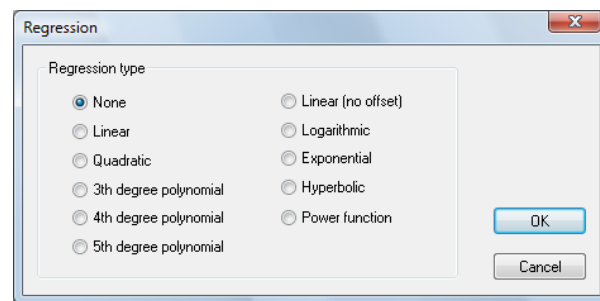


Figure 4-155. The *Regression selection* dialog box, where the type of regression line for the scatter plot can be selected.

Marginal column counts:		
1.250	1	2.13%
1.750	7	14.89%
2.250	17	36.17%
2.750	22	46.81%

Marginal row counts:		
1.250	1	2.13%
1.750	11	23.40%
2.250	17	36.17%
2.750	18	38.30%

Figure 4-153. *Statistics report* for the Chi square test for contingency tables.

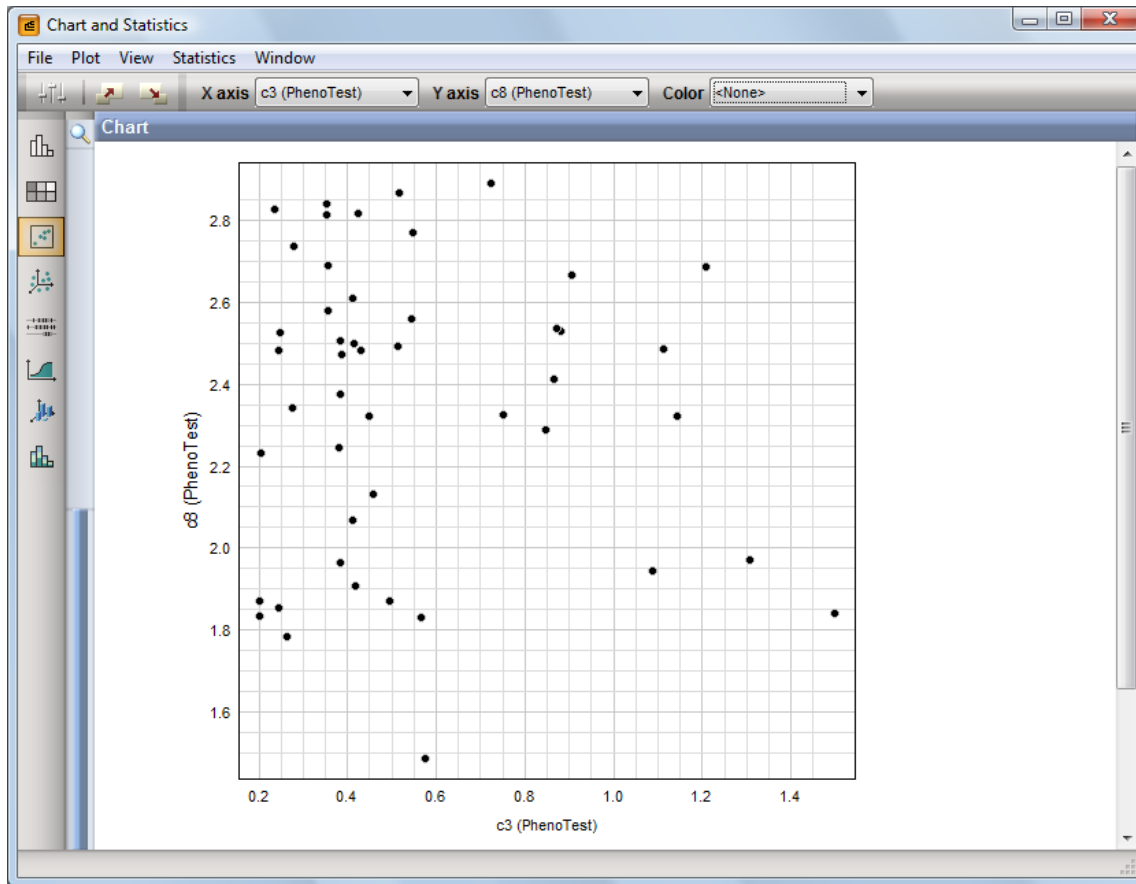




Figure 4-154. 2-D scatterplot for two quantitative variables.


samples), *Wilcoxon signed ranks test (paired samples)*, *Pearson correlation test* and *Spearman rank-order correlation test*. Each of these tests generates a statistics test report. A description of these tests can be found in 4.13.3.5.

4.13.7.6 If one or more categorical variables are present in the *Used components* list, additional information from one of these variables can be displayed in color code by selecting the variable from the color drop-down list. If this is the case, you can change the color labels with the command *View > Label with continuous colors*.

4.13.7.7 For each of the quantitative variables used in this plot, a 1-D distribution function plot can be generated. This can be done by selecting *Plot > 1D distribution function*, or by clicking the  button. For more details on this kind of chart, see 4.13.10.

4.13.8 3-D scatterplot

4.13.8.1 Create a *Chart and Statistics* window with three quantitative variables. This can be done from the *Comparison* window by clicking  or within the *Chart and Statistics* window by editing the plot compo-



nents after clicking the  button. Select three quantitative variables into the *Used components* list. This will create a 3-D scatterplot.

4.13.8.2 The variables that are displayed on the respective axes are indicated beneath the toolbar. The variables can be switched between the *X axis*, *Y axis* and *Z axis*. A drop-down list makes it possible to assign another quantitative variable from the *Used components* list to the respective axes.



4.13.8.3 Dots can be selected in the chart by CTRL + mouse click or by holding the SHIFT key and drawing a rectangle around the dots with the mouse. The corresponding entries are also selected in the *Comparison* window and the *InfoQuest FP main* window. If they are removed from the comparison, the chart is updated automatically. Selections made in the *Chart and Statistics* window are automatically updated in the *Comparison* window and vice versa.

4.13.8.4 By clicking on the plot and holding the left mouse button, the plot can be rotated in different directions. The data points in the plot can be displayed as small dots or as larger spheres, which can be achieved by checking or unchecking the command *View > Show rendered spheres*.

4.13.8.5 If one or more categorical variables are present in the *Used components* list, additional information from one of these variables can be displayed in color code by selecting the variable from the color drop-down list. If this is the case, you can change the color labels with the command *View > Label with continuous colors*.

4.13.8.6 For each of the quantitative variables used in this plot, a *1-D distribution function* plot can be generated. This can be done by selecting *Plot > 1D distribution function*, or by clicking the  button. For more details on this kind of chart, see 4.13.10. For each couple of categorical variables used in this plot, a 2-D scatterplot can be generated. This can be done by selecting *Plot > 2D scatterplot*, or by clicking the  button. For more details on this kind of chart, see 4.13.7.

4.13.9 ANOVA plot



4.13.9.1 Create a *Chart and Statistics* window with one categorical and one quantitative variable. This can be done from the *Comparison* window by clicking  or within the *Chart and Statistics* window by editing the plot components after clicking the  button. Select one categorical and one quantitative variable for the plot. This creates an ANOVA plot like in Figure 4-156. The data for each category is presented on a horizontal line. The scale for the line is indicated at the top. Each data point is indicated with a small vertical mark at the position according to its numerical value and the category it belongs to.

4.13.9.2 The categorical and quantitative variables that are displayed are indicated in the toolbar. A drop-down list makes it possible to assign other variables from the *Used components* list to the respective axes.

4.13.9.3 Vertical marks, indicating the database entries, can be selected in the chart by CTRL + mouse click or by holding the SHIFT key and drawing a rectangle around the marks with the mouse.

4.13.9.4 From the menu item *Statistics*, the ANOVA test (*F test*), the *Kruskal-Wallis test* (in case more than two categorical variables are used), the *T test* or the *Mann-Whitney test* (in case only two categorical variables are used) can be launched. For these tests a statistics report is generated. A description of these tests can be found in 4.13.3.6.

4.13.9.5 If one or more categorical variables are present in the *Used components* list, additional information from one of these variables can be displayed in color code by selecting the variable from the color drop-down list. If this is the case, you can change the color labels with the command *View > Label with continuous colors*.

4.13.9.6 For the quantitative variable used in this plot, a 1-D distribution function plot can be generated. This can be done by selecting *Plot > 1D distribution function*, or by clicking the  button. For more details on this kind of chart, see 4.13.10. For the categorical variables used in this plot, a bar graph can be generated. This can be done by selecting *Plot > Bar graph*, or by clicking the  button. For more details on this kind of chart, see 4.13.5.

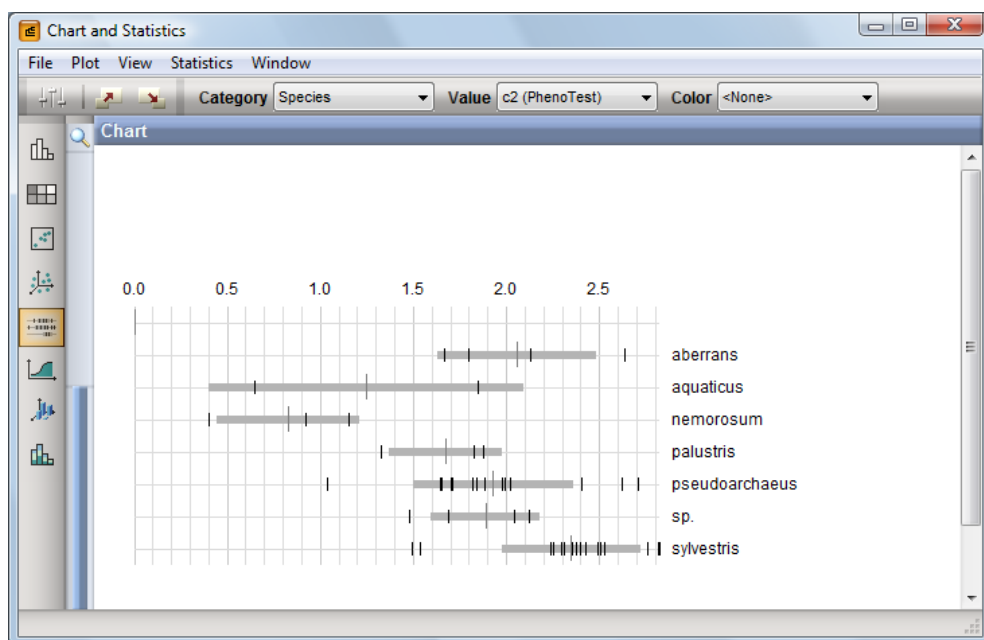


Figure 4-156. ANOVA plot for a categorical and a quantitative variable.

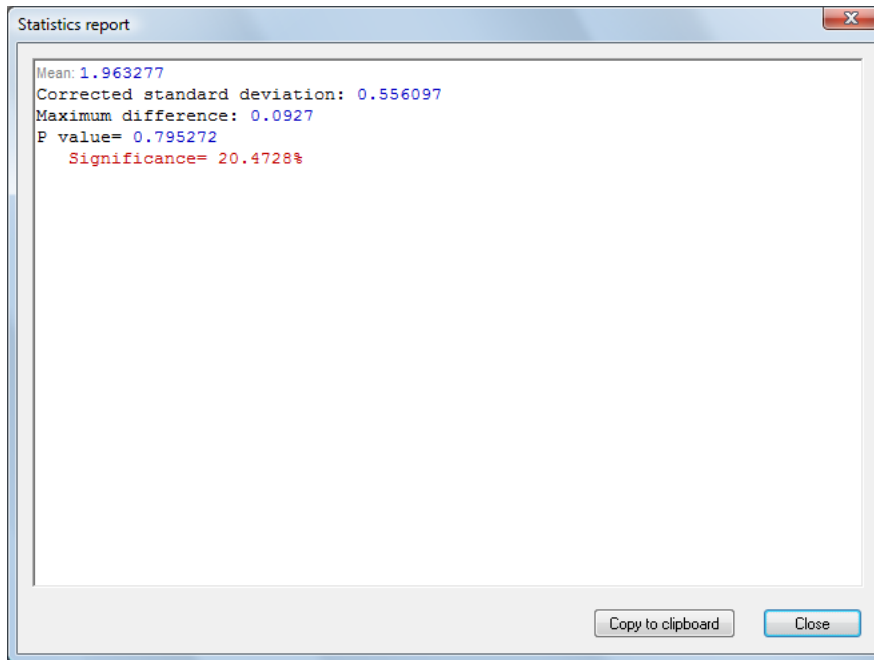




Figure 4-157. Statistics report for a Kolmogorov-Smirnov test.

4.13.10 1-D numerical distribution

4.13.10.1 Create a *Chart and Statistics* window with one quantitative variables. This can be done from the

Comparison window by clicking  or within the *Chart and Statistics* window by editing the plot components after clicking the  button. Select only one quantitative variable for the plot. This will create a 1-D cumulative distribution function plot as previously shown in Figure 4-149. The dots present the data points, each dot has a corresponding vertical mark just above the chart. The smooth green line is the normal distribution that serves as a model for the data.

4.13.10.2 The variable that is displayed is indicated in the toolbar. A drop-down list is available to select another variable in case there is more than one numerical variable in the *Used components* list.


4.13.10.3 Data points can be selected in the chart by CTRL + mouse click or by holding the SHIFT key and drawing a rectangle around the vertical marks with the mouse.

4.13.10.4 Select *Statistics > Kolmogorov-Smirnov test for normality* for applying the statistical test that is available for this kind of plot. This will create a *Statistics report*, as shown in Figure 4-157. A description of this test can be found in 4.13.3.4..

4.13.10.5 Instead of a cumulative distribution, the data can be presented as bar graph by unchecking the command *View > Display cumulative distribution*.

4.13.10.6 Additional information from a categorical variable can be displayed in color code. In this case, with the menu item *View*, you can change the color code into a continuous color code and back


4.13.11 3-D Bar graph

4.13.11.1 For categorical variables, a 3-D bar graph can be plotted, see Figure 4-158 for an example. This can be done by selecting two categorical variables for the plot and by clicking the  button or by selecting *Plot > 3D bar graph* from the menu.

4.13.11.2 Under the menu item *View*, there is the option to *Label the X axis in color*, to *Label the Y axis in color* or to *Label with continuous colors*.

4.13.11.3 By clicking on the plot and holding the left mouse button, the plot can be rotated in different directions.

4.13.12 Colored bar graph

4.13.12.1 For categorical variables, a colored bar graph can be generated, see Figure 4-159 for an example. This can be done by selecting two categorical variables for the plot and by clicking the  button or by selecting *Plot > colored bar graph* from the menu.

The components used as *X component* and *Color* are indicated in the toolbar. The drop-down list can be used to display another variable.

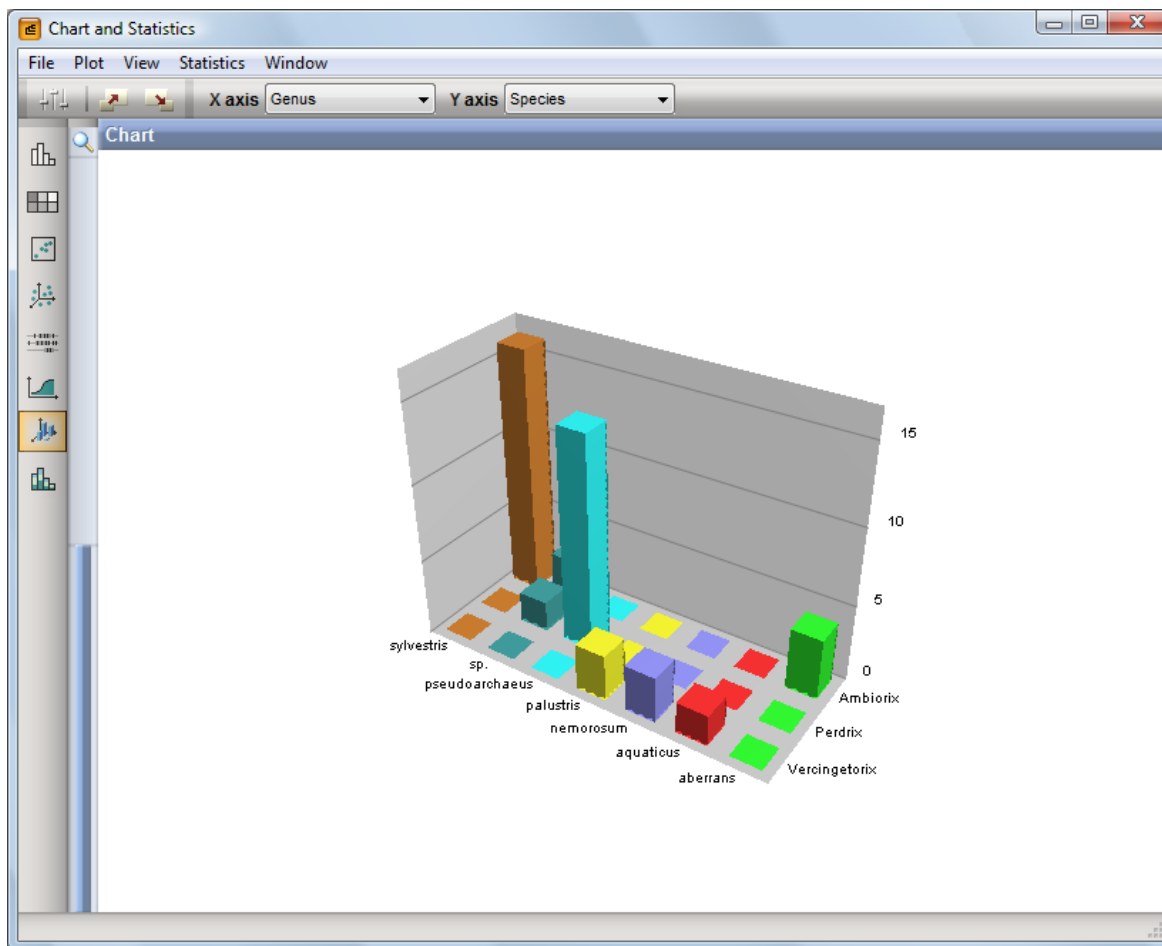


Figure 4-158. 3-D bar graph for two categorical variables.

4.13.12.2 The entries corresponding to the colored bars in the chart can be selected (or unselected) by pressing the CTRL key while clicking or dragging the mouse.

4.13.12.3 Select *View > Show percentages* to scale the colored blocks in a relative fashion.

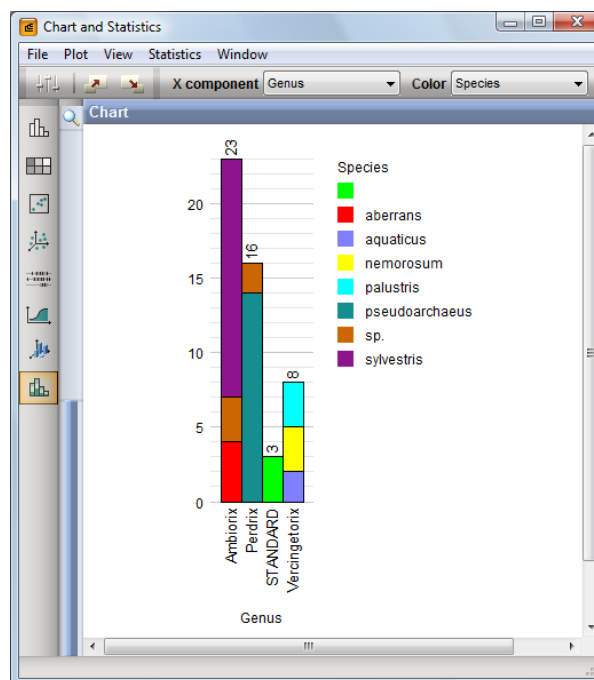


Figure 4-159. Colored bar graph for two categorical variables.

5. IDENTIFICATION


5.1 Identification with database entries


There are three methods for identification available in InfoQuest FP. The most straightforward way (described in this section), is to compare and identify unknown patterns against a selection of database patterns stored on disk. The two more sophisticated methods are to identify unknown patterns against an identification library (see Section 5.2) or identification using a decision network (see Section 5.3).

5.1.1 Creating lists for identification

In InfoQuest FP, a comparison can be used as a “container” for a number of well-characterized database entries that are used to identify against.

5.1.1.1 In **DemoBase**, select all *Ambiorix* entries except the *Ambiorix* sp. entries: First perform a search with *Ambiorix* as genus name, and then perform a second search with *Search in list* and *Negative search* enabled, and sp. as species string. For more information about the automatic search and select functions, see 2.2.8 and 2.2.9.

5.1.1.2 Select **Comparison > Create new comparison** (ALT+C) or press the  button in the *Comparisons* panel toolbar to create a comparison with the selected entries.

5.1.1.3 Select **File > Save as** or press  to save the comparison (shortcut CTRL+S on the keyboard). Enter **Ambiorix** as name for the comparison.

5.1.1.4 Exit the *Comparison* window.


5.1.2 Identifying unknown entries

First we select the entries which we want to identify. We consider the *Ambiorix* sp. entries (those without species name) as unknown, and we will identify them against the known *Ambiorix* entries (the list **Ambiorix**).


5.1.2.1 In the *InfoQuest FP main* window, press F4 to clear the selection.

5.1.2.2 Select all *Ambiorix* sp. entries (in the *Entry search dialog box*, disable *Search in list* and *Negative search* and enter *Ambiorix* in the ‘Genus’ field and sp. in the ‘Species’ field). For more information about the automatic search and select functions, see 2.2.8 and 2.2.9.

5.1.2.3 Copy the selected entries to the clipboard using

Edit > Copy selection or .

5.1.2.4 Open the saved comparison **Ambiorix** by double-clicking on **Ambiorix** in the *Comparison* panel.

5.1.2.5 Paste the selected *Ambiorix* sp. entries into the comparison with **Edit > Paste selection** or .

5.1.2.6 For identification purposes, we do not need the *Dendrogram* panel (left, see 4.1.3 and Figure 4-1), which you can minimize.


5.1.2.7 Create sufficient space for the *Similarities* panel (right, see Figure 4.1.3 and Figure 4-2), where the similarity values will appear.

5.1.2.8 In the *Experiments* panel, select an experiment by means of which you want to identify the unknown entries. Select for example **FAME** (fatty acid methyl esters).

5.1.2.9 Click on the first unknown *Ambiorix* sp. entry in the *Information fields* panel. This entry now becomes highlighted.

5.1.2.10 In the menu of the *Comparison* window, choose **Edit > Arrange entries by similarity**.

The highlighted entry stands on top and all the other entries in the comparison are arranged by decreasing similarity with that entry. The similarity values are shown in the *Similarities* panel.

5.1.2.11 You can click on the  button of **FAME** to display the images and drag the horizontal separator line down to show the complete names of the fatty acids.

The **Arrange entries by similarity** function can be repeated for each experiment type and for composite data sets, in order to compare the different results. The program uses the similarity coefficient which is specified in the *Experiment type* window (see Chapter 3).

5.1.2.12 A printout of the list of similarity values can be obtained with **File > Print database fields**.

5.1.2.13 An export file of the similarity values is created with **File > Export database fields**.

*NOTE: In case of a fingerprint type, you can also show the number of different bands between a highlighted entry and the other entries, by selecting **Different bands** as the default similarity coefficient (see Figure 4-*

29). Before selecting **Edit > Arrange entries by similarity**, you should enable **Layout > Show distances**.

5.1.3 Fast band-based database screening of fingerprints

In case of large databases of fingerprint patterns, the most time-consuming part of a quick database screening of new or unknown patterns is reading or downloading all the fingerprint information. InfoQuest FP offers a tool that overcomes this bottleneck by generating a cache containing band information of all available fingerprints belonging to a fingerprint type. When a database screening is performed, this cache is loaded rather than the full gel information. This cache-based fingerprint screening is extremely fast, even for the largest databases, but is limited to band-based comparisons of fingerprint patterns. In addition, the feature is only available in a connected database environment (see 2.3), where a special column holding the quick-access band information is generated (7.1.17). To try out this feature, you can e.g. install the **DemoBase_SQL** database, as described in 1.3.2.

5.1.3.1 The fast band-based identification can be enabled in the *Fingerprint type* window (*Experiments* panel), by selecting **Settings > Enable fast band matching** (this menu command appears only in a connected database). A question pops up **"Do you want to generate cached patterns for all current fingerprints?"**. By answering **<Yes>**, a cached pattern will be generated for all patterns present in the database that belong to the selected fingerprint type. If you answer **<No>**, a cached pattern will be created only for new patterns that are added to the database.

5.1.3.2 The fast band matching identification tool is launched from the *InfoQuest FP main* window, where a set of selected entries will be identified against all other database entries.

NOTE: For the fast band matching identification tool to work, metrics information (molecular weight regression) needs to be available for the active reference system of the selected fingerprint type (see 3.2.9).

5.1.3.3 A menu command **Identification > Fast band matching** (only in a connected database) pops up the *Fast band matching* dialog box (Figure 5-1). Under **Experiment type**, select the fingerprint type you want to use for the band matching. With **Used range**, you can specify a range of the pattern (in percentage distance from top) within which bands will be compared. The **Tolerance** is the same as the Position tolerance explained in . With **Maximum difference**, you can specify the maximum number of different bands between the unknown pattern and a database pattern to be included in the result set. Furthermore, the **Result set** can be limited to a certain number (default 20). In the input box **SQL query**, it is possible to enter an SQL query, to limit the search to

a subset of entries that match a specific string entered for an information field.

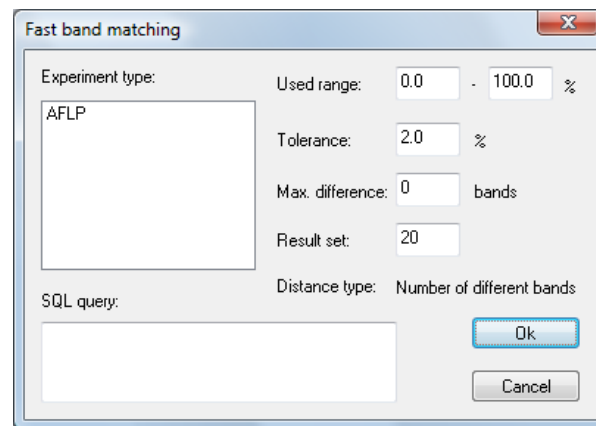


Figure 5-1. The *Fast band matching* dialog box.

The typical syntax of a restricting SQL query is:

```
"GENUS"='Ambiorix'
```


One can also combine statements, for example:

```
"GENUS"='Ambiorix' AND "SPECIES"='sylvestris'  
"GENUS"='Ambiorix' OR 'Perdrix'
```


5.1.3.4 By pressing **<OK>** the fast band matching is executed, and the identification result pops up in the *Fast matching* window (Figure 5-2). This window is subdivided in two dockable panels, of which the *Entries* panel lists the entries to be identified, and the *Matches* panel lists the result set for the selected entry in the *Entries* panel (for display options of dockable panels, see 1.6.4). The only matching criterion used is the number of different bands, which is listed in the 'Distance' column of the *Matches* panel.

NOTES:

(1) In cases where matching patterns are identical, there may be a small decimal distance. For each identical match, the software uses the band pair with the highest shift and adds this shift value to the match (i.e. to zero). This is an additional feature to sort identical patterns according to distance based upon shifts within the defined position tolerance.

(2) In the *Entries* panel and *Matches* panel of the *Fast matching* window, the same information fields are displayed as in the *Database entries* panel of the *InfoQuest FP main* window. To display or hide other information fields in a panel, click on the column properties button  in the information fields header.

5.1.3.5 In both panels of the *Fast band matching* window, you can select or unselect entries using the mouse in combination with the SHIFT or CTRL keys. You can also pop up the *Entry edit* window by double-clicking on an entry or pressing ENTER.

5.1.3.6 A text report can be exported with *File > Export* or by pressing the  button. A tab-delimited text file is opened in Notepad, where the matched entries are listed together with the best matching database entries, sorted according to number of different bands.

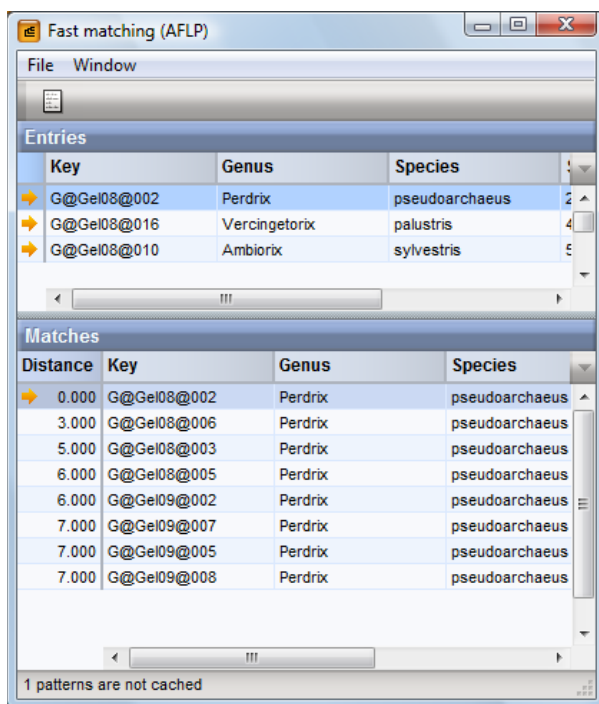


Figure 5-2. The *Fast matching report* window.

5.1.4 Fast character-based identification



Similar as for fingerprints (5.1.3), the software offers a tool for screening a database entry against a connected database, based upon a character type experiment. This identification tool benefits from a bulk-fetching mechanism, which makes it many times faster for identification against large databases. Unlike for a fingerprint type, there is no indexing of the experiment information needed to optimize the speed.

5.1.4.1 With a selection of entries to identify made in the database, select *Identification > Fast character set matching*.

The *Fast character set matching* window appears (Figure 5-3), displaying the available character type experiments under *Experiment type*. With Distance type, the coefficient on which the distance is based can be chosen. Available coefficients include *Pearson correlation*, *Cosine correlation*, *Canberra metric*, *Euclidean distance*, *Manhattan distance* and the *Categorical* coefficient. The distances are calculated as $100 - [\% \text{ similarity or correla-}$

tion]. Under *Max. difference*, a maximum percentage distance can be entered for database entries to be listed as matching. The maximum number of matching database entries to be displayed can be specified under *Result set*.

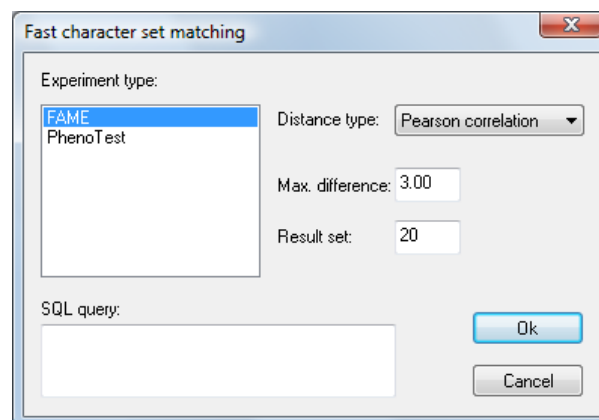




Figure 5-3. *Fast character set matching* dialog box.

Similar as for fingerprints, an SQL query can be entered, to limit the search to a subset of entries that match a specific string entered for an information field. See 5.1.3 for examples of such SQL queries.

5.1.4.2 By pressing <OK> the fast character matching is executed, and the identification result pops up in the *Fast matching report* window (see Figure 5-2 under 5.1.3). This window is subdivided in two dockable panels, of which the *Entries* panel lists the entries to be identified, and the *Matches* panel lists the result set for the selected entry in the upper panel (for display options of dockable panels, see 1.6.4). The matching criterion used is the distance based upon the coefficient used, calculated as $100 - [\% \text{ similarity or correlation}]$. These values are listed in the *Distance* column of the *Matches* panel.

NOTE: In the Entries panel and Matches panel of the Fast matching window, the same information fields are displayed as in the Database entries panel of the InfoQuest FP main window. To display or hide other information fields in a panel, click on the column properties button  in the information fields header.

5.1.4.3 In both panels of the *Fast character matching* window, you can select or unselect entries using the mouse in combination with the SHIFT or CTRL keys. You can also pop up the *Entry edit* window by double-clicking on an entry or pressing ENTER.

5.1.4.4 A text report can be exported with *File > Export* or by pressing the  button. A tab-delimited text file is opened in Notepad, where the matched entries are listed together with the best matching database entries, sorted according to their distance to the matched entries.

5.1.5 Fast sequence-based identification



Similar as for fingerprints (5.1.3) and characters (5.1.4), the software offers a tool for screening a database entry against a connected database, based upon a sequence type experiment. Unlike for a fingerprint type, there is no indexing of the experiment information needed to optimize the speed.

5.1.5.1 With a selection of entries to identify made in the database, select **Identification > Fast sequence matching**.

The *Fast sequence matching* window appears (Figure 5-4), displaying the available sequence type experiments under **Experiment type**. The similarity type is the default setting specified for the experiment type. The distances are calculated as $100 - [\% \text{ similarity or correlation}]$. Under **Max. difference**, a maximum percentage distance can be entered for database entries to be listed as matching. The maximum number of matching database entries to be displayed can be specified under **Result set**.

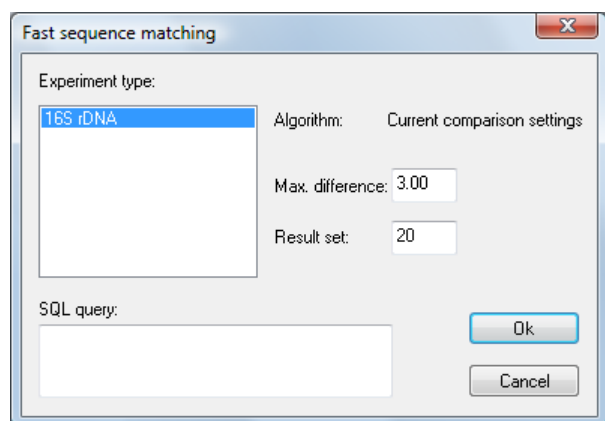



Figure 5-4. *Fast sequence matching* dialog box.

Similar as for fingerprints and character sets, an SQL query can be entered, to limit the search to a subset of entries that match a specific string entered for an information field. See 5.1.3 for examples of such SQL queries.


5.1.5.2 By pressing **<OK>**, the fast sequence matching is executed, and the identification result pops up in the *Fast matching report* window (see Figure 5-2 under 5.1.3). This window is subdivided in two dockable panels, of which the *Entries* panel lists the entries to be identified, and the *Matches* panel lists the result set for the selected entry in the *Entries* panel (for display options of dockable panels, see 1.6.4). The matching criterion used is the distance based upon the coefficient used, calculated as $100 - [\% \text{ similarity or correlation}]$. These values are listed in the *Distance* column of the *Matches* panel.

NOTE: In the Entries panel and Matches panel of the Fast matching window, the same information fields are

displayed as in the Database entries panel of the InfoQuest FP main window. To display or hide other information fields in a panel, click on the column

properties button  *in the information fields header.*


5.1.5.3 In both panels of the *Fast matching report* window, you can select or unselect entries using the mouse in combination with the SHIFT or CTRL keys. You can also pop up the *Entry edit* window by double-clicking on an entry or pressing ENTER.

5.1.5.4 A text report can be exported with **File > Export** or by pressing the  button. A tab-delimited text file is opened in Notepad, where the matched entries are listed together with the best matching database entries, sorted according to number of different bands.

5.1.6 Probabilistic identification

• Concepts

Probabilistic identification is usually applied to the identification of bacteria based upon sets of phenotypic tests. Consequently, in InfoQuest FP, this tool only works in

combination with the *Character types* module . A probabilistic identification matrix is generated from a table of test results of known reference organisms (see Rypka et al., 1967¹). The table has a number of taxa as rows and a number of tests as columns (see Table 1). The numbers in the matrix represent the chance that an organism belonging to the taxon of the respective row has a positive score for the test of the respective column. This chance, p , is given as a percentage between 1 and 99 (for algorithmic-technical reasons, 0 and 100 are not used). By convention, missing results are represented as $p = 50$. A test is considered positive when $p \geq 85$ or negative when $p \leq 15$.

• Performing probabilistic identifications

InfoQuest FP can read probabilistic identification matrices in three different formats:

- *Excel* (.xls format):

The identification matrix is looked for in the following order:

- The *Named range* with the name '**matrix**'
- The first *Named range* with a name
- The sheet with the name '**matrix**'
- The first sheet

- *CSV* (Comma Separated Values): a comma, a dot and a tab-character are recognized as column separator signs.

1. Rypka, E.W., W.E. Clapper, I.G. Bowen, and R. Babb. 1967. A model for the identification of bacteria. *J. Gen. Microbiology* 46: 407-424.

- *Fixed Format*: Text format matrices (.mat) as described and used in the PIB software (Bryant, 1995¹).

A probabilistic identification matrix in Excel format, based on the fictitious organisms used in database **DemoBase**, is provided on the installation CD-ROM and can be found as **Sample and Tutorial data\Probabilistic identification data\Prob_Id.xls**. Alternatively, the same Excel file is also available from the download page of the website (www.bio-rad.com/softwaredownloads).

Before a probabilistic identification can be done, some phenotypic tests needs to be added to entries of the **DemoBase**. Depending on whether your InfoQuest FP software has the *Database sharing tools* module or not, you will need to follow procedure [A] or [B] to add the new experiment type and the tests. You can find out whether you have the Database Sharing Tools by clicking *File > About* in the *InfoQuest FP main* window. If a hyphen appears left from **Database Sharing Tools**, the module is present. For more information about the InfoQuest FP modules, see 1.1.5.

[A] Using the *Database sharing tools* module (DS), you can import the phenotypic tests directly from the CD-ROM (or from the downloaded and unzipped folder from the website) using the XML Tools plugin.

5.1.6.1 To activate the XML Tools plugin, select *File > Install / remove plugins* in the *InfoQuest FP main* window, select the XML Tools and click **<Install>** (see also 1.5.3 on how to install plugins).

5.1.6.2 Select *File > Import selection as XML* in the *InfoQuest FP main* window. A file open dialog box appears that allows you to select the XML files.

5.1.6.3 Two files, **DatabaseEntries_1.xml** and **Database-Layout.xml**, should be selected from the **Sample and**

1. Bryant T. 1995. Software and identification matrices for probabilistic identification of bacteria (PIB). Southampton University, UK. Available from <http://www.som.soton.ac.uk/staff/tnb/pib.htm>.

Tutorial data\Probabilistic identification data directory on the CD-ROM or from the downloaded and unzipped folder from the website.

5.1.6.4 In the *XML import* dialog box, click **<OK>**. A new experiment type, **ID Tests** is generated and three entries have data for it.

[B] Without the *Database sharing tools* module, you will have to create and enter the experiment data manually, as follows.

5.1.6.5 Create a new *binary* and *closed* character type, **ID Tests**, with 5 columns and maximum value 1 (see 3.3.1 on how to define a new character type).

5.1.6.6 Enter the following character tests to the new character type (see 3.3.2; make sure the names are entered exactly):

- Glycerol PWS
- Inositol PWS
- Mannitol PWS
- Sorbitol PWS
- Oxydase

5.1.6.7 Enter the data for character type **ID Tests** as given in Table 2 via the experiment card of the corresponding database entries (see 3.8.3).

Once the data is entered, a probabilistic identification can be obtained as follows.

5.1.6.8 Select the entries in the database for which character type **ID Tests** are defined.

5.1.6.9 In the *InfoQuest FP main* window, select **Identification > Probabilistic identification**.

5.1.6.10 A dialog box lists the available character types; select **ID Tests** and press **<OK>**.

5.1.6.11 Next, you are prompted to select the probabilistic identification matrix, which can be of file type .xls, .csv, or .mat. Select **Prob_Id.xls** in the **Sample and Tutorial data\Probabilistic identification data** directory on

Vercingetorix species	Motility at RT (1)	Glycerol PWS (2)	Inositol PWS (3)	Oxidase (4)
<i>Vercingetorix aquaticus</i>	1	95	1	1
<i>Vercingetorix nemorosum</i>	1	99	1	99
<i>Vercingetorix palustris</i>	1	99	85	20
<i>Vercingetorix maritimus</i>	50	95	1	15
<i>Vercingetorix viridis</i>	94	97	1	15

Table 1: Example of a probabilistic identification matrix.

the CD-ROM or from the downloaded and unzipped folder from the website, and press *<Open>*.

The resulting *Probabilistic identification* window looks as in Figure 5-5. In the left panel, the selected database entries are listed. For the currently selected entry, the identification report is displayed in the right panel. The two columns, 'Taxon' and 'Identification score', list the different taxa in the identification matrix, and the probability score with the selected entry, respectively.

If a probability is higher than 0.95, the score is indicated in green. This is the case for the first entry in the example data set.

5.1.6.12 The list can be sorted according to the taxon name by clicking on the 'Taxon' column header. It can also be sorted according to the score by clicking on the 'Identification score' column header.

5.1.6.13 The identification result for the selected entry can be saved as a Rich Text Format (RTF) file with *<Save>*.

The *Test results* tab displays the test results for the selected entry (Figure 5-6). If the entry has a positive score for a test, a green + is shown under 'Result'.

Strain	Glycerol PWS	Inositol PWS	Mannitol PWS	Sorbitol PWS	Oxydase
42815	+	+	-	-	+
42816	+	-	-	+	+
42853	+	-	+	+	-

Table 2: Example input data for probabilistic identification matrix Prob_Id.xls (see text).

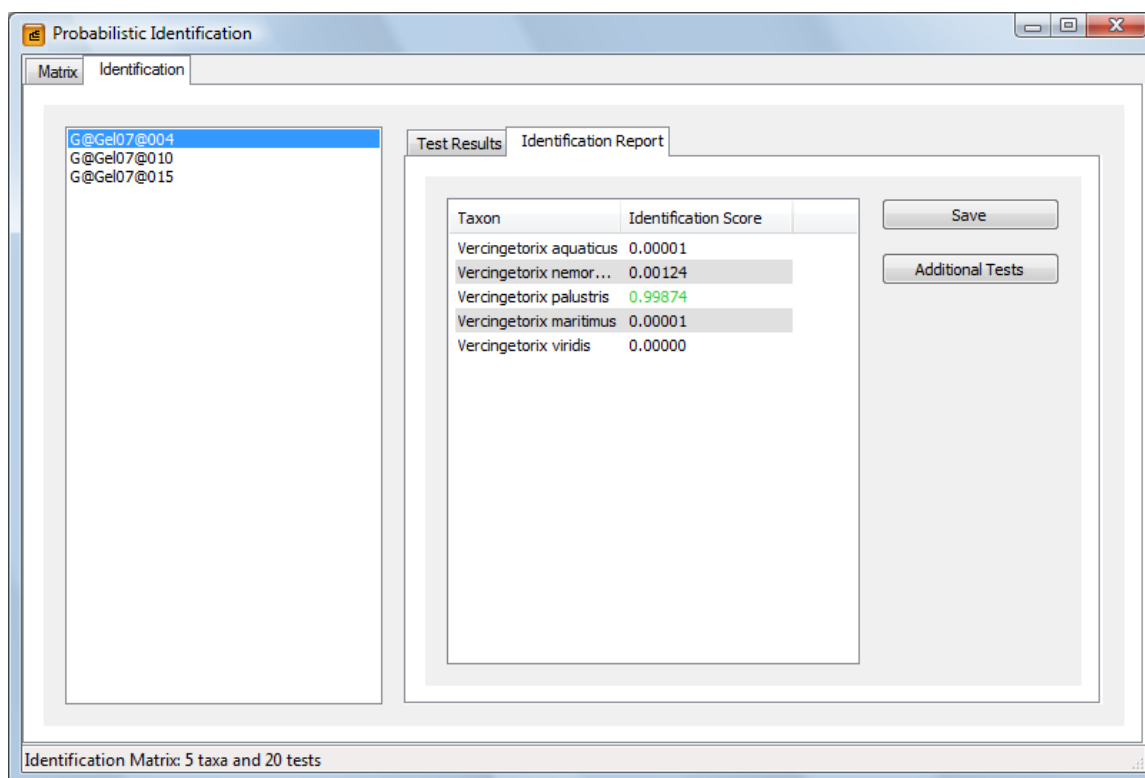


Figure 5-5. Probabilistic identification window, Identification panel.

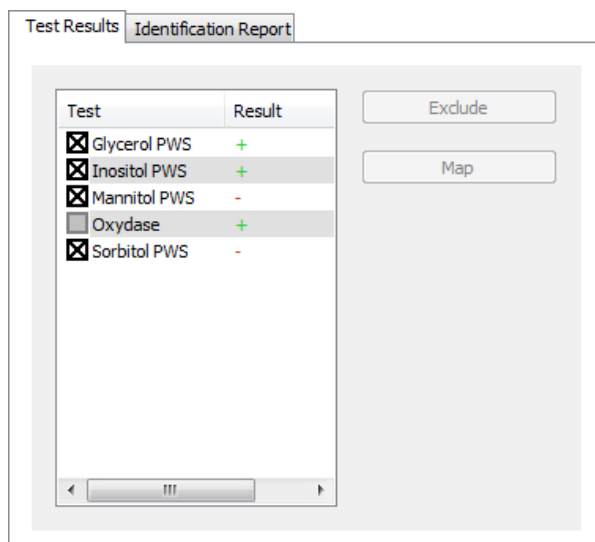


Figure 5-6. Probabilistic identification window, Test results tab in the Identification panel.

Each character has a checkbox (left), which is checked by default, and which means that the character is included in the identification.

5.1.6.14 To exclude a character, select it and press the <Exclude> button, or simply double-click on the character.

5.1.6.15 To see the updated result, click the *Identification report* tab.

In the example identification report, the character Oxidase is grayed in the *Test results* tab. This is because the character was incorrectly entered as Oxydase. The matrix, however, contains a character name Oxidase. To avoid such problems, characters from the database can be mapped on character from the matrix with different spelling or different names:

5.1.6.16 Select the unmapped character and press the <Map> button.

5.1.6.17 A box appears that lists all the characters present in the identification matrix. Select Oxidase and press <OK>.

The *Test result* tab now shows **Oxydase [Oxidase]**, which means that the character is mapped. The updated identification result can be viewed if the *Identification report* tab is clicked.

In case an identification does not provide a clear answer for a given entry, i.e. if the probability score is less than 0.95, the program can calculate the minimum number of additional tests to perform to provide a clear identification for the entry.

5.1.6.18 In the example data set, the third entry (key G@Gel07@015) remains unidentified based upon the 5

characters used. Select this entry and press <Additional tests>.

In the dialog box that pops up (Figure 5-7), you can check or uncheck individual characters. The algorithm will only use the checked characters to search for a minimal set of extra characters. In larger data sets, finding a minimal subset of characters can be a quite tough mathematical problem to solve. Therefore, heuristic methods have been devised:

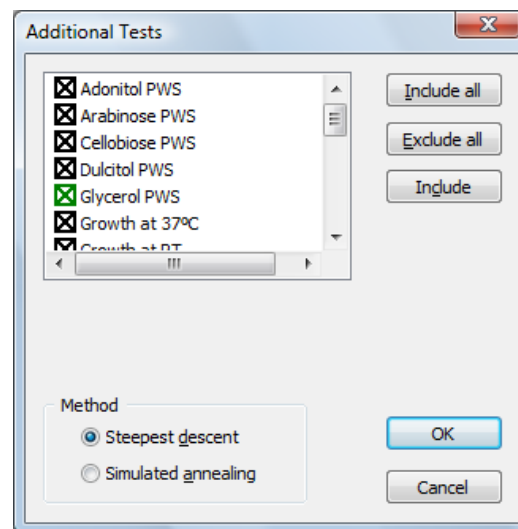


Figure 5-7. Dialog box to search for a minimal set of additional characters needed to identify an entry.

- **Steepest descent:** This method first uses the character that separates the largest number of taxa from the unknown; thereafter combinations of tests are added until the best separation is achieved.

- **Simulated annealing:** This method starts from all characters included and will randomly remove characters, and then progressively add characters, until the separation reaches a maximum. Using this method, different answers can be obtained after repeated calculations.

5.1.6.19 Press <OK> to calculate the additional tests to perform. The result is shown in Figure 5-8.

The characters that were already included in the identification are shown in green, the additional tests to be performed are shown in black.

Smax (maximal *Separation value*) is the total number of pairs of taxa to separate from each other in the identification matrix. S is the actual *Separation value* obtained after adding the proposed additional tests.

5.1.6.20 This report can be saved as a Rich Text Format (RTF) file.

NOTE: If additional reports are saved to the same file, the reports will be appended; the file will not be overwritten.

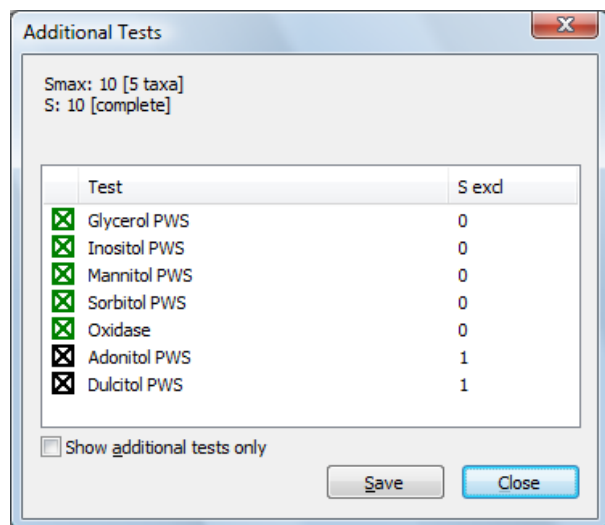


Figure 5-8. Additional tests to perform as calculated using the *Steepest descent* method.

• The identification matrix

5.1.6.21 In the top left corner of the *Probabilistic identification* window, the *Matrix* tab can be selected instead of the *Identification* tab (see Figure 5-9).

As mentioned before, missing results are represented as $p = 50$. A test is considered positive when $p \geq 85$ or negative when $p \leq 15$.

5.1.6.22 The matrix can be replaced by '+' and '-' signs by checking *Symbolic representation*. Undetermined results or missing values are represented by a 'v'.

5.1.6.23 Using the checkbox *Matrix ID scores*, you can allow the program to perform a quality test on the identification matrix.

The test checks whether each taxon can be separated from all other taxa in the matrix, as follows:

1. For each taxon, an artificial entry is generated by assigning 1 to a test with $p > 50$ and zero if $p < 50$. If $p = 50$ the test is excluded.
2. This entry is identified against the matrix and filled in the column 'ID score'. The ID score is 1 if the separation is perfect. ID scores above 0.99 are considered good and are indicated in green.

These two steps are repeated twice: once with all missing values considered as positive, filled in column 'Missing as positive', and once with all missing values considered as negative, filled in column 'Missing as negative'.

If the ID score is less than 0.99, the next best identification is shown in column 'Next taxon'.

5.1.7 BLAST sequence matching

• Introduction and terminology

In this manual, only some basic concepts and terms that are essential to understand the most important settings and parameters in the BLAST implementation in InfoQuest FP will be explained in brief. For in-depth documentation, we refer to the specialized literature¹.

BLAST or *Basic Local Alignment Search Tool* (Altschul et al., 1990²) is a fast sequence comparison algorithm used

1. Korf, I., M. Yandell, and J. Bedell. 2003. BLAST. L. LeJeune (Ed.), O'Reilly & Associates, Inc., 2003.

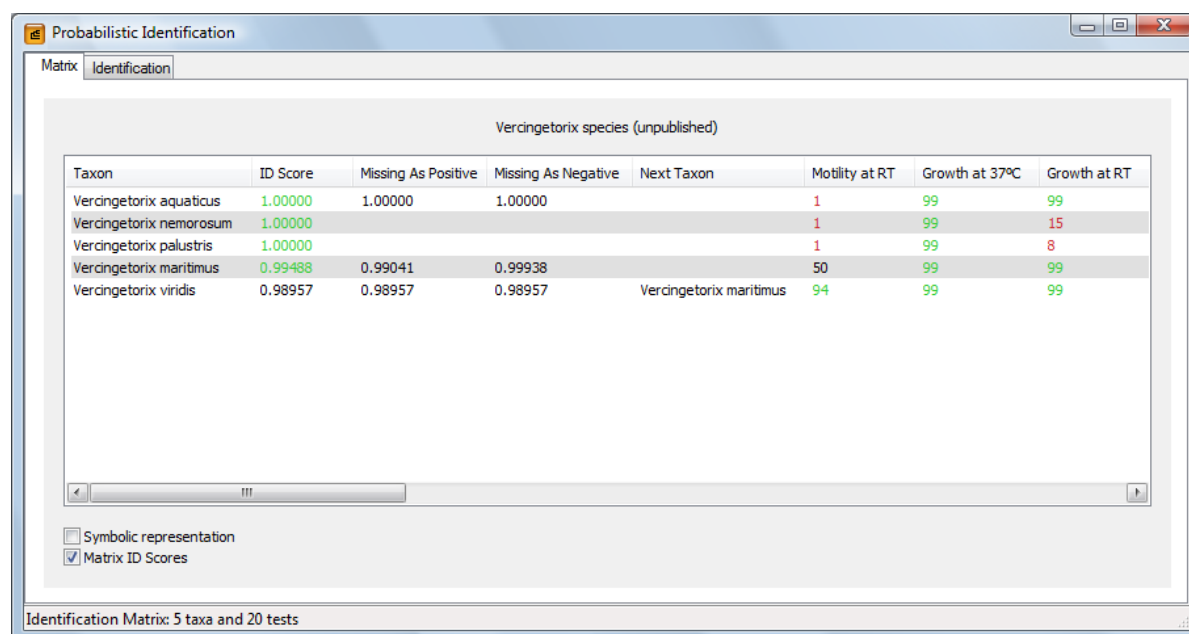


Figure 5-9. Matrix panel in the *Probabilistic identification* window.

to search sequence databases for optimal local alignments to a query sequence. To optimize the speed, a substitution matrix is generated from words of length "W". The initial search is done for a word of length "W" that scores at least "T" residues when compared to the query sequence. The "T" parameter determines the speed and sensitivity of the search. In the gapped-BLAST implementation, *word hits* are then extended in either direction, allowing gaps to be introduced, in an attempt to generate an alignment with a score exceeding a threshold value of "S". One such gapped alignment is called a *high-scoring segment pair (HSP)*. A sequence from the BLAST database that shows one or more HSPs with the query sequence is called a *hit*.

Before BLAST can be performed, the sequences in the database need to be converted into a special indexed format (BLAST database). The BLAST database is not updated when new sequences are added to the InfoQuest FP database. To update a BLAST database with new sequences, it has to be built again.

There are 5 different types of BLAST, depending on the type of sequences to compare:

- **blastp**: compares an amino acid query sequence against a protein sequence database.
- **blastn**: compares a nucleotide query sequence against a nucleotide sequence database.
- **blastx**: compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- **tblastn**: compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).
- **tblastx**: compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Due to the nature of tblastx, gapped alignments are not available with this option.

For the comparison of protein sequences or translated nucleotide sequences, a *Substitution Scoring Matrix* is required. Two families of matrices exist: *PAM matrices* and *BLOSUM matrices*:

- **PAM matrices** (Percent Accepted Mutation) are based on global alignments of closely related proteins. For example, the PAM1 is the matrix calculated from comparisons of sequences with no more than 1% divergence. Other PAM matrices are extrapolated from PAM1.
- **BLOSUM matrices** (BLOock Substitution Matrix) are based on local alignments. For example, BLOSUM 62

is a matrix calculated from comparisons of sequences with no less than 62% divergence. All BLOSUM matrices are based on observed alignments; they are not extrapolated from comparisons of closely related proteins. BLOSUM 62 is the default matrix in BLAST 2.0. Although it is tailored for comparisons of moderately distant proteins, it performs well in detecting closer relationships. A search for distant relatives may be more sensitive with a different matrix.

BLOSUM matrices with high numbers and PAM matrices with low numbers are both designed for comparisons of closely related sequences. BLOSUM matrices with low numbers and PAM matrices with high numbers are designed for comparisons of distantly related sequences.

• Creating a BLAST database

In InfoQuest FP, a BLAST database is created from a selection of database entries and a selected sequence type. The sequence type can contain nucleic acid or amino acid sequences.

5.1.7.1 To build a BLAST sequence database, first select the entries in the database to be included in the database.

5.1.7.2 Select *Identification > Create new BLAST database*. A dialog box pops up (Figure 5-10), prompting for the sequence type to use and a name and path for the BLAST database.

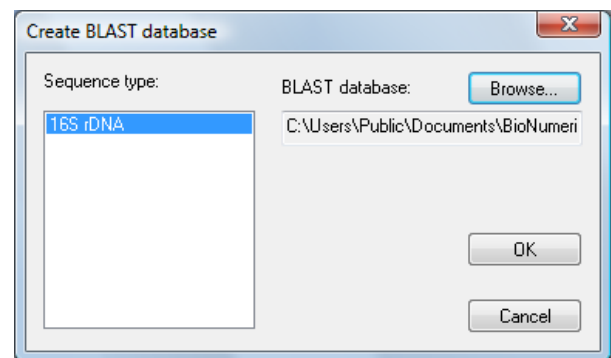


Figure 5-10. Create BLAST library dialog box.

5.1.7.3 Select the sequence type to use for creating the BLAST database.

5.1.7.4 Press the <Browse> button to define a path and a name for the database (do not use extensions).

5.1.7.5 Press <OK> to build the database. A database built from nucleotide sequences consists of three files (.nhr, .nsq and .nin). A database built from amino acid sequences also consists of three files (.phr, .psq and .pin).

2. Altschul, S.F., W. Gish, Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403-410.

• Performing a BLAST search

Once a BLAST database is created, you can select some entries to match against the BLAST database.

With one or more entries selected, choose **Identification > BLAST sequence matching**. The BLAST settings dialog box that pops up is shown in Figure 5-11.

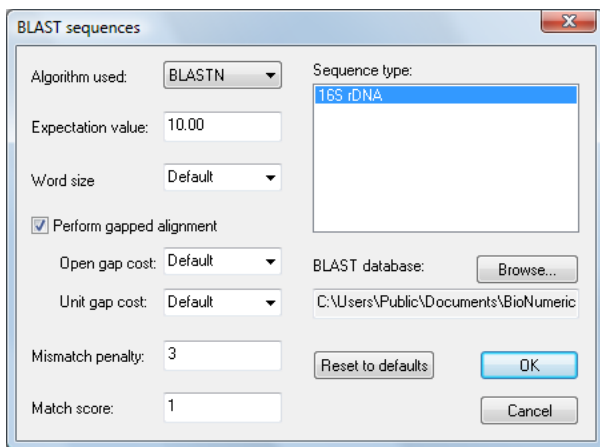


Figure 5-11. BLAST settings dialog box, BLASTN algorithm selected.

- Under **Algorithm used**, you can specify the BLAST algorithm to use. The algorithms and their meaning are described in the introduction. For a nucleotide query sequence you can choose between BLASTN, BLASTX and TBLASTX. BLASTX will only work if the selected database is a protein database. Depending on the alignment option used, some other options in the dialog box will differ.
- The **Expectation value** or E-value is the number of different alignments with scores equivalent to or better than the current one that are expected to occur in a database search by chance. The lower the E-value of an alignment, the more significant the score. The number to be filled in here is the maximum E-value allowed. The default value is 10, which is very tolerant.
- **Word size** is the size of the words to optimize the search speed. The default value is 11 for BLASTN (DNA) and 3 for BLASTP or BLASTX (protein).
- **Perform gapped alignment** is a refinement according to Altschul et al. (1997)¹, allowing gaps to be introduced in the alignment. The option does not

apply to TBLASTX. If Gapped alignment is enabled, two more options are available:

- **Open gap cost**: the cost to create a gap;
- **Unit gap cost**: the cost to increase a gap.

- **Mismatch penalty** and **Match score** are two parameters that determine the total score of a nucleotide alignment (BLASTN). The default values are -3 and 1, respectively.

- **Matrix** defines the substitution scoring matrix which will be used in case of an amino acid alignment (BNLASTX, TBLASTX, BLASTP, TBLASTN). See the introduction for an explanation. The default matrix is BLOSUM62.

*NOTE: The default settings for the **Open gap cost** and **Unit gap cost** depend on the BLAST type chosen and are different for each substitution matrix. For BLASTN (DNA), they are 5 and 2, respectively. For the different matrices they are as follows:*

- BLOSUM45: 14/2
- BLOSUM50: 13/2
- BLOSUM62: 11/1
- BLOSUM62_20: 100/10
- BLOSUM80: 10/1
- BLOSUM90: 10/1
- PAM30: 9/1
- PAM70: 10/1
- PAM250: 14/2

- Under **Sequence type** you can select the sequence type to use for the BLAST search, in case more than one sequence type is available.
- The BLAST database can be chosen with the **<Browse>** button. As a BLAST database consists of multiple files (.nhr, .nin, .nsq or .phv, .pin, .psq), only one of the files should be selected.
- With the **Reset to defaults** button, the parameters are all set back to the defaults for the selected **Algorithm used**.

5.1.7.6 Press the **<OK>** button to launch the BLAST search. The resulting BLAST report summary is shown in Figure 5-12.

1. Altschul S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.

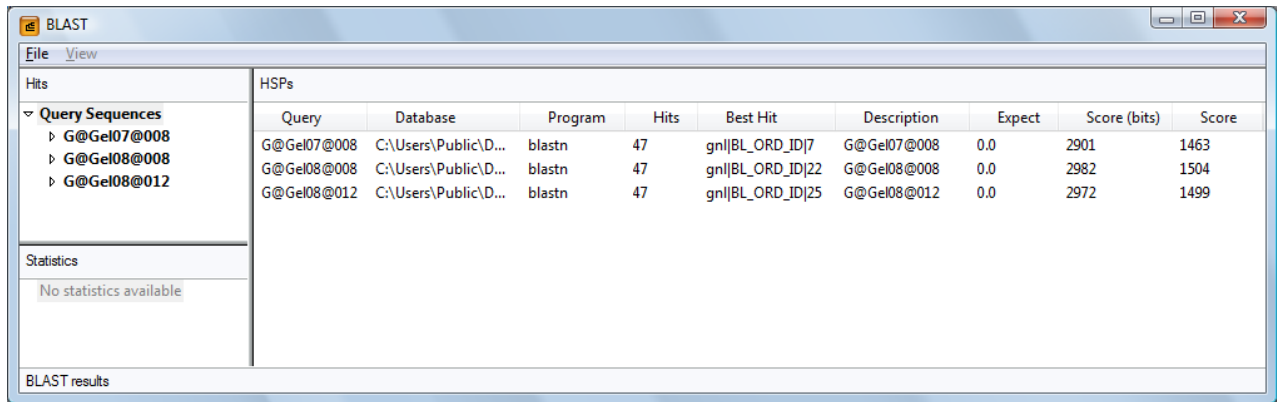


Figure 5-12. BLAST identification report window, the summary panel.

5.1.7.7 The upper left panel lists the query sequences. If you click on **Query sequences**, an overview showing the best hit (best matching database sequence) is shown in the main panel (initial view after calculation, see also Figure 5-12).

The keys are indicated in the 'Description' column. For each hit, the expectation value ('Expect'), the score and the model-independent bit score ['Score (bits)'] is given. The number of hits withheld for each query sequence is shown under 'Hits'.

5.1.7.8 In the left panel, you can click on an individual query sequence to display all the hits found for that sequence (Figure 5-13). Some statistics, parameter

settings and other information is displayed in the bottom left panel.

5.1.7.9 If you click on a hit in the right panel, the HSPs (high-scoring segment pairs) for the hit are listed in the bottom right panel.

5.1.7.10 The upper left panel is actually a tree view: you can double-click on a query sequence to display all its hits, and further on, you can double-click on a hit to display the HSPs it contains. Figure 5-13 shows an example where a hit containing two HSPs is selected.

5.1.7.11 In addition to the *Overview* panel, you can also choose the *Graphics* panel, which graphically displays

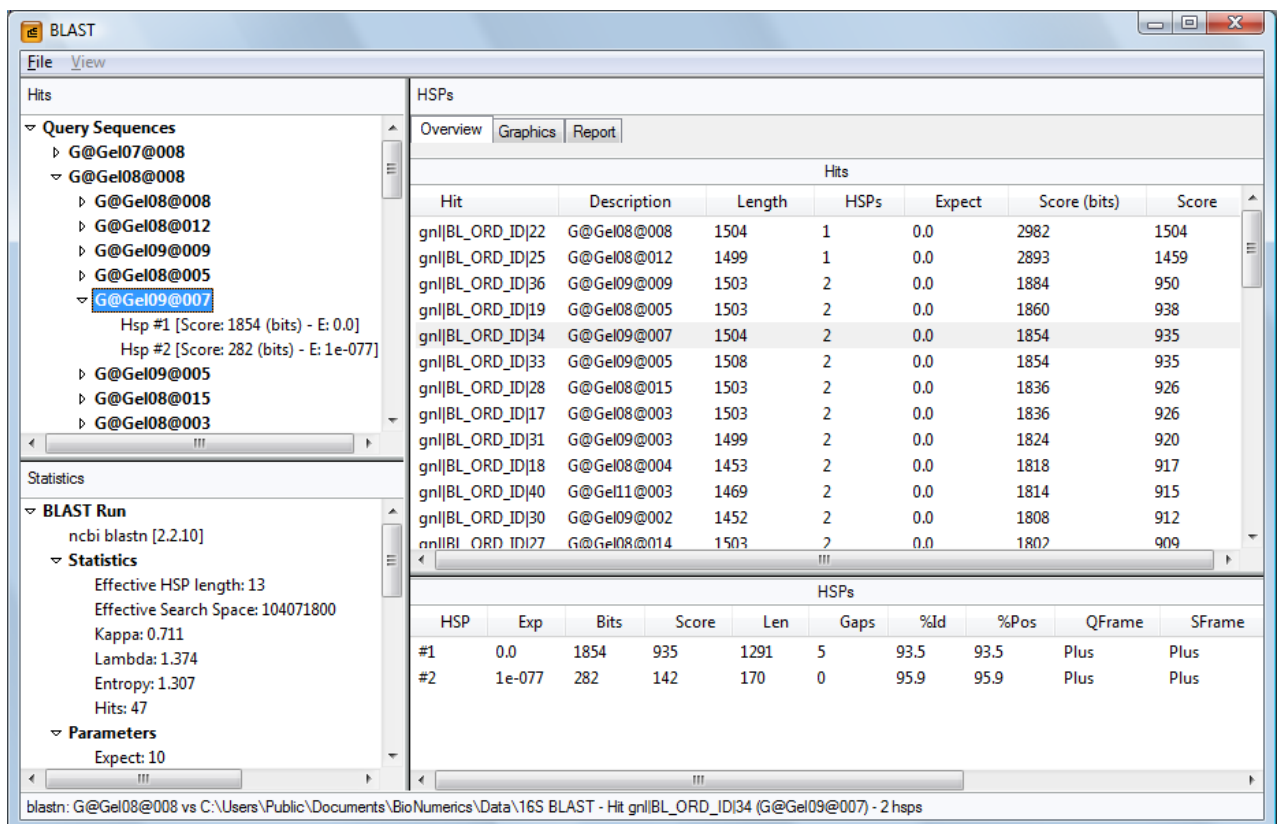


Figure 5-13. BLAST identification report window, the Overview panel.

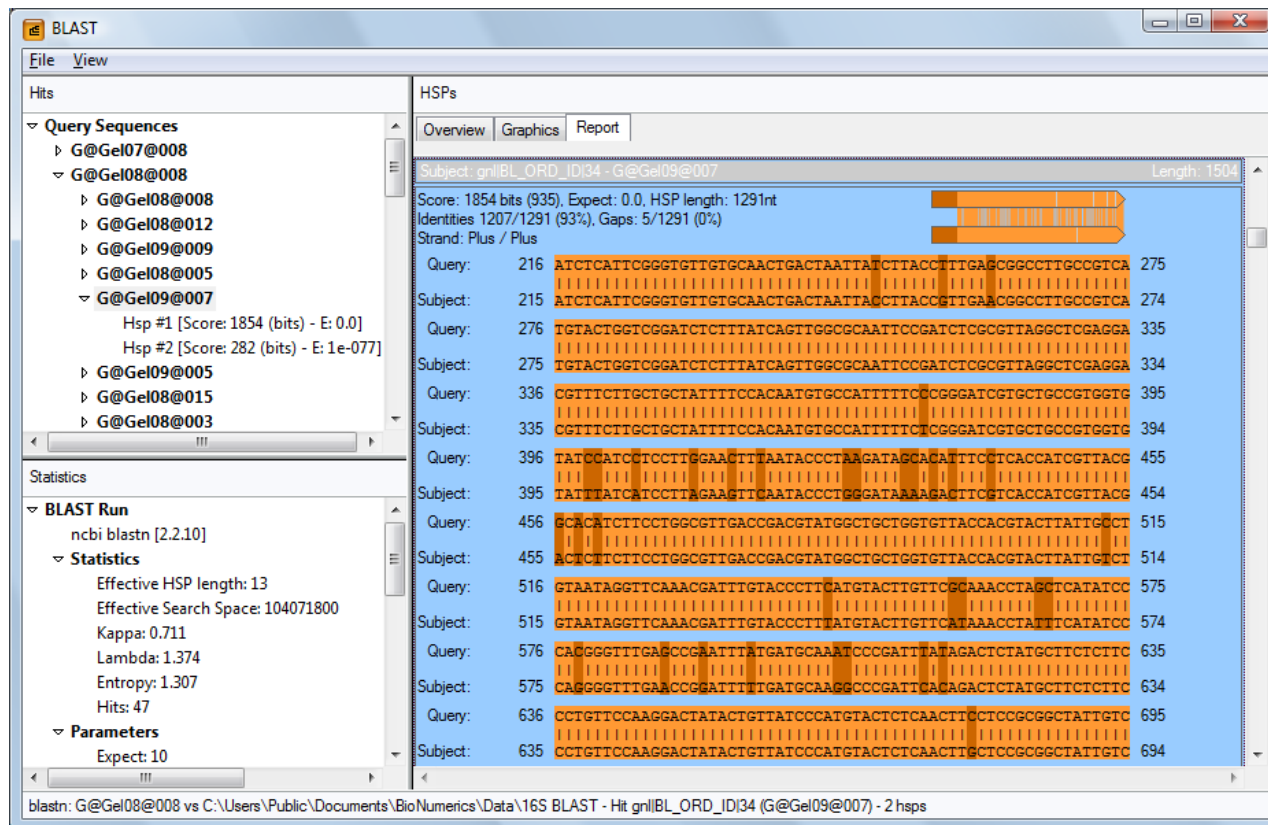


Figure 5-14. BLAST identification report window, the Report view.

the hits and the position of the HSPs on the query and database sequences. The key and length of the query sequence is displayed in the header of each hit plot.

5.1.7.12 In this view and with a hit selected in the right panel, some changes can be made to the visual appearance using the *View* menu.

5.1.7.13 A third possibility is the *Report* view (Figure 5-14), which shows the alignments for all HSPs of a selected hit in detail. Matching positions are shown in

orange, whereas mismatches are shown in a darker shading. Gaps are shown in gray.


This view also shows a small plot for each HSP between the query and hit sequence in the upper right corner. With *View > Fixed scale* disabled, the two sequences are not necessarily drawn proportionally.

5.1.7.14 The entire report can be exported as an XML file, or printed with *File > Export* or *File > Print*, respectively.

5.2 Identification using libraries

A *library* is a collection of *library units*, which in turn is a selection of database entries. A library unit is supposed to be a definable *taxon*. When generating a system for identification, a new library is first created. Then, library units are defined within that library, to which the names of the taxa are given. Within each library unit, a selection of representative entries for that taxon is entered.

5.2.1 Creating a library

5.2.1.1 In the *InfoQuest FP* main window with **DemoBase** loaded, select *Identification > Create new library* from the menu or press  in the toolbar of the *Libraries* panel.

In case of a connected database (see Section 2.3), a dialog box allows you to choose whether you want to store the library in the *Local database* (file-based), or in the *Connected database* (Figure 5-16). In the latter case, other users that are connected to the same database will be able to use the library too.

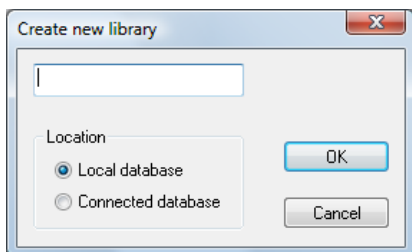



Figure 5-16. The *Create new library* dialog box.

5.2.1.2 Enter a name for the library, for example **DemoLib**.

The *Library* window of the new library appears (Figure 5-17). The *Experiments* panel (left in default configuration) shows the available experiments and the *Units* panel (right in default configuration) shows the library units defined within the library. Both panels are dockable (see 1.6.4 for display options). The layout of the *Experiments* panel can be modified by clicking on the column properties button  in the information fields header. The *Units* panel is initially empty.

Within the library, you can include or exclude experiments. Excluded experiments will not be used for identification.

5.2.1.3 Select an experiment which you do not want use for identification, for example a composite data set.

5.2.1.4 In the menu, choose *Experiment > Use for identification*. Experiments that are used for identification are marked with \checkmark ; experiments that are not used are marked with a red cross.

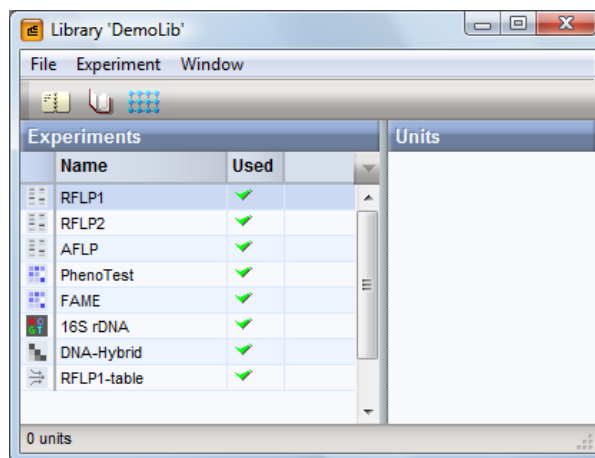



Figure 5-17. The *Library* window of a new library.


5.2.1.5 Select *File > Add new library unit* or .


5.2.1.6 Enter a name of one of the species in the database, for example *Ambiorix sylvestris*.


The library unit now shows up in the *Units* panel.

5.2.1.7 Double-click on the unit, or click on it and select *File > Edit library unit*.

The *Library unit* window which appears, is very similar to the *Comparison* window, and allows all the same clustering functions as in the *Comparison* window (see Section 4.1). This allows you to cluster the members of a library unit internally in order to check the homogeneity of a defined taxon.

5.2.1.8 In the database, select all *Ambiorix sylvestris* entries and copy them to the clipboard using *Edit > Copy selection* or .

5.2.1.9 Paste the entries in the library unit with *Edit > Paste selection* or .

5.2.1.10 Save the library unit with *File > Save* or .

5.2.1.11 Repeat 5.2.1.5 to 5.2.1.10 to create library units for the other named species.

5.2.1.12 When finished, close the library with *File > Exit*.

The library is now listed in the *Libraries* panel of the *InfoQuest FP main* window. You can open the library and add or edit units whenever desired.

5.2.2 Identifying entries against a library

5.2.2.1 In the *InfoQuest FP main* window, clear any selected entries in the database with F4.

5.2.2.2 Select a list of entries, for example all unnamed species (*Ambiorix* sp. and *Perdrix* sp.) and a few entries of the other species.

5.2.2.3 Click on **DemoLib** in the *Libraries* panel and select *Identification > Identify selected entries*.

A dialog box appears, as shown in Figure 5-18. Under Method, you can choose between *Mean similarity*, *Maximum similarity*, *K-Nearest Neighbor* and *Neural Network* (if available; see 5.2.3).

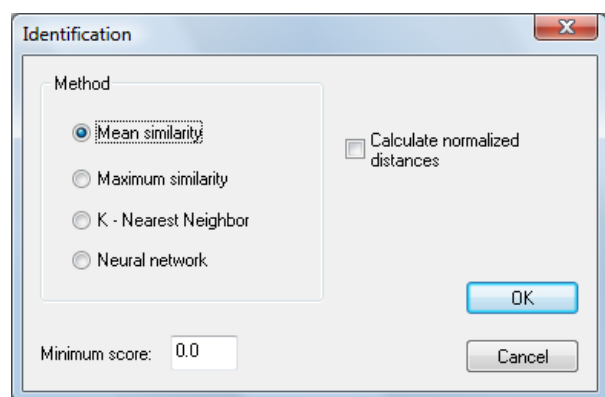


Figure 5-18. The *Identification* dialog box with the *Mean similarity* option selected.

5.2.2.4 With the option *Mean similarity*, the program calculates a similarity between the unknown entry and each entry in the library unit, and then calculates the average similarity for the entire library unit. These average similarities are then used in the identification report.

5.2.2.5 With the option *Maximum similarity*, the program will also calculate all similarities between the unknown and the library unit entries, but only the highest similarity value found is used in the identification report.

5.2.2.6 If *Mean similarity* or *Maximum similarity* is selected, an option *Calculate normalized distances* becomes available.

The *Normalized distance* is an indication for the confidence of the identification. It is achieved by comparing the average similarity between the unknown entry and the library unit's entries with the average similarity of the library unit's entries with each other. If the first value is as high or higher than the second one, the unknown entry fits well within the library unit. Thus this quality indication takes into account the internal heterogeneity of the taxon defined in the library unit.

5.2.2.7 With the option *K - Nearest Neighbor*, the user has to specify a value K, which is a number of entries from the whole library having the highest similarity with the unknown. Suppose that 10 is entered for K, the 10 best matching entries from the whole library will be retained. The library unit having the largest number of entries belonging to these K nearest neighbors is considered the best matching, and gets the highest score. The score is simply the number of entries of the library unit that belong to the K nearest neighbors.

5.2.2.8 If *K - Nearest Neighbor* is selected, an input field *K value* becomes available, where you can enter the number of nearest neighbors to look for.

NOTE: The value for K is supposed to be smaller than the number of entries contained in each of the library units. If this is not the case, the program will warn you for this conflict when the identification is executed.

5.2.2.9 The *Neural network* option is explained in detail in 5.2.3. If this option is checked, a drop-down list becomes available, showing the existing neural networks, from which you can choose one.

5.2.2.10 Optionally, a *Minimum score* can be specified. If a library unit has a score that is lower than the minimum score specified, the library unit will be listed in between brackets and in gray type in the identification report. Obviously, the score depends on the method selected. If a similarity method is selected, the score should be a floating value between 0 and 100; if *K - Nearest Neighbor* is selected, the value should be an integer value between 0 and K.

5.2.2.11 Click *Mean similarity*, check *Calculate normalized distances*, and press <OK>.

The *Identification* window appears, showing the progress of the calculations in the progress bar in the bottom of the window. Once the calculations are done, the window is divided in three panels (Figure 5-19). The *Unknowns* panel (left) lists the unknown entries that you have selected for identification. The *Matches* panel (right) lists for each experiment type (organized in columns) the library unit that matches best with the unknowns. The dockable *Details* panel (bottom panel in default configuration) shows the identification details for the highlighted unknown/experiment type combi-

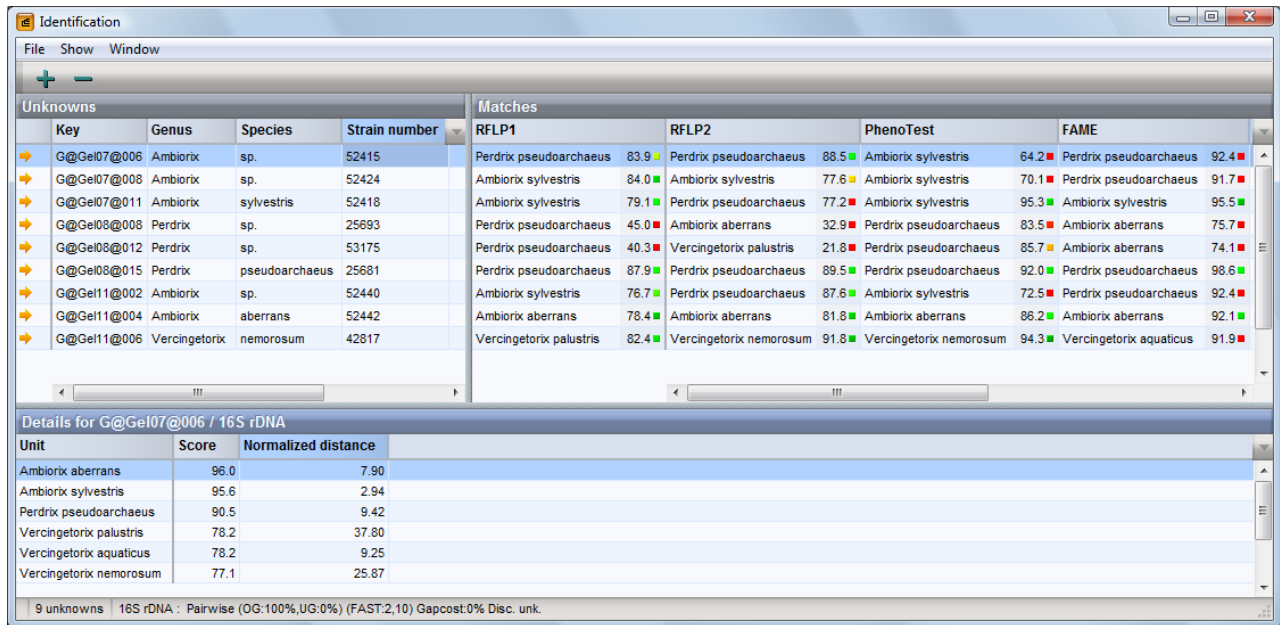





Figure 5-19. Identification window.

nation. See 1.6.4 for general display options of dockable panels.

You can move the separator lines between the panels to make optimal use of the display. Information fields in the *Unknowns* panel, *Matches* panel and *Details* panel can be displayed or hidden by pressing the column properties button  in the information fields header of the corresponding panels. For detailed information about the display options available for grid panels, see 1.6.6.

The columns in the *Matches* panel contain the name of the best matching library units and their identification score. The identification scores are the similarity values obtained using the coefficient which is specified in the settings of the experiment type (see Chapter 3). The normalized distances appear as colored squares next to the identification scores. They range from red (improbable identification) over orange, yellow (doubtful identification) to green (faithful identification).

5.2.2.12 Using *Show > Show more matches* or , the second, third, etc. best match can be shown for each unknown. To display fewer matches per unknown, select *Show > Show less matches* or .

The *Details* panel lists the best matching library units for the selected unknown/experiment type combination, ranked by their identification score. The normalized distance is here displayed as a number. Clicking in the *Unknowns* panel or *Matches* panel updates the *Details* panel with the information of the newly selected unknown/experiment type combination.

5.2.2.13 Double-clicking on a library unit within the *Details* panel opens an *Identification comparison* window.

This window is similar to a normal *Comparison* window, listing the unknown entry and the entries of the library unit.

5.2.2.14 Export the identification overview to a text file with *File > Export overview*, or create a detailed text report with *File > Export details*.

For routine identification purposes, it can be useful to store the identification results for each unknown entry. Thereto, first create a dedicated field in the database:

5.2.2.15 In the *InfoQuest FP main* window, select *Database > Create new information field* and name the new field e.g. ID result.

5.2.2.16 Click on the ID result field in the *Unknowns* panel (if the field is not displayed, press the column properties button and select it from the pull-down menu) and select *File > Fill information field*.

5.2.3 Creating a neural network

• Theory

A neural network is a means of calculating a function of which one does not have a clear description, but of which many examples with known input and output are present. Typically, the input is a set of characters for each example, and the output is the name of a group to which the example belongs. The neural network can be trained with the examples, and if the training succeeds well, the neural network can be used to perform the same calculation with other data of which the output is not known. Usually, all the examples that are fed to the neural network are divided randomly in a *training set* and a *validation set*. The training set is the part of the example set that will be used to calculate the neural

network and the validation set is the part that will be used to validate the network, i.e. check its correctness on other examples than the ones used for training.

A neural network consists of several *layers of neurons or nodes*; mostly there are 2 or 3 layers. The first layer is the *input layer*, the last one is the *output layer*, and the intermediate ones - if present - are called the *hidden layers*. Usually there are 0 or 1 hidden layers. Every neuron or node has a value that is calculated by the neural network. The values of the neurons in the input layer are simply the input of the function. Every neuron in the successive layers takes the value of all the neurons in the previous layer and performs a calculation on it, to obtain its own value. Mostly this calculation is a weighted sum, in which the weights can be different for every neuron. That value will be used by neurons in consecutive layers. The number of nodes in the input layer is equal to the number of characters available for the data set, i.e. the number of characters in the experiment which is used to calculate the neural network. The number of nodes in the output layer is equal to the number of groups defined in the identification system. The number of nodes in the hidden layer - of any - can be chosen and is dependent on the nature and complexity of the data set and identification system.

During the training cycle the input of a known example is fed in the neural network and the calculation is performed. Initially the calculated output will most likely be very different from what it should be. The weights between every pair of consecutive neurons are then slightly adjusted, so that the calculated output becomes closer to the correct output. This is done using a process called *back-propagation*. This means that in the output layer the errors are calculated, which are the difference between the correct output and the calculated output. These errors are then back-propagated to the neurons in previous layers by multiplying the error by the weight that connects two neurons, and summing for every neuron. The weights of the neurons are then adjusted by the error times a number called the *learning ratio*. Furthermore, the weight correction of the previous training cycle times a number called the *momentum* is added. The higher the learning ratio and the momentum, the faster the training, but the higher the risk that the error doesn't decrease.

This training process is repeated many times (typically a few thousand times), each time with another known example chosen randomly from the training set. After sufficient *iterations* the calculated outputs will be very close to what they should be, provided that the number of layers and number of nodes per hidden layer is chosen correctly. A higher number of layers and/or neurons means that training and calculation will take longer, so a trade-off has to be made. Furthermore, there is a danger of *overtraining* when there are too many layers and/or neurons, which means that the neural network would be very good for the examples, but not at all for other inputs. To have an estimate of this, one usually divides the known examples in a *training set* and a *validation set*. The validation set is not used for training,

but only to check how well the neural network performs on this set. If it is significantly worse than for the training set, one knows that there are too many layers and/or neurons.

•Application

A neural network can be applied to many problems, such as control theory, character recognition, statistical analysis and distinguishing patterns. In practice, a neural network is very useful to set up an identification or recognition system based upon complex data sets in which it is not easy or impossible to identify discriminatory keys based upon conventional methods such as calculation of similarity using coefficients, cluster analysis, principal components analysis etc. An important requirement for successfully applying neural networks is that the example data set is sufficiently large and that many examples are present for each group of the identification system.

In our software, it is used for determining to what predefined group or taxon an unknown database entry belongs, based on measurements that could be a character set or a fingerprint. This is thus an example of distinguishing patterns. In this case the output of the neural network is n values, where n is the number of predefined groups. Every group is given a number from 1 to n , and thereby corresponds to one of the outputs. The higher a value in the output, the more likely the sample belongs to that group. In the training and validation set the output values are zero, except for the output that corresponds to the group, which will be one. After the training has succeeded one can use it with measurements on unknown samples. In these, the highest output will be decisive for what group it is.

In InfoQuest FP, the choice in hidden layers is limited to none or one, because more hidden layers usually don't give any advantage. In extensive tests, one hidden layer was always sufficient, in many cases no hidden layer worked just as well. The number of nodes in the hidden layer can be chosen if the user wants to do so. If the user doesn't specify this, the neural network will start without a hidden layer. If it doesn't succeed in lowering the error, a hidden layer will be created. If it still doesn't lower the error, the hidden layer is expanded until the error is below a predefined threshold.

The learning rate and momentum cannot be specified. Instead we fixed these to 0.5 and 0.1 respectively, because in our tests these values gave the optimal trade-off between speed and success.

To train a neural network, a library must be present. See 5.2.1 to create a new library. To obtain a reliable neural network, each of the library units must have sufficient members, many more than just two or three. The number of entries required also depends on the heterogeneity of the group: the more heterogeneous a group, the more entries that will be needed to create a reliable neural network.

5.2.3.1 Double-click on a library to open it.

5.2.3.2 Select *Experiment > Train neural network* or



. A dialog box pops up, listing the existing neural networks for this database, if any.

5.2.3.3 To add a neural network, press **<Add>**.

The *Neural network training* dialog box appears, as shown in Figure 5-20.

5.2.3.4 Under *Select experiment to be used in the neural network*, you can select the experiment to train the neural network.

5.2.3.5 With *Validation samples*, it is possible to specify the percentage of the library entries (i.e. the example data) to be used as validation set. By default this value is 25%.

5.2.3.6 With *Max. number of iterations*, you can specify the maximum number of training cycles to be performed. By default this value is 20000.

5.2.3.7 *Number of hidden nodes* allows you to manually specify the number of hidden nodes. If you leave this field blank, the program will automatically determine whether a hidden layer is required, and if so, the optimal number of hidden nodes. If you enter zero, no hidden layer will be created.

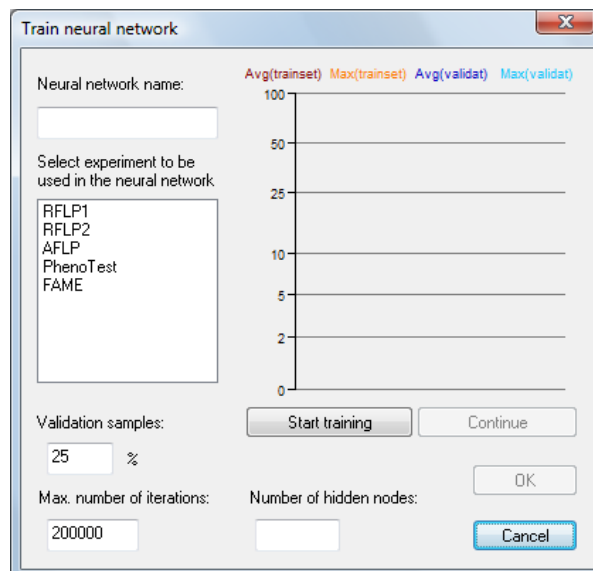


Figure 5-20. The *Neural network training* dialog box.

5.2.3.8 Enter a name for the neural network under *Neural network name*. You can use the name of the experiment type.

5.2.3.9 When all parameters are entered, press **<Start training>** to start the training process. Depending on the size of the library, the training process can take several minutes. An animation of the progress of the training is shown in the x-t diagram (Figure 5-20).

5.2.3.10 During the training, it is possible to interrupt or abort the process by pressing **<Stop>**.

5.2.3.11 If you wish to resume the training process, press **<Continue>**. The program will continue the iteration process until the maximum number is achieved.

5.2.3.12 To save the neural network, press **<OK>**.

5.2.3.13 To identify database entries using a neural network, proceed as explained in 5.2.2.1 to 5.2.2.11, but choose *Neural network* instead. A drop-down list showing the existing neural networks will become available, allowing you to choose one of them for the identification.

5.3 Decision networks

5.3.1 Introduction

Decision networks are operational workflows that carry out [logical] operations and/or actions on the database. The networks consist of *Operators* as building blocks, that form the *Nodes* of the network:

- **Input operators** retrieve specific data, usually experimental data from the database;
- **String, Value and Sequence operators** perform an action on data types, for example find subsequences, count bands, or evaluate character values;
- **Boolean operators** have one or more binary states as input and can e.g. combine them into a new binary state or a string;
- **Output actions** can perform a specific action on the database, for example, write the result of a decision into a database field.

Decision networks should be seen as a construction kit that allows you to build your own automated decision or action workflows, with practically endless possibilities. They can be used to make decisions, predict features, perform queries, fill in fields, create graphs and plots, and much more.

5.3.2 Creating a new decision network

In the default configuration of the *InfoQuest FP main* window (see Section 1.6 for features and display options of the *InfoQuest FP main* window), the *Decision networks* panel is seen as a tab behind the *Comparisons* panel. Click on the tab to bring the *Decision networks* panel to the top (Figure 5-21).

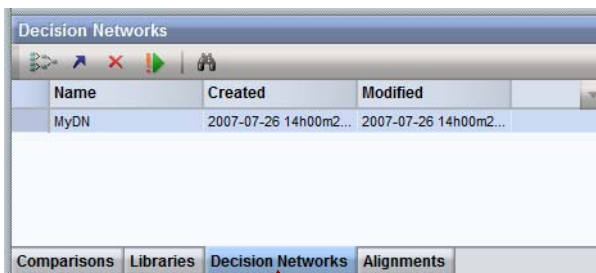




Figure 5-21. The *Decision networks* panel in the *InfoQuest FP main* window.

5.3.2.1 Press the  button to create a new empty decision network. Enter a name in the dialog box that pops up, for example, **MyDN**.



The new decision network is now listed in the panel. When a decision network is opened, it contains by default the current selection of entries. Therefore, it is practical to make a selection of entries you want to use in the decision network before opening it.

5.3.2.2 As an example, select all entries except the ones marked as “STANDARD” (see 2.2.6 to 2.2.9 about selecting entries in the database).

5.3.2.3 Open the decision network by pressing the  button or by double-clicking on its name.

The empty *Decision network* window looks as in Figure 5-22 in its default configuration. The window contains 4 panels, of which the main *Network* panel displays the network scheme. The *Operators* panel lists a tree of all operators that are available to construct the decision network (the building blocks). In *Node properties* panel, the properties and data of the current selected node is given. The *Entry data* panel lists the entries currently used in the network and their selection status. In the right hand subpanel, the output(s) from the network are listed for the entries (currently empty).

Similar as in a *Comparison* window (see 4.1.4), it is possible to add or remove entries from the list.

5.3.2.4 To add entries, copy selected entries in the *InfoQuest FP main* window or in the *Comparison* window using the  button, and paste them in the *Decision network* window using the  button.

5.3.2.5 To remove entries from the decision network, first select the entries to remove, and then press the

 button.

NOTE: As opposed to a comparison (see 4.1.3), the selection of entries in a decision network is dynamic: each time you open or run the decision network, it will act on the current selection.

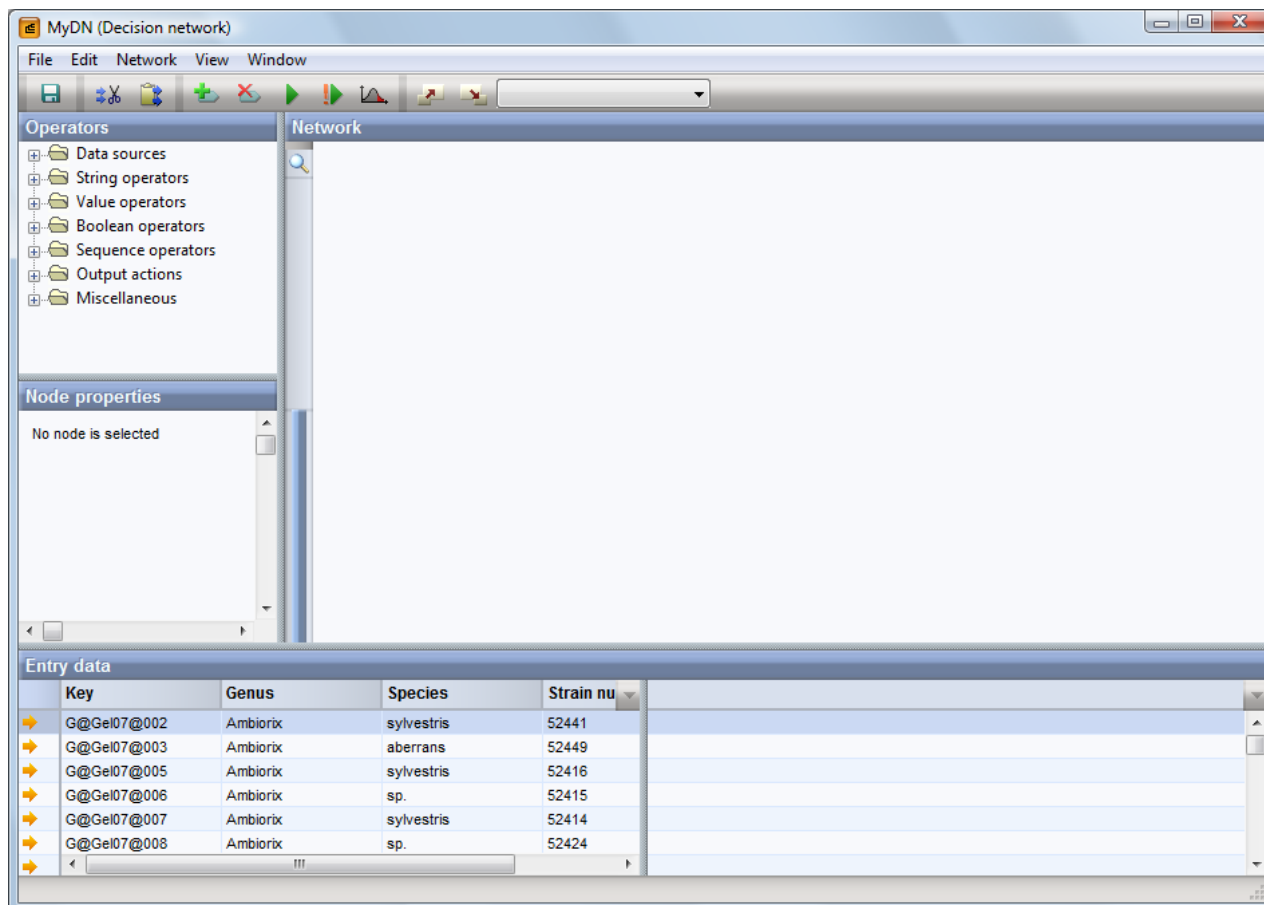




Figure 5-22. The *Decision network* window with a new empty decision network.

5.3.3 Operators

Operators are the building blocks of a decision network. They can be categorized in different groups according to their function. These groups are represented in the expandable tree in the *Operators* panel. Each operator requires a compatible output from another operator as input and delivers a result as output to the network (only operators of type *Data sources* do not require output from an operator).

- **Data sources:** These operators request data components from the database or from the user and deliver it to the network. The component can be a database field, attachment, fingerprint fields, fingerprint bands, a character value, or a sequence. A special subcategory contains the *Fixed values*, which can either be a constant value or a constant string. The subcategory *User prompt* contains operators that prompt the user to enter information of a defined data type.
- **String operators** perform an operation on a string. They include finding a text match, defining regular expressions, comparing two strings, concatenating strings, getting substrings, or converting a string into a value. Note that more powerful string operators exist specifically for sequences (*Sequence operators*).
- **Value operators** perform an operation on one or more values. The result can be a value (in case of calculations and functions), a boolean (*Comparison*, *Value range*), or a string (*Value to string*).
- **Boolean operators** have one or more boolean states as input. Besides the basic operators *AND*, *OR* and *NOT*, there are more advanced boolean operators such as *TRUECOUNT*, which evaluates the number of true states between multiple outputs (see 5.3.7). The categorical combiner will evaluate multiple boolean outputs and list the true output(s) as its own output. *Boolean to string* and *Boolean to value* operators will convert a boolean state into a string or value, respectively.
- **Sequence operators** are specifically designed for sequence data. *Find subsequence* searches for a subsequence in a sequence data type, allowing for mismatches, gaps, and IUPAC notations. *Amino acid translation* translates a nucleic acid sequence into an amino acid sequence using a defined translation table.
- **Output actions** perform an action on the database, which can be writing a result in a field, changing the selection status according to the result, writing a value in a character type experiment, writing into a sequence, or writing to an attachment. Output actions


are only executed if the Execute button  is pressed.

- **Charts** will produce a chart in the *Chart & statistics* window (see Section 4.12) from the outputs of selected nodes. A chart will only be created if the Execute button  is pressed.

- *Miscellaneous* contains a *Duplicator*, which allows one to duplicate a selected operator, e.g. to split up complex networks. The *Is present* operator returns whether a data component is present for an entry (a character, a sequence, or a fingerprint). The *Execute script* option is beyond the scope of this manual.

5.3.4 Building a decision network

As an example, we will create a simple decision network that discriminates between the three genera in *DemoBase*, based upon the 16S rDNA sequences.

5.3.4.1 In the newly created decision network, open the **Data sources** group in the *Operators* panel by clicking on its  icon.

5.3.4.2 Double-click on *Sequence*, which opens a *New operator* box (Figure 5-23).

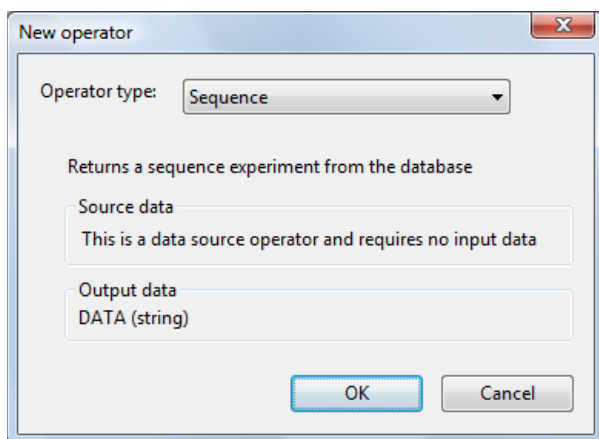


Figure 5-23. The *New operator* dialog box for a *Sequence* operator type.

The dialog box describes the operator and mentions the source data needed and the output data delivered to the network.

5.3.4.3 Press **<OK>** to edit the *Node properties* for the sequence input node (Figure 5-24).

5.3.4.4 Optionally, enter a *Name* for the node. If this is not done, it will be named automatically using a sequential number. In this example, we can enter e.g. '16S' as name.

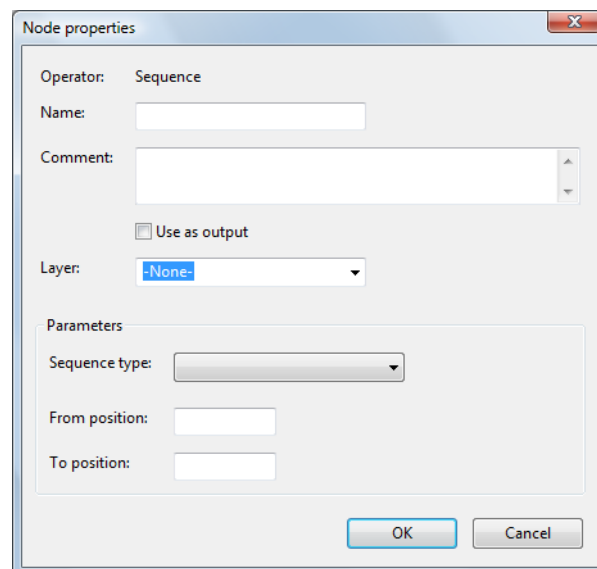


Figure 5-24. The *Node properties* dialog box for a sequence input operator.

5.3.4.5 A *Comment* field can also be entered; this field will be shown in the *Node properties* panel (5.3.2).

When *Use as output* is checked, the result of the node is shown in the *Entry data* panel (right hand subpanel; see 5.3.2). This option makes little sense for a *Sequence* operator type, as only sequences would be returned.

The *Layer* option is explained later (see 5.3.6).

5.3.4.6 Under *Parameters*, select the *Sequence type* to feed the sequences. In *DemoBase*, there is only one sequence type available, **16S rDNA**.

With *From position* and *To position*, a range within the full sequences can optionally be specified to deliver to the network. This option only makes sense if the sequences are pre-aligned, which is not the case in this database.

After pressing **<OK>**, the network contains one node, i.e. '16S'.

5.3.4.7 If you click on an entry in the *Entry data* panel, the node and the *Node properties* panel are updated with the sequence data of the highlighted entry.

5.3.4.8 Select the node '16S' in the network (a selected node is bordered by a red line).

5.3.4.9 Open the **Sequence operators** group in the *Operators* panel and double-click on *Find subsequence*.

The *New operator* box appears, showing that this operator delivers multiple output data:

- **IsMatch** is a boolean reporting whether the subsequence occurs;

- **Start** and **End** are value-type data that return the start and end positions of the matching subsequence on the sequence;
- **Seq1** and **Seq2** return the query sequence and the matching subsequence on the data sequence, respectively. The latter can be different as the operator allows mismatches and gaps.

5.3.4.10 Press **<OK>** to edit the node properties (Figure 5-25).

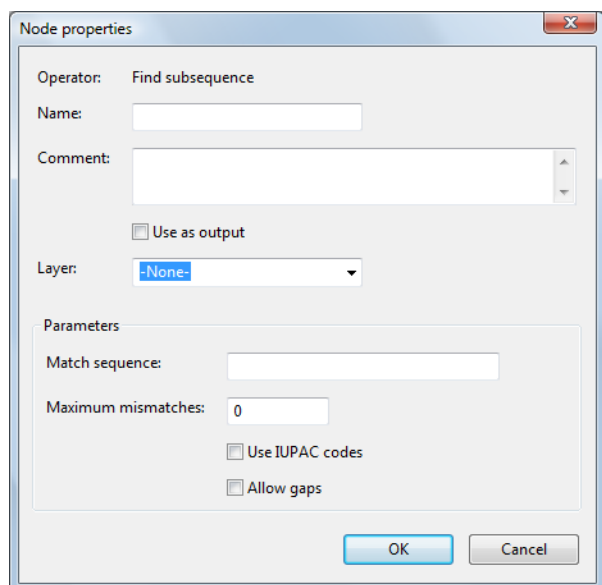



Figure 5-25. *Node properties* dialog box for a *Find subsequence* type operator.

5.3.4.11 Enter as *Name* 'Signature 1', and as *Comment* 'Recognizes Ambiorix', and check *Use as output*.

5.3.4.12 Enter 'GGGTGTAG' as *Match sequence*, with zero mismatches allowed.

5.3.4.13 Press **<OK>** to confirm the node properties. The network is now ready to produce a first result.

5.3.4.14 In the *Decision network* window, press the  button to calculate the network.

The *Entry data* panel now contains one output column, 'Signature 1', showing a boolean TRUE or FALSE for each entry. All Ambiorix entries have the boolean TRUE, the others FALSE.

In the decision network (Figure 5-26), the node 'Signature 1' is marked with a green flag, indicating that it is an output node, resulting in a column in the *Entry data* panel.

For each highlighted entry in the *Entry data* panel, the output node is either colored green (true) or red (false). The percentage of true and false entries in the entry data panel is indicated as a green and red bar, respectively. In

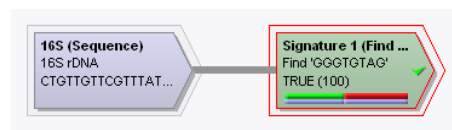


Figure 5-26. Simple decision network with one boolean output.


addition, the percentage of selected entries is indicated with a blue bar.

For each highlighted sequence, the *Node properties* panel displays detailed information about the selected node: the input parameters as a first group and the output data as a second group.

5.3.4.15 Continue to build the network by selecting the data node '16S' again and adding a second 'Find subsequence' node to it.

5.3.4.16 Enter 'Signature 2' as *Name*, 'Recognizes Vercingetorix' as *Comment*, and 'CGATCTCACG' as *Match sequence*, with zero mismatches allowed.

5.3.4.17 Check *Use as output* and press **<OK>**.

5.3.4.18 Press the  button to calculate the network. The network now looks as in Figure 5-27.

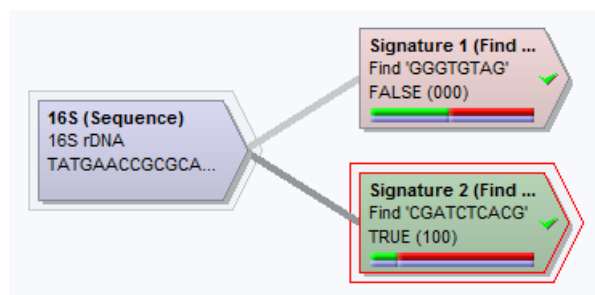


Figure 5-27. Decision network with two boolean output nodes.

A second column, 'Signature 2', is added to the *Entry data* panel. The *Ambiorix* entries have 'Signature 1' true and 'Signature 2' false, whereas the *Vercingetorix* entries have 'Signature 2' true and 'Signature 1' false. Entries of *Perdrinx* have both signatures false.

Although this type of network might allow you to predict the properties for new and unknown entries, the output is not very descriptive or easily interpretable. We will now turn the network into a more descriptive result.

This time we will make use of a duplicator node. A duplicator node duplicates a node in the network, i.e. one output parameter from it, which can be chosen. This tool is useful if a node is to be used in more than one independent operations in the network. Theoretically one could branch the different operations from the same node, but the network could easily become unsurvey-

able. As we need to check 'Signature 1' and 'Signature 2' to be both negative for Perdrix, we will duplicate both booleans and branch the new operation from there.

5.3.4.19 Select the subsequence search node 'Signature 1' and, in the *Operators* panel, double-click on the *Duplicator* operator in the **Miscellaneous** group.

5.3.4.20 Select *Signature 1 (IsMatch)* as the parameter to be duplicated and press <OK>.

5.3.4.21 In the next dialog box, leave *Name* and *Comment* fields blank and leave *Use as output* unchecked, as this is an intermediate node.

5.3.4.22 Make sure *Hide link* is checked and press <OK>. This will place the node on a new line, separated from the other operators.

5.3.4.23 Select the duplicator node (red border) and create a new node using the *NOT* operator from the **Boolean operators** group.

5.3.4.24 Enter 'No Signature 1' as *Name* for this node and press <OK>.

5.3.4.25 Repeat actions 5.3.4.19 to 5.3.4.24 for node 'Signature 2'. However, for the *NOT* boolean node, enter this time 'No Signature 2'.

5.3.4.26 Select both boolean nodes 'NOT' by clicking the first and then, while holding down the CTRL key, clicking the second. Both nodes are now bordered in red.

NOTE: You can also select multiple nodes by dragging the mouse over the nodes to select.

5.3.4.27 Combine the two nodes with an *AND* operator from the **Boolean operators** group.

5.3.4.28 Enter 'Perdrix' as *Name* for this node and press <OK>.

The network now looks as in Figure 5-28.

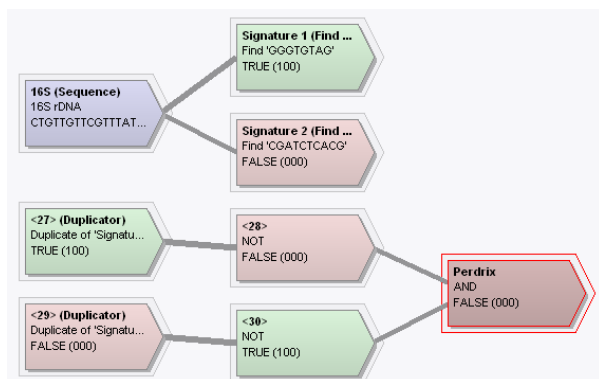


Figure 5-28. Decision network with duplicated nodes.

We now already have one boolean node called 'Perdrix'; we still need to create similar nodes for the two other groups.

5.3.4.29 Select both nodes 'Signature 1' and 'No Signature 2' and combine them with a boolean operator *AND*.

5.3.4.30 As *Name* for this node, enter 'Ambiorix' and press <OK>.

5.3.4.31 Select both nodes 'Signature 2' and 'No Signature 1' and combine them with a boolean operator *AND*.

5.3.4.32 As *Name* for this node, enter 'Vercingetorix' and press <OK>.

The network now looks as in Figure 5-29. This network contains cross-branching operators, but is still surveyable thanks to the duplicator nodes.

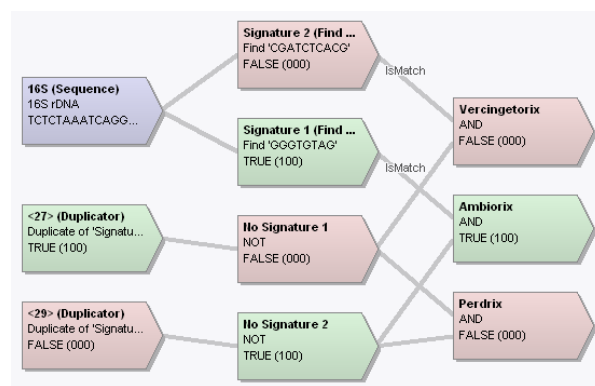


Figure 5-29. Decision network leading to three categorical boolean nodes.

Note that the connectors branching from the 'Signature 1' and 'Signature 2' nodes contain a tag "IsMatch", because these nodes have multiple outputs. "IsMatch" is the boolean that tells whether there is a match. The duplicator nodes do not contain this tag, because a duplicator can only contain one parameter from its parent, for which we chose "IsMatch".

If you click on any of the entries in the network, the end node with its name should be TRUE (green) whereas both others should be FALSE (red).

NOTES:

(1) One of the advantages of a decision network over the advanced query tool (see 2.2.9) is that you can inspect the state for each evaluation or action in the network on the fly. The Node properties panel thereby shows all the details for the non-boolean operations.

(2) If you click on any node, all the connector lines in the network that connect the node to either parent or descendent nodes are shown as bold dark lines. This makes it easier to inspect dependencies.

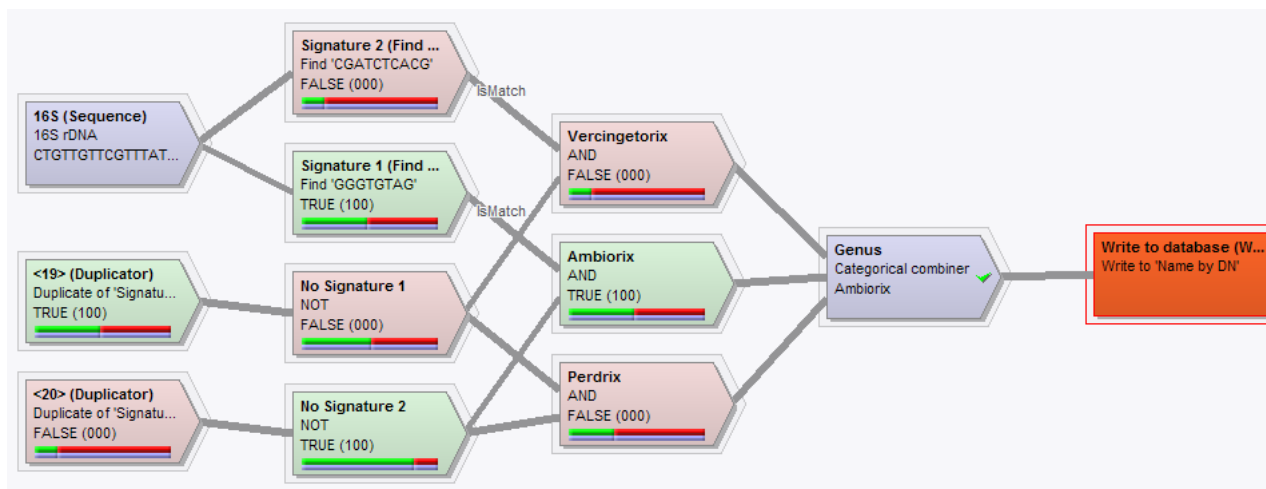



Figure 5-30. Decision network that decides between three groups and writes the output to the database.

5.3.4.33 In the *Decision network* window, press the  button to calculate the network.


The network now contains 3 mutually exclusive boolean nodes (only one out of the three can be true). Such nodes can be combined into a categorical set, i.e. a set of multiple states (categories), of which one state is true for each individual. We will combine these categories into one node that tells the name of the genus:

5.3.4.34 Select all three end nodes ('Vercingetorix', 'Ambiorix' and 'Perdrinx') and create a new node using the boolean operator *Categorical combiner*. The output for this node is a string, containing the category that is true.

5.3.4.35 Enter 'Genus' as *Name* and check *Use as output*.

The parameter *Single choice (highest confidence only)* will allow the node to take out the category with the highest confidence value (see 5.3.7), in case more than one category turns out to be true. In this specific case, we do not need to enable this option.

5.3.4.36 Press <OK> to confirm the node properties.

5.3.4.37 In the *Decision network* window, press the  button to calculate the network.


The *Entry data* panel now contains a new output column, 'Genus', showing the genus name for all entries selected in the decision network.


Finally, we will add an output action operator to write the result of the network in a database field.

5.3.4.38 In the *InfoQuest FP main window*, add a new information field, e.g. 'Name by DN'.

5.3.4.39 In the *Decision network* window, select the categorical combiner node 'Genus' and double-click the *Write to field* operator from the **Output actions** group.

5.3.4.40 In the *Node properties* dialog box, enter a name (optional), e.g. 'Write to database', and select field 'Name by DN' as the *Database field name*.

The finished decision network now looks as in Figure 5-30. The nodes that perform an output action are orange, to indicate that these nodes can perform changes to the database. For safety reasons, the output actions are not executed automatically when the network is calculated using the  button.

5.3.4.41 To calculate the network and execute the output action(s), press the  button.

To alert you that the network will now perform changes to the database, the following warning box appears (Figure 5-31).

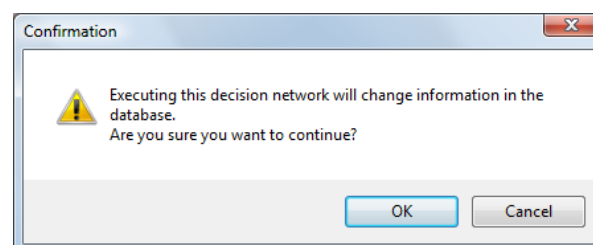


Figure 5-31. Warning box if a network is executed that contains output nodes.

5.3.4.42 Press <OK> to confirm the execution. When finished, the database field 'Name by DN' contains the genus names as defined by the decision network.

NOTE: The entire decision network could have been build without the duplicator nodes. In that event, more cross-connectors would exist and the network would be less surveyable (Figure 5-32). However, the result would be the same.

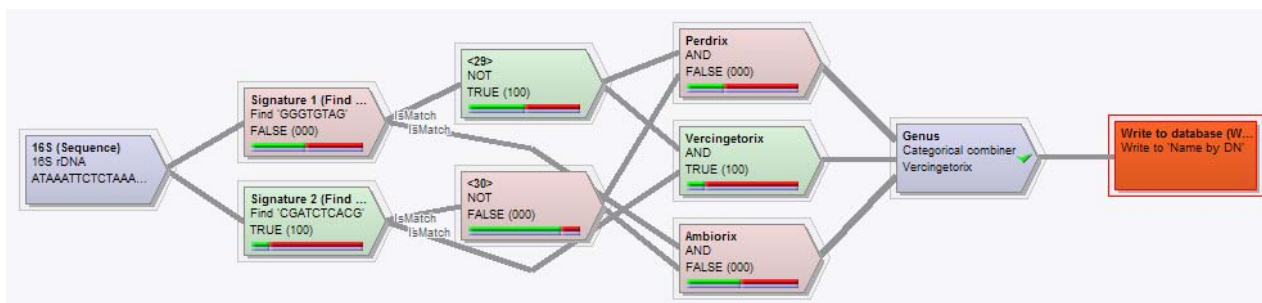




Figure 5-32. The same network as in Figure 5-30, without duplicator nodes.

5.3.5 Display and output options for decision networks

5.3.5.1 It is possible to zoom in or out on a decision network using *View > Zoom in*  and *View > Zoom*

out  or by extending or compressing the zoom slider in the *Network* panel (see 1.6.7 for instructions on the use of zoom sliders).

5.3.5.2 Select *File > Print* to print the decision network. One will be prompted for the printer to use and basic printer settings. Only the content of the *Network* panel is sent to the printer.

The content of the *Network* panel can also be copied to the clipboard for import in other programs, e.g. to create reports:

5.3.5.3 Select *File > Copy to clipboard (metafile)* to export the network as a metafile. Paste the clipboard in a program such as MS Word or PowerPoint to see the exported graphical representation of the network.

5.3.5.4 Select *File > Copy to clipboard (bitmap)* to export the network as a bitmap. The program will prompt for the bitmap resolution, enter e.g. 1,000. The clipboard can now be pasted in a graphics editor such as Adobe Photoshop.

5.3.6 Working with layers in a decision network

The use of *layers* in a decision network is to structure and organize separate subflows in complex networks. Whereas by means of *duplicators* one can duplicate a node to start at a new line and continue the flow from there, a *layer* is a subflow of the network that can be visualized separately from the others.

A layer can only be created along with the creation of a new node. To illustrate the use of layers we will add a subflow to the existing decision network **MyDN**. Suppose we will evaluate two character values to define resistance of the entries.

5.3.6.1 Create a 'Character value' node by double-clicking on the *Character value* operator in the **Data sources** group.

5.3.6.2 In the *Node properties* dialog box, type 'Resistance' in the input field *Layer*.

5.3.6.3 Choose **PhenoTest** as *Experiment* and select 'c4' as *Character*.

5.3.6.4 Enter 'Char 1' as *Name*, and press <OK>.

5.3.6.5 Repeat actions 5.3.6.1 to 5.3.6.4 for a second character 'c12' from **PhenoTest**, entering 'Char 2' as *Name*.

5.3.6.6 Select character value node for 'Char 1' and double-click on the *Value range* operator in the **Value operators** group.

5.3.6.7 In the *Node properties* dialog box, select **Resistance** from the drop-down box under *Layer*.


5.3.6.8 Enter '2' as *Minimum value* and press <OK>.

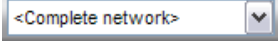
5.3.6.9 Repeat actions 5.3.6.6 to 5.3.6.8 for node 'Char 2', entering the same minimum value '2'.

5.3.6.10 Select both 'Value range' nodes and combine them with a boolean operator **AND**.

5.3.6.11 In the *Node properties* dialog box of the 'AND' node, specify **Resistance** as *Layer* and enter 'Multiresistant' as *Name*.

5.3.6.12 Check *Use as output* and press <OK>.

5.3.6.13 Calculate the network with ; a new column is added to the *Entry data* output subpanel.

The toolbar in the *Decision network* window contains a drop-down box  that allows you to select a layer to visualize. By default, the complete network is visualized.

5.3.6.14 Select **Resistance** from the drop-down box. Only the nodes that belong to the resistance flow are now shown (Figure 5-33).

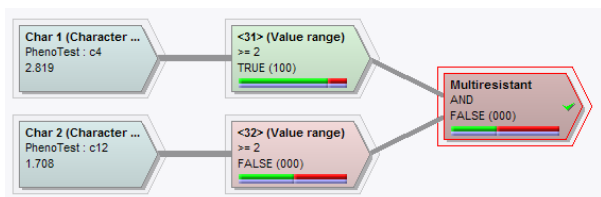


Figure 5-33. Decision network with one layer visualized (see text for explanation).

Nodes from a layer can be connected to other nodes belonging to the complete network or a different layer. In that case, the subpart of the complete network or the other layer(s) that contribute to the outcome of the layer are shown along with the layer.

5.3.6.15 As an example, select *Complete network* again from the drop-down box.

5.3.6.16 Select the boolean nodes 'Multiresistant' and 'Ambiorix', and connect them with a boolean 'AND' node.

5.3.6.17 In the *Node properties* dialog box, specify a *Name* 'Multiresistant Ambiorix' and choose *Resistance* as *Layer*.

5.3.6.18 Select *Resistance* from the drop-down box. The network now looks as in Figure 5-34.

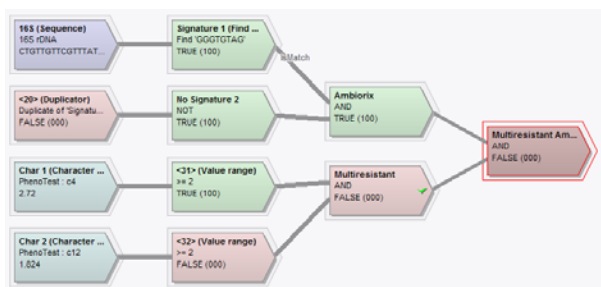


Figure 5-34. Example of a layer visualized along with nodes that contribute to the output.

5.3.7 Using confidence values

Confidence values are useful in cases where answers cannot be formulated clearly as either yes or no. For example, in the network we created in 5.3.6, the decision whether to label the entries as "resistant" or not, depends on the fact that a value is more or less than 2. In reality, however, states cannot be clearly defined from non-binary measurements. Therefore, it is possible to enter a *Fuzzy zone* in a value range operator.

5.3.7.1 In the network layer created in 5.3.6, double-click on the value range node 'Char 1' to re-edit it.

5.3.7.2 In the *Node properties* dialog box, enter '1' as *Fuzzy zone*. Press **<OK>** to confirm the change.

Repeat this for the value range node connected to 'Char 2', also entering '1' as *Fuzzy zone*.

The fuzzy zone extends equally to both sides of the limit(s) entered for the range (see Figure 5-35). For example, if you entered 2 as limit, the answer will still be FALSE for all entries that have the value below 2 and TRUE for all those that have more than 2. However, all values between 1.5 and 2.5 will exhibit a confidence that is bigger than 0 and lower than 100. A value of exactly 2 will have a confidence of 50.

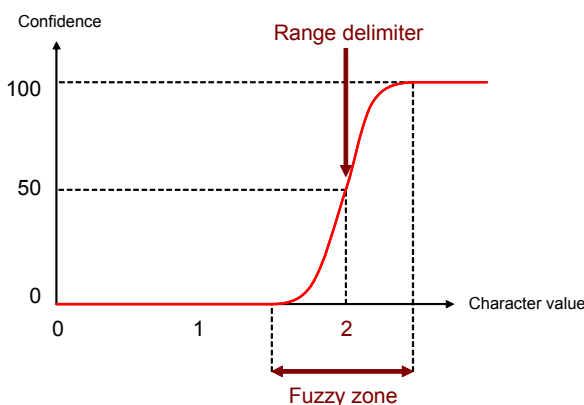



Figure 5-35. Graphical illustration of the effect of a fuzzy zone on the confidence value of a boolean decision.

After recalculating the network with , each output TRUE or FALSE contains a value that ranges between 0 and 100%.

The confidence values are preserved throughout the flow of the network: when, for example, two booleans are combined with AND (Figure 5-34), the lowest of the two values is used. In case two booleans are combined with OR, the highest of the two values is retained.

Confidence values are also optionally used in the categorical combiner operator (5.3.4.34). If the *Single choice (highest confidence only)* option is enabled, and in case multiple categories appear to be true, the network will look for the category with the highest confidence value to retain. If this option is not enabled, multiple true categories will be listed together, separated by semicolons.

5.3.8 Building decisions relying on multiple states

In paragraph 5.3.4 we have already described the *Categorical combiner* operator (5.3.4.34), which evaluates a number of boolean nodes and uses the name of the most true node as its output. In a number of cases, however, it might be required to build decisions upon conditions such as "at least x true states", and/or "at most y true

states". This can be achieved using the *TRUECOUNT* operator.

The *TRUECOUNT* operator (Figure 5-36) combines a number of boolean nodes, and is set to TRUE if at least x booleans are true (*At least true*) and/or at most y booleans are true (*At most true*) ($y \geq x$).

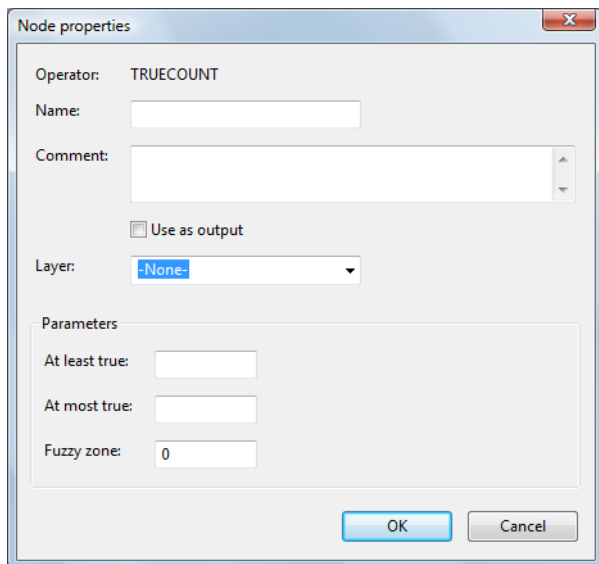


Figure 5-36. The *TRUECOUNT* operator node properties.

For example, suppose that a bacterial strain is multidrug resistant if it exhibits resistance for at least 12 antibiotics out of a set of 20. A network can be built that calculates multidrug resistance by using value range nodes for each of the antibiotics, and combining them using a *TRUECOUNT* operator that has "12" as *At least true* parameter.


The boolean operator *TRUECOUNT* also makes use of confidence values. Optionally, one could enter '11' for *At least true* and use '2' for *Fuzzy zone*, which would mean that entries with resistance towards 12 antibiotics are shown as multidrug resistant with 100% confidence, and entries with resistance towards 11 antibiotics are shown as multidrug resistant with 50% confidence.

NOTE: Since the TRUECOUNT operator only recognizes integer values as input, the fuzzy zone value has to be an even number of at least 2. A fuzzy value of 1 would be divided into 0.5 at either side of the delimiter (see Figure 5-35) and would be truncated to zero.

5.3.9 Creating charts from a decision network

InfoQuest FP can plot the results of a decision network in a chart (scatterplot, bar graph, contingency table, ...) by using its *Chart and statistics tools* (see Section 4.12).


There are two ways a chart can be generated from a decision network:


- The result of every node that is specified as *Output node* can be plotted using the  button. Depending of the content of the node, the *Chart and statistics* window automatically generates the suitable plot type.
- Charts can also be created as an *Output action*, in which case a chart is automatically generated when the network is executed.

To illustrate the manual creation of a graph from a node, we will make some graphs from nodes in the network we created in the previous paragraphs.

5.3.9.1 Make sure *Complete network* is selected from the drop-down menu of the *Decision network* window (see 5.3.6.13).

5.3.9.2 Specify the categorical combiner node 'Genus' as an output node (of not already specified this way) by enabling *Use as output* in the *Node properties* dialog box.

5.3.9.3 Recalculate the network with .

5.3.9.4 Select the 'Genus' node and press the  button or right-click on the node and select *Plot in chart window*. A bar graph appears, showing the relative occurrences of the three genera (Figure 5-37).

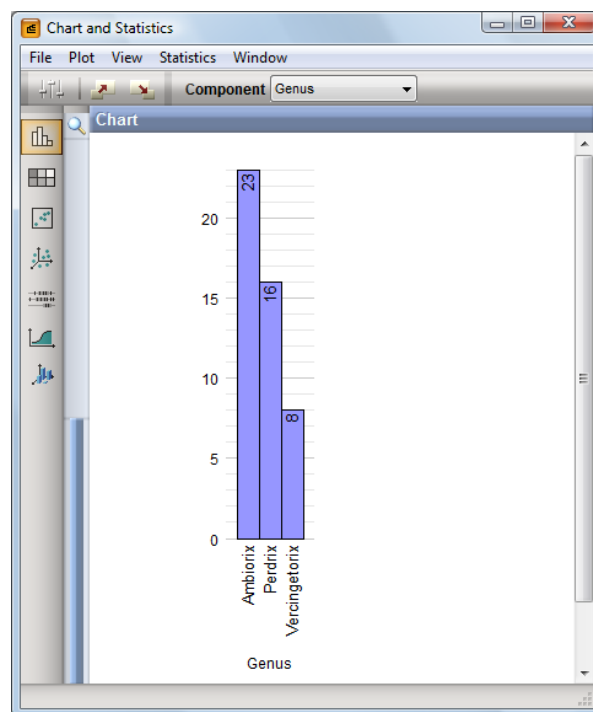




Figure 5-37. Bar graph popped up from categorical combiner output node.

5.3.9.5 Close the *Chart and Statistics* window.

5.3.9.6 Select the data source node 'Char 1' and specify it to be an output node, similar as in 5.3.9.2.

5.3.9.7 Recalculate the network with .

5.3.9.8 Press the  button or right-click on the node and select *Plot in chart window*. A cumulative distribution of the character values appears (Figure 5-37).

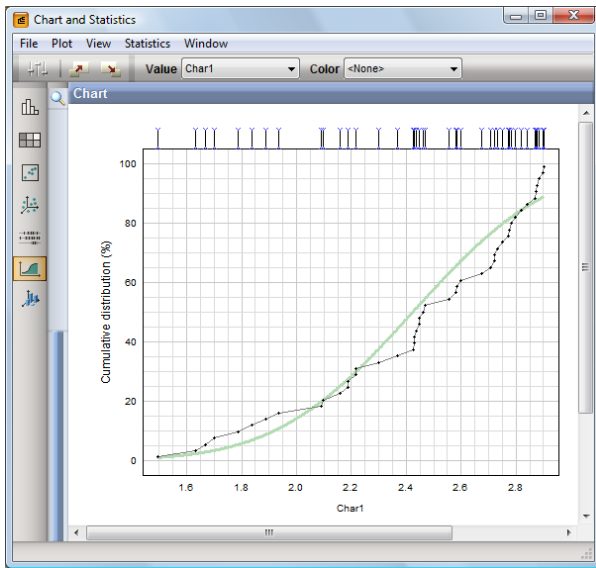



Figure 5-38. Cumulative distribution popped up from a character value output node.

5.3.9.9 Close the *Chart and Statistics* window.

To create a graph each time the network is executed, we will use the nodes 'Char 1' and 'Char 2' as input, and create a scatterplot from them.

5.3.9.10 Select both nodes 'Char 1' and 'Char 2', and create a new output action node by double-clicking on *2D scatter plot* in the **Charts** group.

5.3.9.11 Execute the network by pressing the  button. A scatterplot is generated in a *Chart and Statistics* window, comparing the two characters for each entry used in the network (Figure 5-39).

5.3.9.12 Close the *Chart and Statistics* window.

5.3.10 Executing a decision network from the *InfoQuest FP main window*

Once build and saved, a decision network can be used as a tool to perform certain manipulations on the database

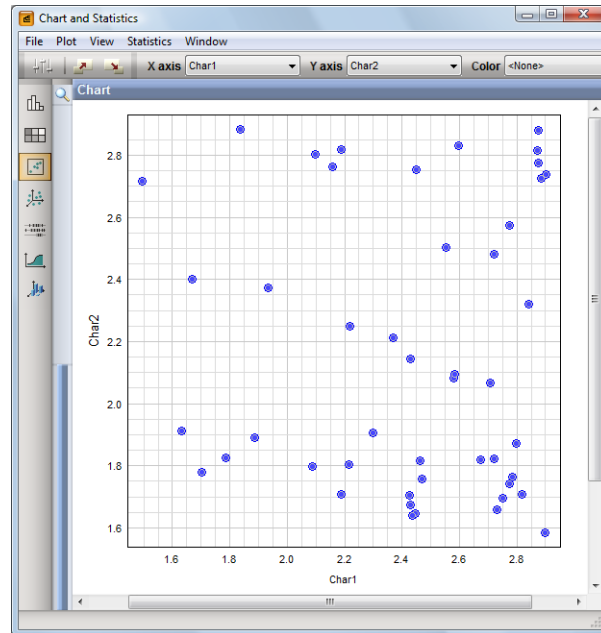



Figure 5-39. Scatterplot graph obtained from an output action node.

in an automated way. These manipulations are the output actions defined in the network.

For example, if you save and quit the network created in the previous paragraphs, you will notice that the network is listed in the *InfoQuest FP main window*, in the *Decision networks* panel (see Figure 5-21).

5.3.10.1 You can directly execute the network from the *InfoQuest FP main window* by pressing the  button in the toolbar of the *Decision networks* panel.

A dialog box pops up (Figure 5-40), offering three choices for executing the decision network: on *All entries*, on *Currently selected entries only*, or on *Non-selected entries only*.

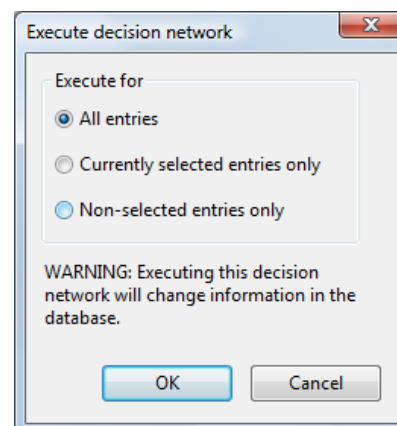


Figure 5-40. Choices for executing a Decision network directly from the *InfoQuest FP main window*.

5.3.10.2 Check *All entries* and press <OK>. All output actions defined in the network are executed on all entries in the database: output strings are written in defined information fields, and charts are generated.

One obvious application for executing a decision network directly from the *InfoQuest FP main* window is to use it as a kind of an advanced query tool: the output action operator *Change selection* will change the selection status of the entries into selected if the input for the 'Change selection' node is TRUE, or non-selected if the input for the 'Change selection' node is FALSE.

5.3.11 Decision trees

A decision tree is mainly used to build bifurcating decision schemes based on a number of TRUE/FALSE evaluations. A decision tree does not provide possibilities which cannot be achieved in a normal decision network as described in the previous paragraphs. However, it allows the scheme to be presented in a more intuitive way and can be a suitable asset for e.g. taxonomic identification schemes.

The operators to build a decision tree can be found in the **Boolean operators** group, where they are grouped in a category **Decision trees**. The tree always starts with a *Decision tree root*, which will contain the output of the tree as well.

In the example below, we will create a decision tree that performs the same task as the decision network created in 5.3.4, i.e. identifying entries at the genus level based upon a signature sequence.

5.3.11.1 Create a new decision network as described in 5.3.2.1 and further. Enter **Decision tree** as *Name*.

5.3.11.2 Select a number of database entries and open the decision tree (see 5.3.2.3).

5.3.11.3 Under **Boolean operators**, open the category **Decision trees** and double-click on *Decision tree root*. Enter 'Genus name' as *Name* and select *Use as output*.

The data evaluations and the definitions of the criteria have to be done in a separate flow of the network, and the decision tree is built on the outcome of those evaluations. We will now create the criteria needed for the identification tree (see also 5.3.4).

5.3.11.4 Under **Data sources**, double-click on *Sequence* to create a sequence data source node.

5.3.11.5 Select **16S rDNA** as *Sequence type* and press <OK>.

5.3.11.6 With the 'Sequence' node selected, double-click the *Find subsequence* operator in the **Sequence operators** group. Enter 'Ambiorix signature' as *Name*, and 'GGGTGTAG' as *Match sequence*.

5.3.11.7 Again with the 'Sequence' node selected, create a second 'Find subsequence' node. Enter 'Vercingetorix signature' as *Name*, and 'CGATCTCAG' as *Match sequence*.

Back in the decision tree, we will create a bifurcation based upon the presence of the Ambiorix signature, as follows:

5.3.11.8 Select both the decision tree root and the 'Ambiorix signature' node and create a *Bifurcation* (under **Boolean operators** > **Decision trees**).

The *New operator* dialog box looks as in Figure 5-41.

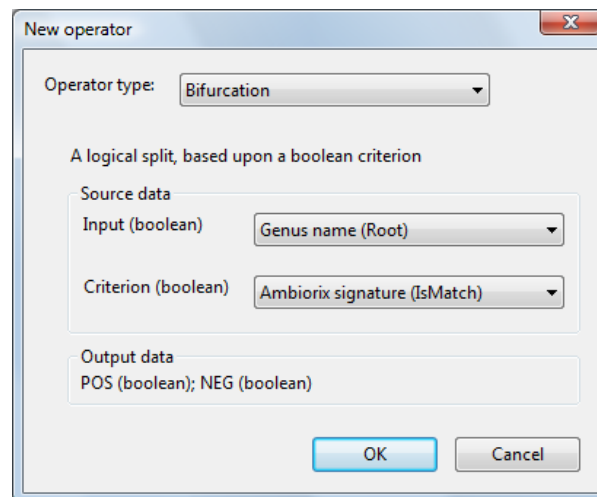


Figure 5-41. *New operator* dialog box for a bifurcation.

As *Input (boolean)*, the tree root should be selected, whereas as *Criterion (boolean)*, the Find subsequence node 'Ambiorix signature' should be selected.

As a *Name*, you can enter 'Is Ambiorix'. The decision tree network now looks as in Figure 5-42.

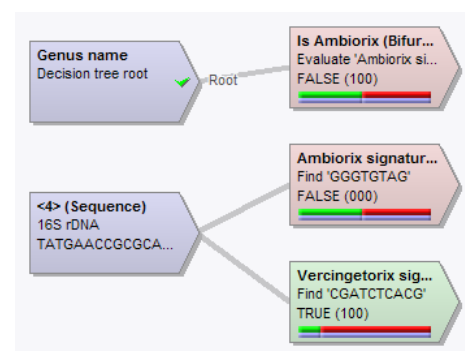


Figure 5-42. Decision tree in construction, containing one bifurcation.

We will now create two *leaves* branching off from this bifurcation: one for true and one for false.

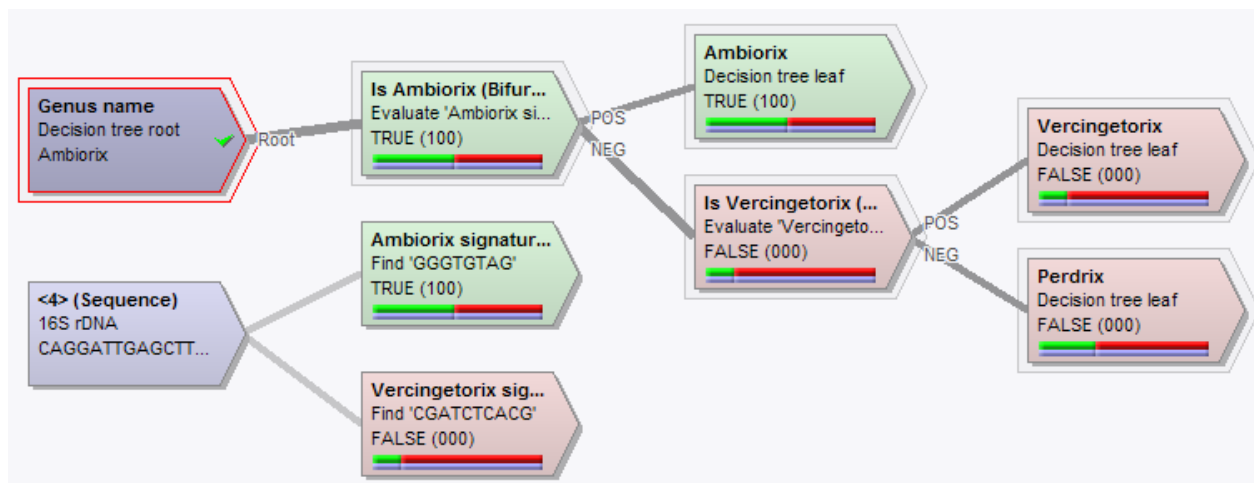


Figure 5-44. Decision tree that decides between three possible genera, in two dichotomic steps.

5.3.11.9 Select the bifurcation 'Is Ambiorix' and double-click on the *Decision tree leaf* operator. The *Source data* can either be *Is Ambiorix (POS)* or *Is Ambiorix (NEG)*, standing for a TRUE and FALSE condition, respectively.

5.3.11.10 Select *Is Ambiorix (POS)* as *Source data*, and enter 'Ambiorix' as *Name*.

5.3.11.11 To create the second bifurcation, select the bifurcation 'Is Ambiorix' again and create a second decision tree leaf.

5.3.11.12 Select *Is Ambiorix (NEG)* as *Source data*, and enter 'Not Ambiorix' as *Name*.

This is an example of the simplest decision tree that exists, evaluating one criterion and identifying entries as belonging to a taxon or not (Figure 5-43).

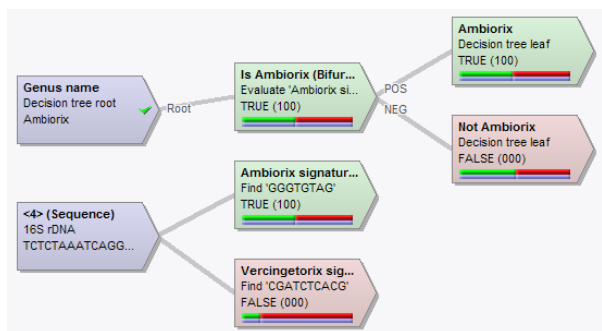



Figure 5-43. Simple decision tree based upon one criterion.

5.3.11.13 If you calculate the decision tree with , you will notice that for each entry you select, the decision tree root either shows 'Ambiorix' or 'Not Ambiorix'.

As we defined the root as an output node (see 5.3.11.3), the right subpanel of the *Entry data* panel also displays this result for all the entries used in the network.

To create a decision tree that identifies the three genera, we have to insert further criteria at the first bifurcation, rather than a leaf:

5.3.11.14 Delete the leaf node 'Not Ambiorix'.

5.3.11.15 Select both the bifurcation 'Is Ambiorix' and the Find subsequence node 'Vercingetorix signature', and create a new bifurcation from these nodes.

5.3.11.16 Select *Is Ambiorix (NEG)* as *Input (boolean)* and *Vercingetorix signature (IsMatch)* as *Criterion (boolean)*.

5.3.11.17 Enter 'Is Vercingetorix' as *Name*.

To finalize the tree so that it identifies the three genera, we further have to insert two leaves:

5.3.11.18 Select the 'Is Vercingetorix' bifurcation node and create a new decision tree leaf node.

5.3.11.19 As *Input (boolean)* for the leaf, select *Is Vercingetorix (POS)*.

5.3.11.20 Enter 'Vercingetorix' as *Name*.

5.3.11.21 Finally, select the 'Is Vercingetorix' bifurcation node again and create a second decision tree leaf node.

5.3.11.22 As *Input (boolean)* for the leaf, select *Is Vercingetorix (NEG)*.

5.3.11.23 Enter 'Perdrix' as *Name*.

The finalized tree now looks as in Figure 5-44, which is simpler to interpret than the comparable decision network depicted in Figure 5-30.

Optionally, you can add an output action node to the root, writing the result to a database field, as explained in 5.3.4.40.

The criteria defined for this decision tree could have been created in a separate layer (see 5.3.6), which would

allow us to better separate the tree from its criteria and display either the tree or its criteria.

6. INFOQUEST FP 2D

6.1 Analyzing 2D gels

6.1.1 Proteomics in a broader context: the InfoQuest FP Platform

The InfoQuest FP 2D application, developed for the analysis and comparison of two-dimensional, spot-oriented bitmap files, is physically an integral part of the InfoQuest FP software suite. Therefore, it is available as a module of InfoQuest FP, referred to as the *2D gel types* module or **InfoQuest FP 2D**. Along with two other applications that act as plugins of InfoQuest FP: GeneMaths XT and Kodon, the InfoQuest FP software forms the basis for an integrated bioinformatics platform: the *InfoQuest FP Platform*. The obvious advantage of integrating a 2D image analysis application within a broad bioinformatics platform, is the possibility to link genomics, proteomics, metabolomics and phenotypic data in one powerful database.

By its integration in the InfoQuest FP Platform, the combined use of InfoQuest FP 2D and the GeneMaths XT software (Bio-Rad) will allow the co-evaluation of the expression of specific proteins with the simultaneous expression of homologous genes as evidenced by microarray experiments. Also, the proteins detected and identified can be linked to DNA and protein sequences that are kept in the InfoQuest FP database and which are amenable to all kinds of sequence analysis tools such as structural comparison, chromosome mapping, vector cloning, primer design, secondary structure analysis using the Kodon software (Bio-Rad), which is also fully integrated in the InfoQuest FP Platform.

Another advantage of integrating 2D gel analysis in a broad bioinformatics analysis platform is the availability of numerous powerful analysis tools. These include cluster analysis of organisms or samples based upon their (combined) experimental data, or cluster analysis of characters such as genes or protein spots; a wide range of dimensioning techniques such as principal components analysis, discriminant analysis, MANOVA, or Self-Organizing Maps, are all available to compare in two ways: organisms/samples amongst each other, or protein spots and genes amongst different samples.

In InfoQuest FP, all biological experiments are functionally classified in six different classes, called *experiment types*:

- **Fingerprint types:** Any densitometric record seen as a one-dimensional profile of peaks or bands can be considered as a fingerprint type. Fingerprint types can be derived from TIFF or bitmap files as well, which are two-dimensional bitmaps. The condition is that one must be able to translate the patterns into densitometric curves.
- **2D gel types:** Any two-dimensional bitmap image seen as a profile, spots or defined labelled structures. Examples are e.g. 2D protein gel electrophoresis patterns, 2D DNA electrophoresis profiles, 2D thin layer chromatograms, or even images from radioactively labelled cryosections or short half-life radiotracers.
- **Character types:** Any array of named characters, binary or continuous, with fixed or undefined length can be classified within the character types. The main difference between character types and electrophoresis types is that in the character types, each character has a well-determined name, whereas in the electrophoresis types, the bands, peaks or densitometric values are unnamed (a molecular size is NOT a well-determined name!).
- **Sequence types:** Within the sequence types, the user can enter nucleic acid (DNA and RNA) sequences and amino acid (protein) sequences.
- A fifth type, **matrix types**, is not a native experiment type, but the result of a comparison between database entries, expressed as similarity values between certain database entries.
- **Trend data types:** Reactions to certain substrates or conditions are sometimes recorded as multiple readings in function of a changing factor, defining a trend.

Each experiment type is available as a module of the InfoQuest FP software. In the following chapters, **InfoQuest FP 2D** will refer to the *2D gel types* module within InfoQuest FP.

Through its integration with InfoQuest FP, the InfoQuest FP 2D software is a perfect tool to be used in applications such as proteomics, protein expression studies, drug discovery, functional genomics and proteome mapping, metabolomics, protein interactions research, signal transduction pathways, molecular oncology and clinical screening.

6.1.2 Data sources for InfoQuest FP 2D

InfoQuest FP 2D can handle a variety of file formats including 8-bit, 12-bit, and 16-bit. The software is able to cope with images of any size and OD depth. The software can be used with a variety of staining and labelling protocols, using different support materials. For the capture of 2D images a variety of densitometers, cameras or radiation detection devices are used. These

devices do not only differ in cost but also in resolution and dynamic range. Examples of commonly used equipment for the digitalization of gels are:

- Polaroid photo
- Autoradiography film
- CCD (video) cameras
- CCD document scanners
- Fluorescence cameras
- Phosphor-imagers
- Laser densitometers
- ...

Expensive laser densitometers have mostly been replaced by document scanners and video cameras. Most document scanners have a rather limited OD range (of about 2.0 OD units covered by a 8-bit gray scale or 256 gray values). These values, however, cover quite well what is generally obtained using the most commonly used staining methods or with X-ray film irradiation. New types of document scanners or imagers may offer a considerably higher dynamic range performance (up to 3 OD units) with 12-,14- or 16-bit gray scale levels. The InfoQuest FP 2D software is able to import the TIFF files from all these types of scanners.

6.1.3 Applications for InfoQuest FP 2D

The most obvious application for InfoQuest FP 2D is the analysis of 2D protein gel electrophoresis experiments. Separating, detecting, and quantifying proteins is the main purpose of modern proteomics research. In order to correctly identify changes of protein expression levels (e.g. of disease related proteins), it is extremely important to use procedures that will allow high resolution separation and a proper staining or labelling method.

2D gel electrophoresis separates proteins based on their iso-electric points (pI values) in a so-called first dimension performed in a carrier that contains an IPG (immobilized pH gradient), followed by a second dimension in a carrier that separates on molecular weight in a traditional electrophoresis process (second dimension). There are currently two techniques available for the first dimension of 2D gel electrophoresis: NEPHGE and IPG. NEPHGE stands for non-equilibrium pH gradient electrophoresis, and is a technique with high resolution but lower levels of reproducibility, while IPG (immobilized pH gradient) has a lower resolution but is more easy to handle. The lack of resolution of the latter technique has been circumvented by the use of multiple gels with more limited pH ranges. By using InfoQuest FP 2D it is possible to assemble these different pH ranged gels into a synthetic gel that will contain the overall information for the subject being studied.

When coupled to existing databases of known proteins, characterized according to the above mentioned parameters, 2D gel electrophoresis can be used to identify cellular proteins, new cell or tissue components or to

detect alterations in protein expression, metabolic or physiological activities and will also assist in the quantitative and qualitative comparison of gels run on samples obtained under different conditions.

By using the InfoQuest FP 2D software all relevant and supporting information can be stored in a structured database format. This information can be used for selection and for comparison purposes.

6.1.4 Automated workflow for experiments with repeats

A plugin tool is available for the InfoQuest FP 2D software which allows the analysis of experiments containing repeated gels to be automated. Average spot profiles are calculated automatically for the repeated gels, and standard deviations are used to enhance the reliability of the analysis and data mining steps.

The plugin is available with InfoQuest FP and can be installed as described in . The plugin tool also comes with an easy and comprehensive tutorial manual, illustrating the analysis of repeat experiments.

6.1.5 Automated workflow for multiplex experiments (DIGE)

The plugin tool described in 6.1.4 also provides tools for the automated analysis of multiplex 2D gels such as DIGE.

This plugin is available with InfoQuest FP and can be installed as described in . The plugin tool also comes with an easy and comprehensive tutorial manual, dealing with the analysis of DIGE gels.

6.1.6 Getting started with InfoQuest FP 2D

The next paragraphs will guide the user stepwise through the different functions of InfoQuest FP' 2D gel analysis application. In order to benefit from all the possibilities of the software, first time users are recommended to read this guide thoroughly.

•The Demo databases

In order to assist the user in setting up a database system, a small sample database is included with the InfoQuest FP 2D software. This sample database, which can be installed from the CD, contains four other examples of 2D gel TIFF files (**Furhigh.tif**, **Furlow.tif**, **Wthigh.tif**, **Wtlow.tif**) that will be used as examples in this guide. These files are obtained with kind permission from Dr. A.H.M. van Vliet¹. They represent a wild type *Campylobacter jejuni* strain exposed to low iron concentration (Wtlow) and high iron concentration (Wthigh),

and a *Fur* protein¹ mutant exposed to low iron concentration (Furlow) and high iron concentration (Furhigh). These gels will be further used for demonstration purposes throughout the 2D gel sections. A database **Demobase 2D**, containing these four gels fully analyzed, is installed with the software.

The plugin tool for 2D gel analysis as described in 6.1.4 also offers automated workflows for experiments containing multiple gels. Please refer to of this manual on how to install this plugin. Along with the plugin, a comprehensive manual containing tutorials with automated workflows is provided. You can use this tutorial manual to quickly and easily learn the basics of 2D gel analysis, including the automation of workflows. However, this manual remains a valuable tool to explore all the features of the *2D gel types* module in InfoQuest FP.

6.1.7 Creating a new database

As explained earlier (), InfoQuest FP databases are designed to store information in a structured way. New databases will be added to this structure, automatically creating the necessary files and folders to allow proper management and back-up of your data. We will create a new database for setting up some 2D gel experiments see also (1.5.2).

6.1.7.1 In the InfoQuest FP Startup screen, press the



button to enter the *New database* wizard.

6.1.7.2 Enter a name for the database, e.g. **Demo2D**, and press **<Next>**.

6.1.7.3 Press **<Next>** again without changing anything to the directory defaults.

6.1.7.4 You are now asked whether or not you want to create log files. If you enable InfoQuest FP to create log files, every change made to a database component (entry, experiment, etc.) is recorded to the log file with indication of the kind, the date, and the time of change (see 2.1.6).

6.1.7.5 Press **<Finish>** to complete the setup of the new database.

6.1.7.6 Before creating the final files, InfoQuest FP will need to know the type of database you like to prepare. Four options are available: *New connected database (automatically created)*, *New connected database (custom created)*, *Existing connected database*, or *Local database (single user only)*.

Details on the use of connected and local databases are given in .

6.1.7.7 Select **Local database (single user only)** and press **<Proceed>** in the *New database* dialog box. Press **<Yes>** in the next dialog box to confirm your selection and to quit the setup of the new database.


6.1.7.8 The *Plugin installation* toolbox appears. The available 2D gels plugin provides additional functionality for workflow automation (see 6.1.4 and 6.1.5). In this case, you can proceed without installing any plugins. For more information on the installation of plugins, see 1.5.3.

6.1.8 Defining a new 2D gel type

Similar as for the other experiment types in InfoQuest FP, it is possible to create different *2D gel types* within the same database. This option is very interesting to set up different kinds of 2D gel experiments within the same database. All options and parameters defined for a given kind of 2D gels will be stored within the 2D gel type. Other options and parameters may be stored within other 2D gel types, without having to overwrite carefully defined settings.

Within a specific 2D gel type, gels are normalized to match each other through a *Reference system*, similar as for 1-D fingerprints (). In a 2D gel type, a reference system is created by choosing a good quality gel with clearly resolved spots, and defining all spots, or a subset, as *reference spots* in the reference system. Other gels can then be aligned to the reference system, and thus to each other, by linking a number of corresponding spots to the reference spots in the reference system. Such linked spots are called *landmarks*. Based upon a number of landmarks defined by the user, the program can match all the non-landmark spots of the gel with the remaining reference spots of the reference system. This matching is done within certain tolerance boundaries, which can be specified by the user. In this way, corresponding spots on different gels are linked to each other by linking them to the same reference spots on the reference system.

Within the same 2D gel type, however, it is possible to define more than one reference system. This possibility is useful when creating, e.g., multiple gels from the same sample, composed of gels with e.g. different pI ranges.

6.1.8.1 To create a new 2D gel type, select **Experiments > Create new 2D gel type** in the *InfoQuest FP main* window. Alternatively, press the  button in the toolbar of the *Experiments* panel or right-click in the *Experiments* panel and select the option **Create new 2D gel type** from the floating menu that appears.

6.1.8.2 The *New 2D gel type* wizard prompts you to enter a name for the new type. Since we are going to work with the *Fur* experiments of *C. jejuni* (see 6.1.6), enter for instance **"Fur"** as name.

1. van Vliet, A.H.M., K.G. Wooldridge, and J.M. Ketley. 1998. *J. Bacteriol.* 180: 5291-5298.

1. The *Fur* protein controls the expression of iron-regulated proteins.

6.1.8.3 Press **<Next>** and select the correct optical density depth (OD) of the fingerprint data files. The default setting corresponds to the most common case, i.e. two-dimensional TIFF files with 8-bit OD depth (256 gray values).

6.1.8.4 After pressing **<Next>** again, the wizard asks whether the 2D gels have inverted densitometric values. This is the case when your image appears as white spots on a dark background.

Since InfoQuest FP 2D recognizes the darkness as the intensity of a spot, the wizard therefore allows you to invert the densitometric values.

6.1.8.5 Since the example consists of normally registered gels, i.e. dark spots on a white background, check **<No>**.

Furthermore, the wizard allows you to adjust the color of the background and the bands to match the reality. The red, green and blue components can be adjusted individually for both the background color and the band color. Usually, you will leave the colors unaltered. In case you like to mimic e.g. the blue of Coomassie Blue, you can move the Band color adjuster for Red (R) to left, for Blue (B) to right and for Green (G) to an intermediate position that produces the requested color.


6.1.8.6 In the next step, you are prompted to allow a **Background subtraction**. At this time, we leave the background subtraction disabled, by checking **<No>**. Later, we will see how to subtract background (6.1.10.22).

6.1.8.7 Press **<Finish>** to complete the creation of the new 2D gel type.

NOTE: You will be able to adjust all of these parameters later.

6.1.8.8 The *Experiments* panel now lists "Fur" as a 2D gel type experiment of the database **Demo2D**.

6.1.9 Importing 2D gel image files

6.1.9.1 Select **File > Add new experiment file** in the InfoQuest FP main menu, or press the  button in the toolbar of the *Files* panel.



6.1.9.2 Browse to the **[HOMEDIR]\Demobase 2D\Gel2d** folder. The **[HOMEDIR]** tag thereby points to the home directory as defined in the Startup screen (see 1.5.1). Select the **Furhigh.tif** file and press **<Open>**.

The gel **Furhigh** is now available in the *Files* panel. Repeat steps 6.1.9.1 to 6.1.9.2 for the image files **Furlow.tif**, **Wthigh.tif**, and **Wtlow.tif**.

For a local database, you can also copy the files **Furhigh.tif**, **Furlow.tif**, **Wthigh.tif**, and **Wtlow.tif** from the **[HOMEDIR]\Demobase 2D\Gel2d** folder directly

to the folder **[HOMEDIR]\Demo2D\Gel2d** using the Windows explorer.

The gels **Furhigh**, **Furlow**, **Wthigh**, and **Wtlow** are now available in the *Files* panel. The gels are marked with a red **N**, which means that they have not been edited or normalized yet.

*NOTE: Experiment files added to the Files panel can also be deleted by selecting the file and choosing **File > Delete experiment file** from the main menu or by pressing the  button from the Files panel. Deleted experiment files are struck through by a red line, but are not actually deleted until you exit the program. As long as you haven't closed the program, you can undo the deletion of the file by selecting **File > Delete experiment file** or pressing the  button again.*

6.1.10 Processing 2D gel images

The gel analysis workflow of InfoQuest FP 2D is arranged in a number of consecutive steps. These steps will guide you through the process of spot detection, OD calibration, normalization, quantification, and finally to the storage of all spot information in the database. The information is then available for the matching of multiple gels and for quantitative comparison or analyses with specific expression analysis tools.

Before processing work can begin, a TIFF file of a 2D gel experiment usually needs to be 'cleaned' before it can be used for spot detection and quantification. Basic image file editing and cleaning can be done in any image processing package. However, the treatment of a 2D gel image can involve some very specific routines such as background removal, spike removal, streak removal and filtering that are not readily available in traditional image processing software. InfoQuest FP 2D is equipped with a number of useful tools to perform various 'cleaning' activities on gel images and provides algorithms with user-adjustable parameters to deal with these important corrections. These algorithms include **Filtering** (median, Gaussian), **2D background subtraction** (rolling ball principle), **Streak removal** (horizontal and vertical), and **Spike removal**.

Following are the consecutive steps in a 2D gel processing using InfoQuest FP 2D:

1. **Spot detection** will find spots on the gel and quantify them by fitting a 2D Gaussian distribution to the spot. The result is a spot location (determined by the spot's mass centre), a spot size (average size in the X and the Y direction), a spot maximum, and a spot volume. Since overlapping spots are commonly found in 2D gels, InfoQuest FP 2D will detect these automatically and will propose the best possible separation. The spot search algorithm contains a number of parameters that can be varied and therefore it has been equipped with a very useful preview window.

Convenient editing tools allow for further manual correction, such as merging or splitting spots, adding, deleting or redrawing spots.

2. **Calibration** is an optional process that generates a *calibration curve* expressing the relationship between densitometric values on the scanned image file and real OD value. An image is usually calibrated by applying OD calibration strips delivered with the scanning device. These strips are processed along with each gel and will compensate for variation observed between different scans. After calibration, spot volumes are also shown as *relative volumes*. These relative volumes can be recalculated into absolute quantities in step 4 (Defining metrics, "Step 4: Defining metrics").
3. **Normalization** (gel alignment). The third step of the gel processing routine allows the mapping of a gel to a *reference system*. This reference system can be seen as an artificial gel to which the others are aligned. Normally, it is constructed on the basis of a real experiment, but it can be gradually extended (i.e. more spots are added) and modified as more gels are being analyzed and compared with that reference system. A number of tools are available that will allow spots on the gel to be matched with the corresponding spots on the reference system. Based on a number of easily recognizable homologous spots (*landmarks*), InfoQuest FP 2D will align the gel to the reference system and will allow *all* the spots of the gel to be linked to corresponding spots on the reference system, within a user- adjustable position tolerance.
4. **Assignment of metrics**. This step has two different purposes. At first, the mobility properties of each spot will be calculated in both dimensions using a regression. To establish the regression, marker or reference proteins or other easily recognizable physical points (incision, dots, scale indications, colored molecular weight markers for blots, etc.) can be used, which correlate with positions of known molecular weight or pI values. InfoQuest FP 2D can use linear or exponential fitting algorithms of first till fifth degree, with or without logarithmic dependence. The 2-dimensional mobility grid thus appearing can be rotated by the user to correct for artifacts like non-horizontal shots or mobility inclination. The assigned mobility properties (metrics) will assist in comparing a specific spot (with an unknown protein) to known proteins with known molecular weights and pI values. Secondly, it is possible to enter concentration values for spots containing fixed quantities of protein (expressed in ng or μg). The values of these known concentration marker spots are used to establish a regression that calculates spot concentrations for all spots, based on their calibrated volumes. In this way, any calibrated volume in the gel can be read in 'amount of protein', e.g. expressed in μg or ng.
5. **Database construction**. Identified spots can be provided with descriptive information and stored in the database. The descriptive information can be

obtained from existing databases (e.g. 2DPage of SwissProt) by using the accession number or spot reference number. Alternatively, the user can build an own protein reference database by entering specific information fields. A total of 8 fields of unlimited length can be selected. Within the frame of the InfoQuest FP Platform it is possible to link the 2D gel information to other experimental data in the database and to perform multi-experimental comparisons and data mining (e.g. comparison with micro-array data). The protein spot query tool allows the selection of specific proteins from many 2D gels belonging to the same or different 2D gel types.

6. **Matching 2D gel spots**. The 2D gel spot information fields are accessible for searches in order to retrieve subsets of spots. By using the advanced query tool of the InfoQuest FP 2D software on a selection of several gels, it is possible to create subsets of spots that can be analyzed by a variety of comparison tools. The result of a query, covering several gels, can be displayed as a histogram and can be analyzed statistically. In combination with the GeneMaths XT software, it is possible to evaluate protein expression profiles, study time course relationships, etc. on the selected set of proteins.

We will now start analyzing the first 2D gel within the newly created database **Demo2D**.

- 6.1.10.1 Click on **Wtlow** in the *Files* panel and press



from the *Files* panel toolbar or select **File > Open experiment file (entries)** from the *InfoQuest FP main* window. Alternatively, just double-click on the file name.

- 6.1.10.2 Since the gel is new (not processed), InfoQuest FP 2D doesn't know what 2D gel type it belongs to. Therefore, a list box is first shown, listing all available 2D gel types. Select the 2D gel type '**Fur**' and press **<OK>**. By clicking **<Create New>** you can create a new 2D gel type that fits the gel file.

NOTE: The same gel cannot be used in two different 2D gel types.



- 6.1.10.3 The gel file is loaded. Depending on the size of the image, this may take some time. The *2D gel processing* window appears (Figure 6-1), showing the image of the gel.

The *2D gel processing* window consists of two panels: the *Image* panel, displaying the data gel image, and the dockable *Entries* panel, displaying the database entry to which the 2D gel file is linked to (see 6.1.15 for more information on this panel).

As explained above, processing a 2D gel is a multistep process. The current step of 2D gel processing is shown as tabs in the bottom part of the *Image* panel. Initially, the *Spot detection* tab is active.




Figure 6-1. The 2D gel processing window. Step 1: Spot detection.

6.1.10.4 With **Edit > Zoom in (+)** or **Edit > Zoom out (-)** the 2D gel image can be sized to fit the screen. Throughout the gel analysis procedure this tool can also be activated by the respective buttons  and  or the + and - keys, respectively.

*NOTE: Manual tools can be used with greater precision on an **enlarged image**. The **zoom factor** can be read from the bar, which also lists the file type, image size and OD depth. E.g., **TIFF: 1110x804x8 (x1.00)** means 1110 pixels horizontally by 804 pixels vertically and an OD depth of 8 bits per pixel. The **zoom factor** is 1.00.*

A powerful tool to edit the appearance of the image is the *Gel tone curve* editor. While the *Image brightness and contrast* settings act only at the screen (monitor) level, the *Gel tone curve* editor acts at the original TIFF file information. Although the original tiff file is never physically changed, the settings that apply to it will be modified by its proper *Tone curve*, which is saved along with each particular gel. In case a 2D gel image was scanned in 16-bit mode, the tone curve settings are applied to the full 16-bit (65,536) gray scale information, allowing much more information to be revealed in areas with lower contrast.

6.1.10.5 In the 2D gel processing window, select **Edit > Edit tone curve** or press . The *Gel tone curve* editor appears as in Figure 6-2. The upper panel is a distribution plot of the densitometric values in the TIFF file over the available range. The right two windows represent a part of the 2D image **Before correction (upper)** and **After correction (lower)**.

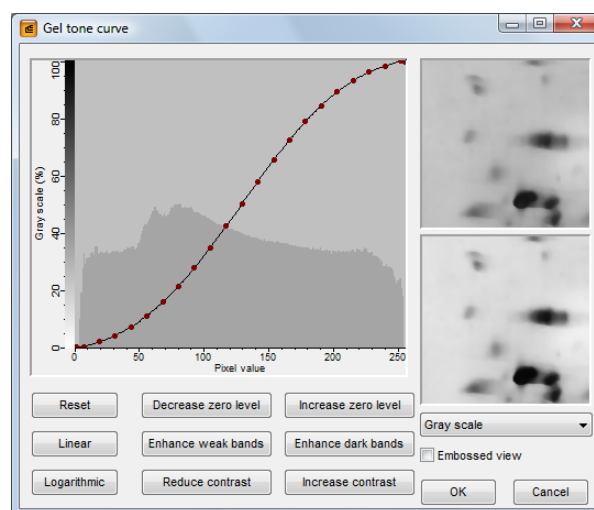


Figure 6-2. The *Gel tone curve* editor.

6.1.10.6 You can scroll through the preview images by left-clicking and moving the mouse while keeping the mouse button pressed.

6.1.10.7 Select a part of the preview images which contains both very weak and dark bands.

On the left, there are two buttons **<Linear>** and **<Logarithmic>**. Both functions introduce a number of distortion points on the tone curve, and reposition the tone curve so that it begins at the grayscale level where the first densitometric values are found, and ends at its maximum where the darkest densitometric values are found. This is a simple optimization function that rescales the used grayscale interval optimally within the available display range. The difference between linear and logarithmic is whether a linear or a logarithmic curve is used.

6.1.10.8 In case of 8-bit gels, a linear curve is the best starting point, so press **<Linear>**. The interval is now optimized between minimum and maximum available values, and the preview *After correction* looks a little bit brighter.

There are six other buttons that are more or less self-explanatory: **<Decrease zero level>** and **<Increase zero level>** are to decrease and increase the starting point of the curve, respectively.

<Enhance weak bands> and **<Enhance dark bands>** are also complementary to each other, the first making the curve more logarithmic so that more contrast is revealed in the left part of the curve (bright area), and the second making the curve more exponential so that more contrast is revealed in the right part of the curve (dark area).

<Reduce contrast> and **<Increase contrast>** make the curve more sigmoid so that the total contrast of the image is reduced or enhanced, respectively.


6.1.10.9 For the image loaded, pressing three times **<Enhance weak bands>** provides a more contrastive picture.

In standard mode, the gel is displayed as a continuation of gray levels. The *Rainbow palettes* and *Contour palettes*, which exist of multiple color transitions, can reveal more visual information in areas of poor contrast (weak and oversaturated areas). In InfoQuest FP 2D, 6 different palettes have been pre-defined besides the gray level representation. In the Contour palettes, each range of the five colors is bordered by a dark transition, which is useful to delineate the *contours* on the 2D gel image. *Contour Palette (I)* has 5 and *Contour Palette (II)* 9 different discrete color ranges while *Contour Palette (III)* has five colors characterized by a discontinuous transition. The use of the different palette views is useful for the evaluation of slight intensity gradients that are invisible in grayscale, such as e.g. the efficacy of the

background removal settings used, judgement of double/single spots.

6.1.10.10 If you press **<OK>**, the tone curve is saved along with the gel.

Another interesting viewing mode is the *embossed view*.

6.1.10.11 Select *Edit > Edit tone curve* or press  again.

6.1.10.12 Click the **<Embossed view>** check box and press **<OK>**.

The use of the embossed option (Figure 6-3) adds a third dimension to the display. The gray levels are transformed to a shaded 3-D shape that enhances the distinction between higher and lower intensities, for example to separate spots in high-intensity areas which look uniformly black in normal grayscale mode.

NOTES:


(1) *Embossed view* cannot be shown in the preview panel in the Gel tone curve editor.

(2) *The embossed view effect is largely lost when a strong zoom factor is used (>x4.00).*



Figure 6-3. The Embossed view option from the *Gel tone curve editor*.

6.1.10.13 Call the *Gel tone curve* editor again, uncheck the embossed view, and press **<OK>**.

6.1.10.14 To save the work done at any stage of the process, you can select *File > Save*, press the **<F2>** key or the  button. It is recommended to save the gel at regular intervals.

A last option that will improve the interpretation and visualization of the 2D gels is the 3-D viewing mode.

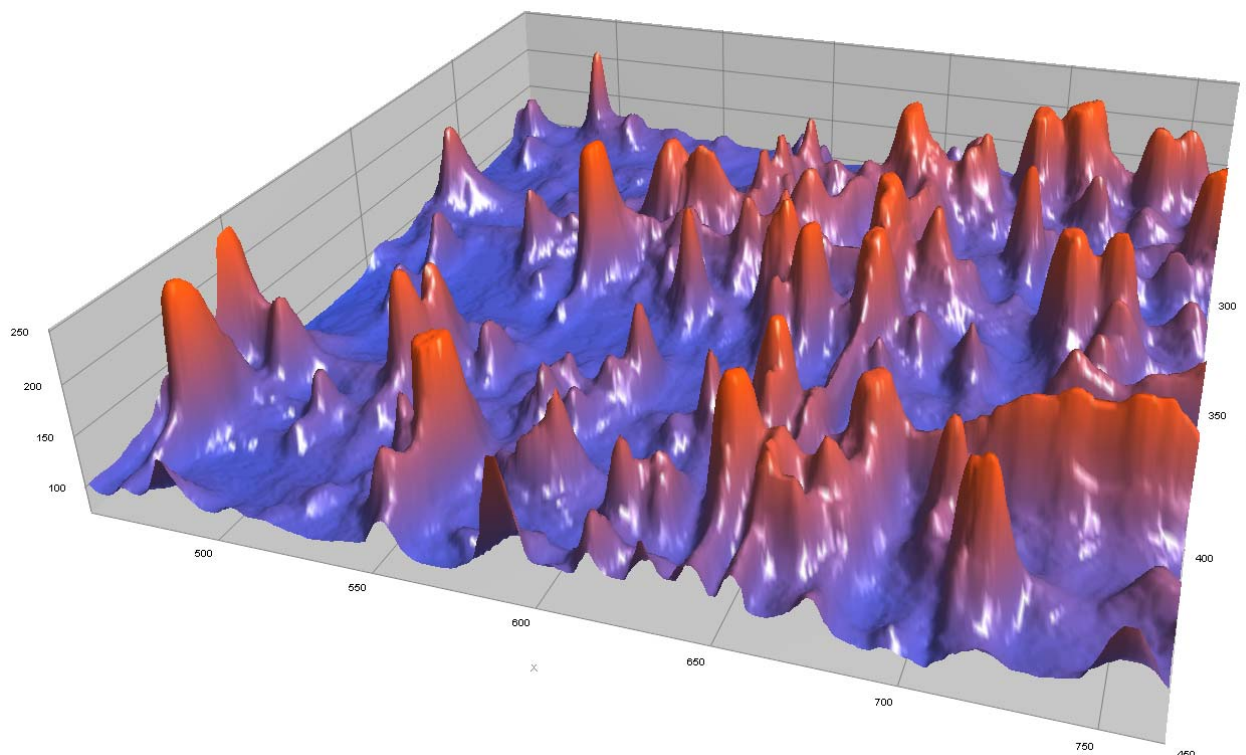




Figure 6-4. 3-D view of a zoomed area of a 2D gel.

6.1.10.15 Zoom in on an area of the gel with many overlapping spots of high intensity, using the  button or *Edit > Zoom in (+)*.

6.1.10.16 Display the 3-D view window using the 3-D view button  or by *File > View 3D image*.

This option opens a new window that contains a scalable three dimensional view of the image. The Z-axis is used to display the pixel intensity of each individual point in the gel. The 3-D view is particularly suited to evaluate and judge individual spots for possible overlap, presence of spikes or visualization of noise (Figure 6-4). Therefore, the view can be used to judge the effect of background removal, spike removal or streak removal (see below 6.1.10.22). To that extent it is possible to keep several 3-D representation windows open at the same time, allowing a side by side comparison for the study of the effect of specific actions on the 2D gel.

6.1.10.17 Select *View > Show spot outlines* to plot the spot contours on the 3-D image (see 6.1.11).


6.1.10.18 By using the **Left**, **Right**, **Up** and **Down** arrows keys on the keyboard, the position of the image can be manipulated in all directions. The image can also be rotated horizontally and vertically by dragging the image left/right or up/down using the mouse.

6.1.10.19 Use the **PgUp** and **PgDn** keys to zoom in or out of the image.

6.1.10.20 The **Insert** and **Delete** keys can be used to higher or lower the peaks, by resizing of the Z-axis.

NOTE: In contrast to the Embossed view, the 3-D view is best used on an image with strong zoom. The zoomed area will selectively be displayed and can be viewed from all sides.

6.1.10.21 Close the 3-D view window with *File > Exit*.

6.1.10.22 Edit the general settings of the 2D gel processing window with *Edit > Settings* or .

In this window, the *Image* and *Metrics* settings can be defined (Figure 6-5). We will discuss the *Metrics* settings later (see paragraph 6.1.14).

- **Image coloring.** With *Inverted values*, gels with bright spots on a dark background can be inverted to dark spots on a bright background. The *OD range* of the gel can be specified in number of grayscale levels (256 = 8 bit; 1024 = 10 bit; 4096 = 12 bit; 65536 = 16 bit). As explained in 6.1.8, the initial settings related to the image color display (on a monitor or screen), and defined during the setting-up of the 2D gel type, can still be changed. The RGB (red-green-blue) contributions can be changed for the *Background color* and the *Foreground color*. Using this tool it is possible to mimic e.g. Coomassie blue or silver stain.
- **Background subtraction** is based on the “rolling ball” principle, i.e. a ball of a certain size is rolled against

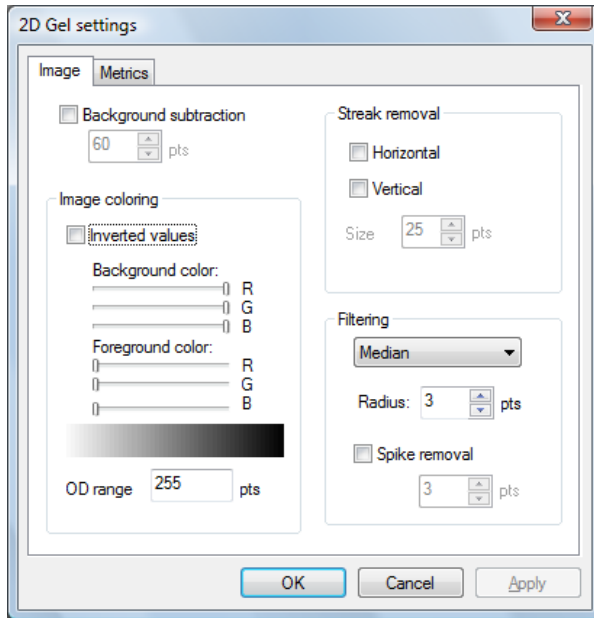


Figure 6-5. 2D gel settings dialog box.

the inner side of a 3-D surface of the gel image. Depths the ball could enter are removed from the image. The size of the ball, in pixels, can be entered. The larger the size of the ball, the less background will be subtracted, but the faster the calculations will be. Background removal may be very effective in removing smear that makes spot quantification more difficult. Removing too much background holds a potential danger of removing or excavating large protein spots from the gel.

- **Spike removal** is a filtering technique with a similar mechanism as the rolling ball. A very small ball size is taken, so that the ball can enter into all regular protein spots, but not into spikes and noise caused by dust, scratches, etc. Those depths the ball could not enter into are removed. The size of the ball can be entered in pixels. The size should be chosen very small, usually less than 4 pixels.
- Two types of **Filtering** have been implemented in InfoQuest FP 2D to smoothen the image: the Median and the Gaussian filtering. **Median filter** is a method which reduces irregularities that constitute less than 50% of the number of values to average. The Median filter is therefore very efficient in removing noise and isolated spikes. The **Gaussian filter** can be used to filter out more continuous noise. Gaussian filtering will remove the continuous noise rather than the accidental noise spikes. When using a Gaussian filtering on the latter, the spike intensity may be reduced but not eliminated. The **Radius** of the filter can be entered in pixels.
- **Streak removal** is a similar mechanism as the rolling ball, but an ellipse is used instead, in order to separate streaks from spots. The streak removal algorithm can look in a **Horizontal** and **Vertical** direction for the presence of continued smear of protein. The **Static**

(length of the zone) to be considered as smear can be entered in pixels.

NOTES:

(1) All the above filtering algorithms will not change the TIFF files permanently but will have an influence on the 2D gel representation and on the spot detection and quantification algorithms. Since the original TIFF files will not be changed, the settings applied to a specific gel can be modified at any step at any time.

(2) Since these settings will have a considerable impact on the spot detection and quantification procedures, we recommend to use these options with care. The spike removal and streak removal algorithms in particular should be handled with consciousness of the effect of the algorithm on all spots. These algorithms inevitably cause some distortion on the protein spots as well. The smaller the level of the spike / streak removal, the less the distortion.

6.1.10.23 For the gel **Wtlow** select median filtering with averaging 3 (6.1.10.22), background subtraction of 30 pixels, horizontal streak removal of 25 pixels, and spike removal of 3 pixels.

6.1.10.24 Press <OK> to save the settings.

NOTES:

(1) It may be useful to redefine the Tone curve after applying the image enhancement settings.


(2) The above settings can be stored for global use by the function **Edit > Save as default settings**. The default settings can be reloaded at any time by **Edit > Load default settings**.

6.1.10.25 With the command **File > Print image** or **File > Copy image to clipboard**, you can print the unprocessed 2D gel, or copy it to the clipboard.

6.1.11 Step 1: Spot detection

Spot detection is an important feature in the creation of a protein database. It needs some experience to include / exclude specific spots from a variety of gels in a consistent way.

Besides the possibility to pre-optimize an image for better spot detection (see 6.1.10), InfoQuest FP 2D provides a preview-based automatic assignment of spots. The normal procedure is to allow the software to assign spots automatically, after adjusting the parameters using the preview window, and further inspect the assigned spots and correct manually where necessary.

6.1.11.1 By selecting **Spots > Automatic search** or by pressing the  button, the *Automatic spot search* dialog box is opened (Figure 6-6).

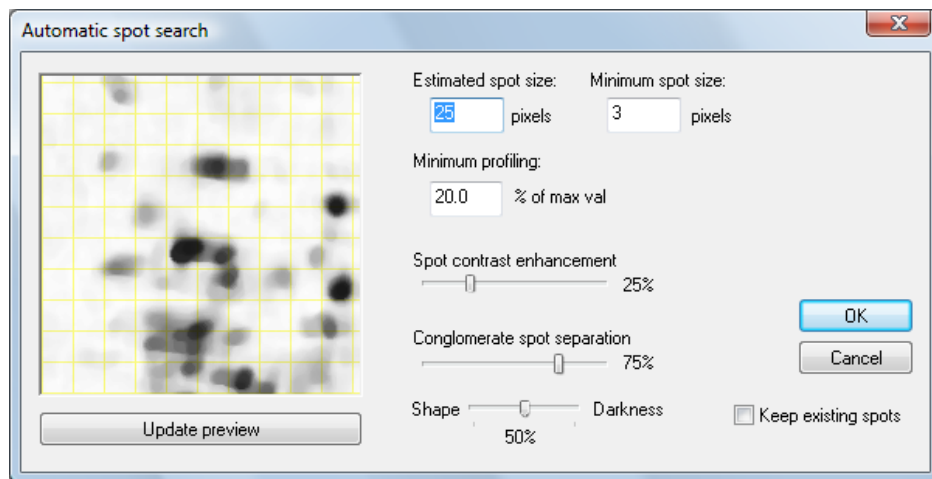


Figure 6-6. Automatic spot search dialog box.

The dialog box allows you to see the influence of any parameter you change on a selected area of the underlying gel image. The area of the gel can be changed by dragging the window with the mouse to any desired position of the gel. The yellow positioning frame is the unit size which covers 10 pixels.

6.1.11.2 By choosing **<Update preview>**, all spots in the selected zone will be detected, using the current conditions.

The spots detected are indicated by red circular indications in the spot preview window (Figure 6-6). After changing the target position of the 2D gel you will have to update the detection by clicking the **<Update preview>** button again. Amongst the spot detection parameters that can be adjusted, are the *Estimated spot size*, the *Minimum spot size*, the *Minimum profiling*, the *Spot contrast enhancement*, the *Conglomerate spot separation*, and the *Shape/Darkness* sensitivity.

The *Estimated spot size* assists the search algorithm in finding spots of approximately the size specified by the user. Depending on the resolution and the size of the gel, there may be considerable differences in the pixel size of a single spot. By estimating the average spot size (or diameter of a virtual circular spot) the software can start to screen the 2D image for individual spots. The default value is set at **25**.

The *Minimum spot size* is an additional help for the algorithm to discriminate spikes from real spots and to optimize the search algorithm. Spots which have a size below the indicated minimum spot size will not be considered by the algorithm. The default value is set at **3**.

The *Minimum profiling* is the elevation of the spot compared to the highest 2% intensity found on the gel. The higher the value is set, the darker a spot should be before it will be found. The default value is 30.

Spot contrast enhancement is an algorithm to reduce the spot surface relative to the intensity of the spot. This

means that dark spots will be clipped at higher gray levels than weak spots. The algorithm also has an influence on the final number of spots found: when small spots are clipped at higher or lower grays, they may fall within or without the minimum spot size. Since the algorithm is applied before the *Conglomerate spot separation*, a large *Spot contrast enhancement* will cause more small spots to be found around high intensity spots or areas. The slider bar, in combination with the preview window, will allow you to quickly evaluate the most suitable *Spot contrast enhancement* setting for each gel.

NOTE: *Spot contrast enhancement* settings may vary according to the gel image processing parameters that have been used. Changing the background subtraction level may have an impact on the optimal *Spot contrast enhancement* settings.

Increasing the *Conglomerate spot separation* factor will force the algorithm to pay more attention to the detection of multiple spots in a core that initially has been recognized as a single spot. In Figure 6-4, the frequent overlap of not completely separated protein spots is illustrated. InfoQuest FP 2D can be forced to explore each spot for the presence of subtops which could be the core of a non completely resolved protein peak (increase conglomerate spot separation) or can be instructed to consider any continued elevation clearly separated from the background as a single spot (decrease conglomerate spot separation).


As an additional parameter in this process you can indicate whether the decision for splitting conglomerate spots will be based on an evaluation of the basic *Shape* (*constriction* sensitivity), or on the *Darkness* of the conglomerate (*depression* sensitivity). In the latter case, the presence in the spot surface of multiple cores or subtops, separated by a valley, will thrive the splitting process. In case shape is selected as a major criterion for splitting, irregularities in the contour (e.g. a spot with a typical 8 -shape) will trigger the splitting process. The slider allows you to modify the importance of either criteria by changing the ratio Shape / Darkness.

6.1.11.3 Choose the following parameter settings for the gel **Wtlow** and press <OK>:

- **Estimated spot size:** 25 pixels
- **Minimum spot size:** 5 pixels
- **Minimum profiling:** 15%
- **Spot contrast enhancement:** 70%
- **Conglomerate spot separation:** 80%
- **Shape/darkness:** center

Assigned spots are now contoured by a green border-line, and semi-transparently colored in green. Selected spots are colored in red.

The semi-transparent overlay can be toggled on or off using the function **Edit > Show filled spots**. To examine weak spots, you may prefer to show the contours only.

6.1.11.4 With **Edit > Spot info** or  you can display a small pop-up window that shows information about the selected spot (Figure 6-7).

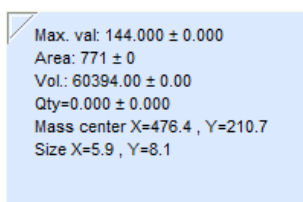



Figure 6-7. Spot information pop-up window.


6.1.11.5 The pop-up window can be moved by dragging the mouse anywhere inside the window, and always remains in front of the 2D gel editor.


6.1.11.6 You can close the spot information window by clicking in the upper left triangular button.

*NOTE: It is also possible to display a label for each spot in the 2D gel processing window, using **Edit > Label spots with**, which offers the choice between a number of information fields that can be assigned to a spot (see 6.2.1). However, such a label can only be displayed if the gel is fully processed according to the reference system. This option can be useful too, if already processed gels are re-edited.*


After the automatic spot search, some spots may still remain undetected, while others may be indicating small background elevations. Also, some conglomerates may have been found which are clearly composed of multiple spots that have not been identified separately with the current settings. A number of manual spot editing tools are available to add, remove, separate, merge, and redraw spots.

6.1.11.7 **Add spot tool.** Click the button  or press SHIFT+F2 to change your cursor into a spot adding tool. It is sufficient to click in the center of an unassigned spot to assign an additional spot. The software will find the correct shape of the newly created spot automatically.



6.1.11.8 At any time you can return to the pointer status of your cursor by clicking  again, or pressing SHIFT+F1.


6.1.11.9 **Remove spot tool.** Click on the button  or press SHIFT+F3, to change the cursor into the spot removing tool. Simply click on an assigned spot to delete it.


6.1.11.10 Selected spots can also be deleted using the DEL key or with **Spots > Delete selected spots**.

6.1.11.11 Groups of spots can be selected at once by dragging the mouse in pointer mode () over the area to select. These spots can be deleted at once by pressing the DEL key.

6.1.11.12 With **Spots > Select all spots** or CTRL+A, all spots on the gel can be selected at once.

*NOTE: The 2D gel processing window has a **multi-level undo function** that will allow you to undo a large number of previously performed manipulations. This undo option therefore enables you to evaluate safely a number of processing steps on your 2D gel. In case you are not satisfied with the result of your last modifications you can get back to the status that had your last approval by consecutive use of the 'undo' function. Press the undo button  or the redo button  or use the commands **Edit > Undo last action** (CTRL + Z) or **Edit > Redo last action** (CTRL + Y) from the menu. The number of steps you can undo, however, is not unlimited. It is therefore advised to save your approved work regularly. Caution: saving your data will erase the 'undo' memory.*


6.1.11.13 **Drawing tool (add pixels).** Press  or SHIFT+F4 to turn the mouse pointer tool into a drawing pencil.

The tool is always linked to a specific pen size as displayed by the pen size tool buttons (). The selected pen size is highlighted.

6.1.11.14 You can use the pencil to mark new spots on the 2D gel or to extend the contours of an existing spot.

NOTE: It is advised to use the drawing pencil in combination with the zoom buttons (+ and - keys)

(6.1.10.4). Using the zoom function, the program will automatically zoom on the selected spot.

6.1.11.15 **Drawing tool (remove pixels)**. Press  or SHIFT+F5 to turn the mouse pointer tool into a pixel removing pencil.

6.1.11.16 You can use the pencil to delete (parts of) selected spots or to separate larger spots into two or more smaller spots.


When used in the appropriate zoom mode and with **small pen size**, the tool allows you to split conglomerate spots following a precise user-defined trace. When used with a **large pen size**, the tool can be used to erase complete spots. When used with an **intermediate pen size** it can be used to delete parts of the spots that should not be considered for quantification.

6.1.11.17 **Split selected spot**. Select a conglomerate spot, which you may want to split up into two spots.

6.1.11.18 Press the button , <F7> or select **Spots > Split selected spot**.

Using this function, you can force the program to calculate the most probable trace to split a spot. When the tool refuses to split a spot it means that no well delineated trace has been discovered that could be used for splitting the spot. Therefore, the software considers it as a single spot that cannot be further divided. If you still like to split such a spot, you can use the tool described in 6.1.11.15.

6.1.11.19 **Merge selected spots**. Select two spots which are more likely to belong to one protein (e.g. the program sometimes identifies smear as a second spot). Two spots can be selected together by holding down the CTRL key while selecting the second.

6.1.11.20 With the button , <F8> or with **Spots > Merge selected spots** you can merge two or more spots that have been selected in advance.

Four more functions are available to edit the selected spot(s):

- **Spots > Fill internal spot holes (CTRL+F)**: This feature is useful if you have drawn spot contours with the pen (6.1.11.13). You can draw the contour using a small pen size without having to fill the spot up. As soon as the contour is closed, CTRL+F will fill up the holes for the selected spot(s) (see Figure 6-8).
- **Spots > Smooth selected spots (CTRL+M)**: This feature is similar to the previous in that it fills up internal holes. In addition, it smooths the spot contours so that drawing imperfections are corrected automatically (see Figure 6-8).
- **Spots > Grow selected spots (CTRL+PgUp)**:

- **Spots > Shrink selected spots (CTRL+PgDn)**:

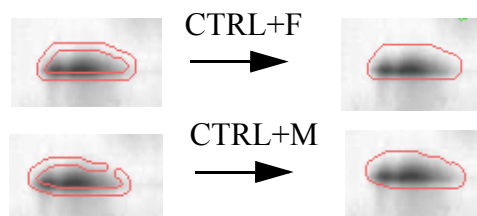




Figure 6-8. The effects of the Fill and Smooth functions.

6.1.11.21 Once you are satisfied with the assigned spots, you can save the work done by selecting **File > Save**, pressing the <F2> key or the  button. It is advisable to save the gel at regular times.

6.1.11.22 Press the next step button () to move to the next step, the **Calibration** step.

6.1.12 Step 2: Calibration

Calibration is an essential step when different gels need to be compared quantitatively. Calibration will improve the inter-gel comparability by correcting for intensity differences between scanned gels due to differences in digitizing of the gel. During the process of calibration, the relation between the original OD (or counts of radiation) and the intensities on the TIFF file image can be defined. Through calibration, each gray level on an image TIFF file can be assigned a calibrated value on the basis of a non-linear regression curve. Calibration can therefore be used to compensate for non-linearity of scanners in the high OD range, or a non-linear response of radiographic film to exposing radiation. To that purpose it is possible to link specific known OD (or radiation) levels to an area of raw pixel values of the TIFF file image.

The most obvious way to define this calibration curve is in combination with a scanner or CCD camera, by the use of calibration strips that are applied on, or next to, the gel. These strips represent well-known physical properties, e.g. OD values. The calibration zones in the strips can be defined by the user, for which the physical value can be entered. After calculating the non-linear calibration curve, every pixel on the 2D gel image can be translated into a new calibrated value with some physical property. For quantification purposes, it is recommended to compare calibrated gels only with calibrated gels.

IMPORTANT NOTE: Calibration is always performed on the raw, unprocessed image file. In order to have a realistic view on the calibration strips on the scanned image, you may need to switch off the background

subtraction, as well as other filters such as streak removal.

Since the present gel has no calibration strips applied, we will perform a fictitious calibration using different intensity areas on the raw gel image.

6.1.12.1 First, switch off the background subtraction (*Edit > Settings* and uncheck *Background subtraction*).

6.1.12.2 Calibration rectangles can be defined using the *Add new calibration rectangle tool* in the toolbar




6.1.12.3 With the calibration rectangle tool cursor selected, draw a small rectangle in the brightest area of the gel (bottom left area).

6.1.12.4 A dialog box pops up, prompting to enter a calibration value for the defined rectangle. Enter zero (0).

6.1.12.5 Next, select a very dark spot, e.g. the lowest spot in the left molecular weight lane, and draw a small rectangle in the center of that spot. Enter 3.0.

6.1.12.6 Lastly, draw a rectangle in the upper center background part of the image, and enter 0.5.

The calibration rectangles are indicated as blue rectangles, with a node in the upper left corner and in the bottom right corner. They can still be modified, after selecting the pointer tool ().

NOTE: The Undo function does not work in this step of 2D gel processing.


6.1.12.7 If you click inside a rectangle, it becomes selected (pink color).

6.1.12.8 To move a calibration rectangle, drag it to a new place using the upper left node.

6.1.12.9 To resize a calibration rectangle, drag the bottom right node to obtain the desired shape.

6.1.12.10 To delete a calibration rectangle, select it and press the DEL key.

6.1.12.11 If you want to change the calibration value of a rectangle, double-click on the rectangle or select it and use the option *Calibration > Change calibration value*.

6.1.12.12 When all values have been entered you can calculate the calibration curve by clicking the *<Edit calibration curve button>* () or by selecting *Calibration > Image calibration*.

This will bring the *Image calibration* window on the screen as displayed in Figure 6-9.

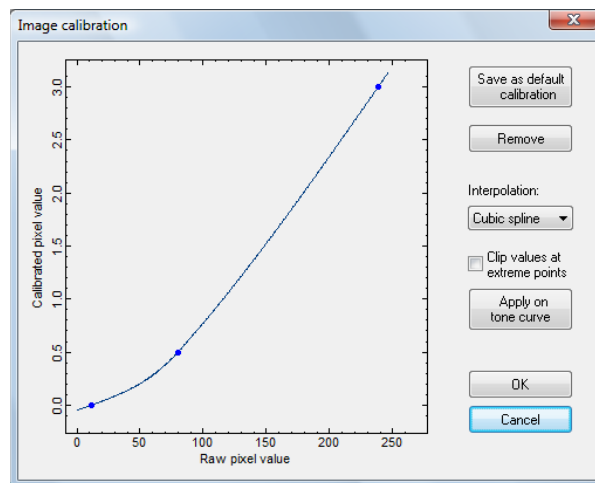



Figure 6-9. The *Image calibration* window after the definition of 3 calibration values.

6.1.12.13 To calculate the calibration curve, click the *'Interpolation'* checkbox and select one of the five fits: *Cubic spline* and *Polynomial (1) to (5)* (the number is the degree of the polynomial fit).

6.1.12.14 Select the *Cubic spline* fit for this calibration curve.


Other applications of the calibration curve include:

- When checking the option *Clip values at extreme points* in the *Image calibration* window (Figure 6-9), all values higher or lower than the respective highest and lowest calibration value entered will be clipped to these respective values. This can be useful if the dynamic range of the film or the scan is not reliable in these higher or lower ranges.
- In case not all gels that need to be compared have a calibration strip, you can save the present curve as the standard calibration curve. This curve can then be used for gels without calibration strip, still enabling reliable quantitative comparison. The gels should be processed and digitized in a similar way. In addition, a comparable amount of total protein should have been loaded. The calibration curve can be saved by selecting the option *Save as default calibration* in the *Image calibration* window (Figure 6-9). This curve will be automatically loaded when the next 2D image for that experiment type is analyzed. You can verify at any moment the active calibration curve by pressing the *<Edit calibration curve button>* () or by selecting *Calibration > Image calibration*.
- When you press *Remove* in the *Image calibration* window, the presently calculated curve is removed. Calibration rectangles will be preserved and a new curve can be calculated at any time.
- When clicking the button *<Apply to tone curve>*, the present tone curve settings (see the tone editor

description in 6.1.10.5) will be overwritten and will be replaced by the calibration curve you have defined. The advantage of applying the calibration curve to the tone curve is (1) that every gel will be displayed in a similar way, improving the immediate visual evaluation of quantitative differences on the screen, and (2) that OD levels where the scanner provides poor discrimination can be linearized to offer a better visual OD depth on the screen.

6.1.12.15 Press **<OK>** in the *Image calibration* window to save the calibration curve along with the gel.

After calibration, for any point of the 2D gel, the status bar will show you the gel information as discussed in section 6.1.10.4, as well as the **raw** value before background subtraction/value after background subtraction as well as the **calibrated** value.

6.1.12.16 This concludes part 2, Calibration. Press the next step button () to move to the next step, the **Normalization** step.

6.1.13 Step 3: Normalization

Normalization of 2D gels is necessary to locate homologous spots on different gels. In InfoQuest FP 2D, normalization makes use of a *reference system*. The reference system is a collection of *reference spots* with their coordinates, mass centers and X-Y sizes. These reference spots have a dual function: (1) They can be linked to homologous spots on other gels, which are then called *landmarks*. Once a number of landmarks have been defined for a gel, the program can map the gel on the reference system. Mapping of a gel is a process of distorting the 2D image so that all linked positions on the gel and the reference system fit each other. Every pixel on the gel is recalculated based upon the relative distance and the magnitude of the known displacement vectors. The process of linking spots to reference spots and recalculating the image is called *normalization*. (2) Once a gel is normalized, the remaining (non-landmark) spots can be matched with the corresponding reference spots. When a spot on one gel is linked to the same reference spot as a spot on another gel, these spots are considered the same protein. This is the key to compare different gels with one another.

Usually, the reference system is initially built from a 2D image of a representative gel by defining easily recognizable protein spots as *reference spots* on the reference system. As more gels will be matched with this reference system, you can add additional reference spots to the reference system. Adding new spots from additional gels will allow new spots to be added to the database and will have no influence on previously normalized gels. A reference system is shown as a synthetic gel with spots created from the mass center, height, X-size and Y-size of the spots that were added to the reference system.

NOTE: Normalization of the image, i.e., recalculating the image to fit the reference system, is only performed for easier visual evaluation. Internally, all quantification is done on the non-distorted image.

The sequence of steps in the normalization procedure is schematically summarized as follows:

1. Defining new reference system

②

2. Adding reference spots to the reference system

②

3. Defining landmarks (linking spots to reference spots to align gel)


②

4. Linking all gel spots to corresponding reference spots

Steps 1 and 2 are done once initially, whereas steps 3 and 4 are done for each new gel. However, new reference spots may be added to the reference system as new gels are analyzed (step 2).

Within the same 2D gel type, it is possible to create different reference systems. This makes it possible to merge gels with different pH ranges of the same sample into one multiple experiment. Within a 2D gel type, one reference system is the active reference system, to which a new gel is assigned by default.

• **Creating a reference system**

When the normalization step is reached for the first time in the current 2D gel type (with ), there is no reference system available. Therefore the software will automatically open the *Assign to reference system* dialog box (Figure 6-10). This dialog box shows all available reference systems. Since there is no reference system available yet the option **<Add new>** will be needed.

6.1.13.1 We add a new reference system by clicking **<Add new>**.

6.1.13.2 In the dialog box '*Add new reference system*' type the name of the new reference system: **Fur** and click **<OK>**.

The program now asks "**Do you want to add all spots of the current gel to this reference system?**". If you answer **<Yes>**, all the spots of the current gel will be defined as reference spots in the new reference system. This can save you the work of adding the spots manually afterwards (see 6.1.13.8 to 6.1.13.11). If you plan to add only a selected number of spots to the reference system, choose **<No>**.

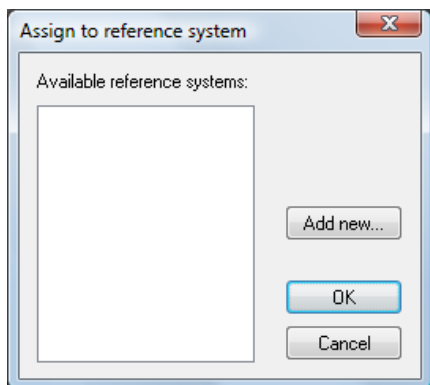


Figure 6-10. The *Assign to reference system* dialog box.

NOTE: If the program is allowed to add all the spots of the current gel automatically to the reference system, the spots of the current gel are automatically landmarked as well (see page 350).

6.1.13.3 Answer **<No>** to the question to add all spots of the current gel.

6.1.13.4 Now the reference system appears in the *Assign to reference system* dialog box and it becomes automatically selected.

6.1.13.5 Close the dialog box by clicking **<OK>**.

In the *Normalization* step, an additional *Reference* panel becomes available in the *2D gel processing* window. The

Reference panel (left in default configuration) will display the reference system, and the *Image* panel (right in default configuration) shows the current gel image. Since the reference system is empty as yet, nothing is shown in the *Reference* panel, except for the contours of the spots from the current gel (Figure 6-11). When you select a spot in the *Image* panel (click on the spot), its contour in the *Reference* panel also becomes highlighted (red) (see Figure 6-11).

• **Defining reference spots**

Setting up a reference system for the first time will require a number of *reference spots* to be defined. Reference spots are protein spots that will allow matching of future gels with each other. Since gel **Wtlow** is the first gel we are analyzing, we will have to define the reference spots from this gel, and thus, gel **Wtlow** will automatically become the *reference gel*. It is obvious that the reference spots present in the reference system should be well spread and covering all areas of the gel as uniformly as possible. Usually, there is no drawback in adding all spots of the gel to the reference system: reference spots that are not linked to spots on the gel can be left untouched.

6.1.13.6 Select a spot on the gel **Wtlow**. The spot is highlighted in red, and the contour of the spot is highlighted in the *Reference* panel as well.

Before we can add reference spots to the reference system, we will need to view the gel in *normalized mode*. The logic behind this step is that the reference system is

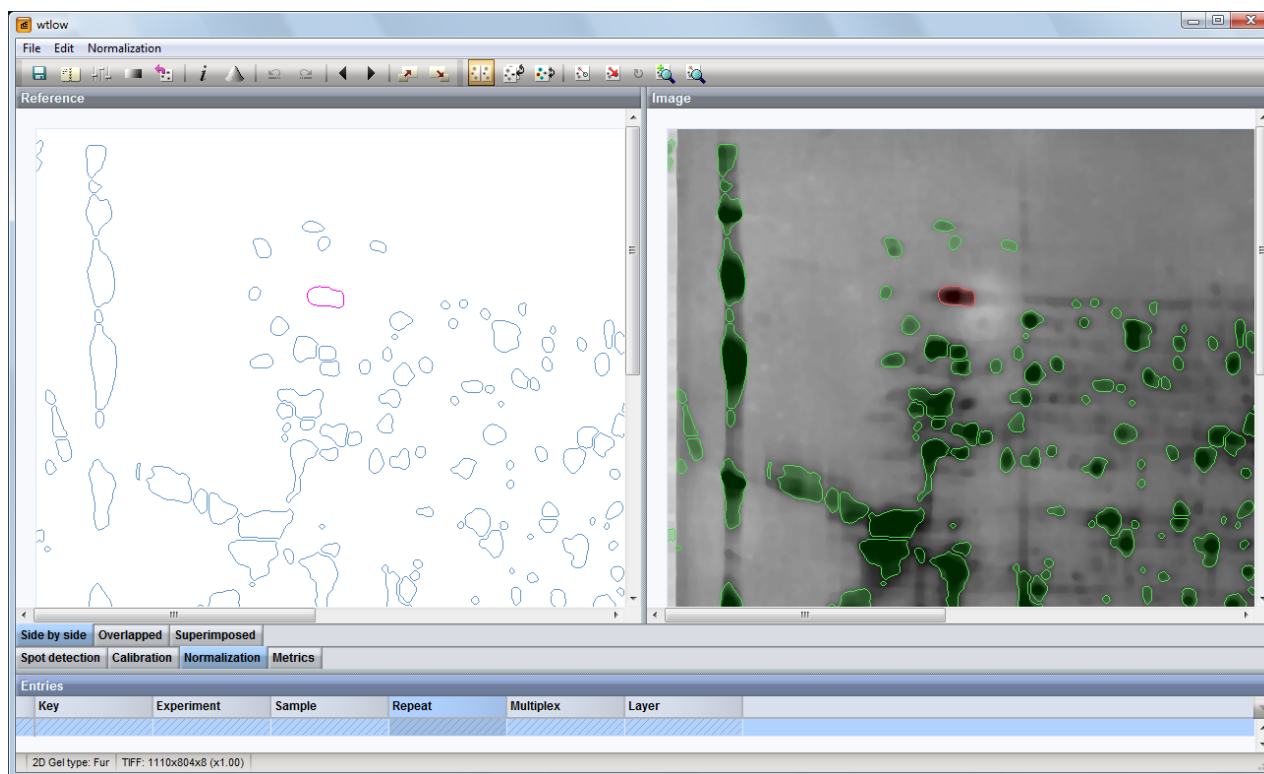


Figure 6-11. The *Normalization* step, initial view.

the basis for normalizing gels, and thus its spots should be taken from a normalized gel.

6.1.13.7 Turn the gel into normalized view by selecting **Normalization > Show normalized view** or pressing



6.1.13.8 To define the selected spot as a *reference spot*, select **Normalization > Add selected spot(s) to reference system**.

The reference spot now becomes visible in the reference system as a synthetic spot, of which the shape is derived from the mass center, the intensity, the X-size and the Y-size of the original spot (Figure 6-12).



Figure 6-12. Reference spot shown on the reference system: "synthetic" view mode.

6.1.13.9 By holding down the CTRL key you can select many spots at once, by clicking them one by one, and add them all together to the reference system (6.1.13.8).

6.1.13.10 Alternatively, you can select all spots in a rectangular zone by dragging a rectangle over the gel image.

6.1.13.11 Define all spots on the gel as reference spots using the commands as described in 6.1.13.10 and 6.1.13.8 (you can first zoom out to make selection of all spots easier).

The reference system is now shown as a synthetic gel. In an alternative viewing mode, the reference system can be shown as the original gel from which the spots were derived (the reference gel).

6.1.13.12 Toggle between synthetic reference system and original reference gel using **Normalization > Show synthetic reference system** and **Normalization > Show reference gel**.

NOTE: The purpose of a synthetic reference gel is to be able to combine and display spots from different gels. If additional reference spots were defined from other gels, these spots will not be shown when you display the original gel rather than the synthetic reference system.

Normalization of a gel happens in two steps. In the first step, homologous spots on the gel and the reference system are assigned, so that the image can be corrected until all homologous spots fall more or less together. In the second step, the program automatically searches for

the remaining homologous spots on the gel and the reference system, and links them.

• **Creating landmarks for normalization**

Since the reference system is derived from the current gel, the gel is already perfectly normalized. Hence, most features related to this step can be skipped for the first gel. We will discuss them when analyzing a second gel. However, to be able to compare 2D gels between different reference systems, the program needs at least a few well-distributed landmarks to be defined for each gel. For that purpose, we will simply perform an automatic search of landmarks, using a spot matching algorithm. Since gel and reference system are the same, the search will involve no manual editing.

6.1.13.13 Select **Normalization > Automatically find landmarks** or press



The program asks for the number of landmarks to be found. It is not recommended to enter a too high number. Instead, a moderate number, e.g. between 5 and 20 will allow the program to assign only the most pronounced and best corresponding spots. After updating the normalization, the user can further assign spots manually.

6.1.13.14 After entering a number and pressing <OK>, the program assigns a number of landmarks, which are indicated by a green cross on the landmark spot (Figure 6-13). The cross becomes red if the spot is selected.

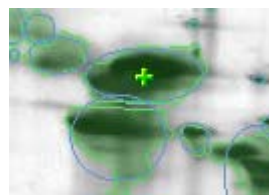


Figure 6-13. Spot defined as landmark.

6.1.13.15 Select **Normalization > Update normalization** or press



to update the normalized view according to the landmark data.

The image should not change if all landmarks were assigned correctly.

• **Linking spots with reference spots**

Once the gel is properly aligned to the reference gel, all the spots that have a homologous reference spot should occur very close to that reference spot. It is then relatively easy to allow an automatic matching algorithm to link spots and corresponding reference spots. However, since alignment is never perfect, some tolerance needs to be allowed. The size of that tolerance, entered in pixels, depends on the accuracy of the normalization, and the resolution of the gels.

Since all spots of the current gel were defined as reference spots, the program has automatically linked the spots with the reference spots. A linked spot is recognizable by a small green square in the center of the spot (Figure 6-3). Initially, all linked spots are unconfirmed, which is recognizable by a small hole in the center of the square (Figure 6-3). A spot or a selection of spots can be marked as confirmed using *Normalization > Mark selected spot(s) as confirmed* or simply by pressing CTRL+C. Conversely, confirmed spots can be marked as unconfirmed with *Normalization > Mark selected spot(s) as unconfirmed* (CTRL+U).

The square of a linked spot becomes red when the spot is selected (Figure 6-3). When selected, the linked reference spot also becomes selected (visible as a red mask). Conversely, when you select a reference spot in the *Reference* panel, the linked protein spot in the *Image* panel (if any) also becomes selected.

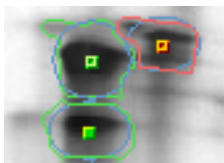



Figure 6-14. Spots that are linked to the reference system: unselected (green) and selected (red). Filled squares represent spots that are confirmed; non-filled squares are unconfirmed spots.


6.1.13.16 Select *Normalization > Automatically link spots* or press the  button.

The program asks to *enter the maximum deviation* (in pixels).

6.1.13.17 Leave the default value of 10 and press <OK>.

6.1.13.18 If the program has made a wrong linkage (not possible with this gel), you can select the spot and choose *Normalization > Unlink selected spot(s)*, or press the DEL key.

NOTE: You can link spots manually by dragging the spot node to the node of the corresponding spot on the other gel. Spots linked in this way automatically become landmarks. If you want spots to be linked without becoming landmarks, hold down the CTRL key while dragging the mouse pointer (see 6.1.16.37).

6.1.13.19 This finishes the *Normalization* part for this gel. Press the next step button () to move to the next step, the **Metrics** step.

6.1.14 Step 4: Defining metrics


As explained above (6.1.10), during the *Metrics* procedure, each spot can be identified by a metric in the horizontal direction and in the vertical direction. For protein gels this will usually be a pI value and a molecular weight, respectively. Since spots are identified using the reference system, the metrics definition is only an optional step, which is not necessary for any of the comparison tools in *InfoQuest FP 2D*. The metric descriptions will facilitate the comparison of spots from different databases. The X and Y metrics are calculated based upon spots with known X and/or Y metrics and using polynomial regressions for which the user can choose the degree (1-5) and a logarithmic dependency.



A third specification which can be defined for each protein spot is its quantity. The quantity (also referred to as Z-metric) is calculated from spots with known quantity, using a polynomial regression of degree 1 to 5. As opposed to the X and Y metric, the quantity can have influence on the comparisons, since spot volumes derived from scanned 2D images are usually not linear with spot quantities. By applying spots with known quantities on the gel, the user can let the program linearize the spot volumes into physical quantities.


6.1.14.1 As an exercise, you can enter the molecular weights and pI values of four known spots, as depicted in Figure 6-15.

6.1.14.2 Double-click on a spot with known MW and pI. A dialog box will prompt you to enter a pI value (X), a MW value (Y) and a quantity (Z).

6.1.14.3 Enter the appropriate values for pI and MW, and leave the quantity field blank. Press <OK>.

NOTE: In case a spot used as calibration spot in Figure 6-15 is not defined in your gel, you can return to step 1 to add this spot at any time, either by pressing repeatedly the  button or by clicking on the Spot detection tab in the bottom of the Image panel.

When an X value has been entered for a spot, it is marked with a bidirectional horizontal arrow: . Likewise, when a Y value has been entered for a spot, it is marked with a bidirectional vertical arrow: . In case both the X and Y values were entered for a spot, it is marked by the combination of these two arrows, as shown in Figure 6-15.

6.1.14.4 When the pI and MW values have been entered for the 4 known spots, call *Edit > Settings* or press the  button.

6.1.14.5 In the *2D gel settings* dialog box, select the *Metrics* tab, which looks as in Figure 6-16.

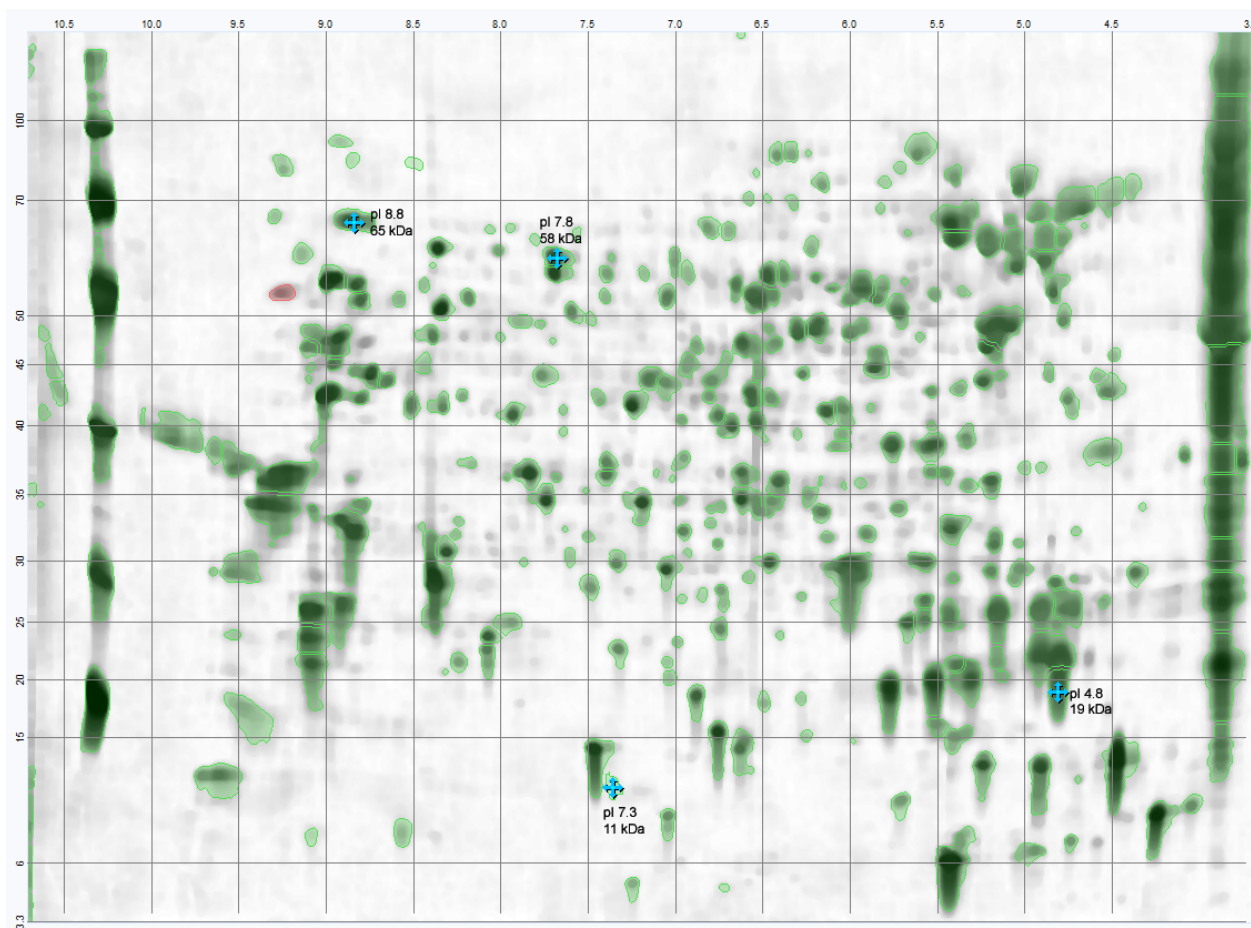


Figure 6-15. Example for entering pI values and molecular weights for known protein spots in gel Wtlow.

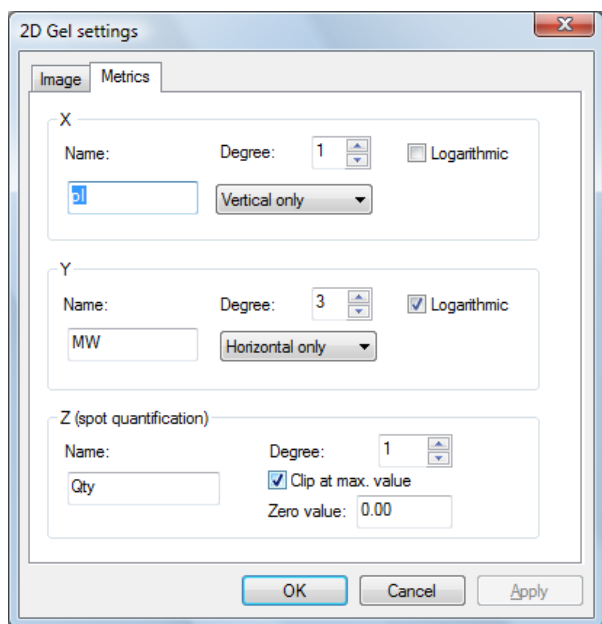


Figure 6-16. The InfoQuest FP 2D gel settings dialog box, Metrics tab.

6.1.14.6 Under *Name*, enter the X and Y metrics, which are pI and MW, respectively.

The entered X values (usually pI) and Y values (usually MW) will be used by the software to calculate a linear or exponential fit which can be used to identify exact spot metrics all over the gel. InfoQuest FP 2D uses fitting algorithms of 1st to 5th *Degree* and can add a *Logarithmic* dependency to the fitting algorithm.

6.1.14.7 For the X metric (pI) choose 1 as *Degree*, and do not check *Logarithmic*.

6.1.14.8 For the Y metric (MW), select 3 as *Degree*, and check *Logarithmic*, since MW electrophoresis runs usually exhibit logarithmic dependence.

As an additional option, one can specify whether the isometric values of the X metric should be strictly vertical or not. In a pull-down box you can choose between *Vertical only*, *Rotated*, and *Rotated & curved*.

- When *Vertical only* is selected, the program assumes no rotation of the gel, and isometric values are vertical.
- When *Rotated* is selected, the program tries to find the best fit through the given marker points with an additional rotation freedom. This means that, if a better fit can be found by rotating the X-isometric lines over a certain angle, this angle will be used.

- With *Rotated & curved*, the program is allowed to add some curvature to the isometric lines to provide an even better fit.



It is obvious that the latter two options require enough input values to become reliable, especially the option *Rotated & curved*. It is generally not recommended to use this option.

The same options apply to the Y-metric.

6.1.14.9 Since the gel is somewhat rotated clockwise, you can try choosing *Rotated* for both X and Y metrics.

6.1.14.10 Press <OK> to close the *2D gel settings* dialog box and confirm the changes.


The 2D gel image now displays a grid, defining the pI values in the horizontal direction and the MW in the vertical direction. You will notice that - if the metrics were entered as in Figure 6-15 - the pI isometrics (vertical lines) are slightly rotated clockwise.


6.1.14.11 Rotation of a gel can also be compensated for by manually rotating the isometrics grid. This can be done by clicking the  and  buttons to rotate counterclockwise and clockwise, respectively.

*NOTE: Manual rotation should not be performed when **Rotation** was selected as an option in the Metrics settings.*

The third or Z metric is intended to express spot intensities as physical quantities.

6.1.14.12 Although there is no real quantitative information available for this gel, you can consecutively select a dark spot, an intermediate spot, and a weak spot, each time entering a lower Z metric value, for example 100, 50, and 10, respectively.

Spots for which a Quantity (Z) metric has been entered are marked with a weight symbol: .

6.1.14.13 Call the *2D gel settings* dialog box again with *Edit > Settings* or by pressing the  button.

6.1.14.14 Select the *Metrics* tab (Figure 6-16).

6.1.14.15 Under **Spot quantification**, you can enter a *Name* for the metric, for example "Quantity".

6.1.14.16 As *Degree* for fitting, enter 2 (second degree exponential fitting).

The option *Clip at max. value* makes it possible to restrict the calculation of quantity to the range within the marker spots entered. Spots that have more volume than the highest calibration spot will be clipped at the

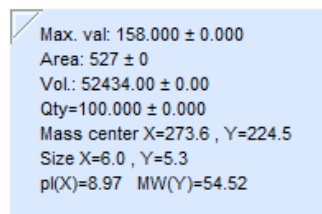


Figure 6-17. Spot information box.

maximum value. This option is useful in case you want to avoid extrapolation of the fit beyond the maximum value entered.

The option *Zero value* allows one to specify a quantity value on the image that corresponds to a zero intensity value. In other words, it will add a spot to the regression with zero intensity value and a quantity value of X.

6.1.14.17 Press <OK> to close the *2D gel settings* dialog box and confirm the changes.

6.1.14.18 In order to view the information stored for every spot you can press the <Show spot info> button



or select the menu option *Edit > Spot info*.

As a result, a small information box will be opened that will display the following information (Figure 6-17):

- The maximum intensity of the spot (maximum gray level value in image file).
- The total area (or surface taken on the image file) in number of pixels.
- The volume as calculated using all gray values of the spot.
- The quantity as calculated from the metrics.
- The absolute position of the mass center using the X and Y positions in the image file.
- The average X and Y size in pixels of the protein spot.
- The X and Y metrics, respectively as calculated from the spot's mass center.

Note that standard deviations are calculated for the maximum value, the area, the volume and the quantity, in case the gel is a synthetic average of several individual gels.

Processing of gel Wtlow is now finished. The gel can be saved.

6.1.14.19 Press <F2> or the  button or select *File > Save*.


6.1.15 Step 5: Describing the 2D gel in the database

Before this gel can be analyzed, i.e., queried and compared with other 2D gels, it needs to be described in the database. We will need to create a database entry that has this gel linked to it. The entry holds the descriptive information of the gel: the organism name, sample number, the experiment conditions, the running conditions, or whatever information that is applicable. The way database information fields are created and filled in is described in and .

In the *2D gel processing* window, the link to a database entry is shown in the dockable *Entries* panel (for display options of dockable panels, see 1.6.4).

There are two ways to link the 2D gel to an associated database entry: (1) if the organism or sample is not available yet, by letting the program automatically create a new entry for the gel, or (2) if the organism or sample is already described in the database, by linking the gel to that existing entry.

Since our database is empty, we will create a new entry for the gel *Wtlow*.

6.1.15.1 In the *2D gel processing* window of *Wtlow*, press the button  or select *File > Add to database*.

A dialog box pops up asking you to enter the database key of the new link. Define a key and press **<OK>**.

6.1.15.2 Close the *2D gel processing* window for *Wtlow*.


6.1.15.3 Using *Database > Add new information field* from the *InfoQuest FP main* window, create three information fields in the database: **Organism**, **Taxon**, and **Condition**.

6.1.15.4 Double-click on the new database entry and enter the following information under:

- Organism: **Wild type**
- Taxon: **Campylobacter jejuni**
- Condition: **Low Fe concentration**

6.1.16 Step 6: Normalization of other 2D gels


When processing the first 2D gel, we have defined the *Reference System* based upon that gel (6.1.13), so the normalization step could be skipped in that gel. We will now process a second gel to illustrate the normalization features in particular.

6.1.16.1 First, create a new database entry with *Database > Add new entries* or .

6.1.16.2 Press **<OK>** to have the software automatically assign a key to the entry.

6.1.16.3 Double-click on the new database entry and enter the following information under:

- Organism: **Wild type**
- Taxon: **Campylobacter jejuni**
- Condition: **High Fe concentration**

6.1.16.4 In the *Files* panel of the *InfoQuest FP main* window, select file **Wthigh**, and press the  button to open the file (or double-click on the file name).

A dialog box prompts you to select the 2D gel type to which the gel should belong.

6.1.16.5 Select **Fur** (the only existing 2D gel type) and press **<OK>**.

NOTE: Opening a second gel in a specific experiment type will automatically apply the settings that have been specified in the experiment. Consequently, opening a second gel or consecutive gels may appear to be slower since background and spikes may be removed, smear removed and filtering applied.

6.1.16.6 Perform Step 1 and Step 2 of the 2D gel processing as described earlier (6.1.11 and 6.1.12).

When moving to Step 3 (Normalization), a dialog box prompts you to select the *Reference System* to which the gel should be assigned.

6.1.16.7 Select **Fur** (the only existing reference system) and press **<OK>**.

In the *Normalization* step, an additional *Reference* panel becomes available. The *Reference* panel displays the reference system **Fur**, and the *Image* panel shows the current data gel **Wthigh**. When you select a spot in the *Image* panel (click on the spot), its contour in the *Reference* panel also becomes highlighted (red instead of blue) (see Figure 6-11).

The reference system is currently shown as a synthetic gel. In an alternative viewing mode, the reference system can be shown as the original gel from which the spots were derived (the reference gel).

6.1.16.8 Toggle between synthetic reference system and original reference gel using *Normalization > Show synthetic reference system* and *Normalization > Show reference gel*.

NOTE: If additional reference spots were defined from other gels than the initial reference gel, these spots will

not be shown when you display the original gel rather than the synthetic reference system.


Normalization of a gel happens in two steps (6.1.13.12): (1) homologous spots on the gel and the reference system are assigned (*landmarks*), and (2) the program automatically searches for the remaining homologous spots on the data gel and the reference system, and links them.

• Creating landmarks for normalization

InfoQuest FP 2D contains an automatic searching tool for landmarks, using a spot matching algorithm. We have described this feature earlier (6.1.13.13). For very different gels, however, this feature will not always provide satisfactory results.


NOTE: You can try the automatic landmark finding tool at any time; if the result is not satisfactory, press the Undo button.

When matching a new gel to a reference system manually, the correctness and the distribution of the chosen landmarks will be a crucial factor in determining the quality of the match. InfoQuest FP 2D offers a number of viewing modes to facilitate the definition of landmarks and verify their effect on the matching. By default, the reference system and the data gel are displayed side by side. In InfoQuest FP 2D, however, there are two additional modes for display: the *Overlapped mode* and the *Superimposed mode*.

6.1.16.9 Show the gels in overlapped mode by clicking on the *Overlapped* tab in the bottom of the *Image* panel or by pressing the  button (menu option: *Normalization > Show overlapped images*).


In this mode, you only see one gel at a time, initially the gel to be matched.

6.1.16.10 By pressing the TAB key or by clicking on the *Reference* or *Image* tab in the bottom of the *Image* panel (menu item: *Normalization > Swap data/Reference*), you can toggle between viewing the data gel and the reference system.

6.1.16.11 Show the images in *Superimposed mode* by clicking on the *Superimposed* tab in the bottom of the *Image* panel or by pressing the  button (menu option: *Normalization > Show superimposed images*).

The gels are now shown in two colors: the data gel to be matched in orange and the reference system in blue. Spots that overlap each other in both gels become black.

Creating landmarks in side-by-side mode.

6.1.16.12 Press the *Side by side* tab in the bottom of the *Images* panel or the  button in the toolbar.

The spots of the reference gel are also shown on the data gel (*Image* panel) as blue contours. Likewise, the spots of the data gel are also shown as blue contours on the reference gel in the *Reference* panel (Figure 6-12).


6.1.16.13 Select a spot on the data gel (right panel).


The selected spot is highlighted by a red contour. The contour of this spot in the *Reference* panel also becomes red.

6.1.16.14 Select the corresponding spot on the *Reference* panel.

The selected spot is highlighted by a red contour. The contour of this spot in the *Image* panel also becomes red. We will now create a landmark by linking the spot with its homologous reference spot.


6.1.16.15 Select *Normalization > Link spot as landmark*, or press ENTER, or press .

We have now created a *landmark*, which is indicated by a green cross on the spot (Figure 6-13). The cross becomes red if the spot is selected. Since the spots on the data gel and the reference system are linked, both spots are highlighted if either of them is selected. If a landmarked spot is selected, the landmark button becomes highlighted (.


6.1.16.16 To show the data gel in normalized view, i.e., to warp the image so that all landmarked spots fall together with their reference spots, press  or select *Normalization > Show normalized view*.

Once normalized to one or a few landmarks, it may become easier to assign additional landmarks.

6.1.16.17 Select a few more spots on the gel and homologous spots on the reference system, to create more landmarks.

6.1.16.18 Select *Normalization > Update normalization* or press  to update the normalized view according to the current landmark data.

Creating landmarks in overlapped mode

6.1.16.19 Show the gels in overlapped mode by clicking on the *Overlapped* tab in the bottom of the *Image* panel or by pressing the .

6.1.16.20 By pressing the TAB key or by clicking on the *Reference* or *Image* tab in the bottom of the *Image* panel, you can toggle between viewing the data gel and the reference system.

There are two ways to create landmarks in this mode. The first way is the same as in the side-by-side mode.

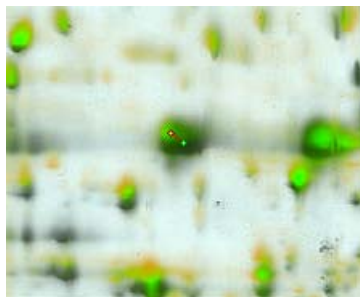



Figure 6-18. Reference gel superimposed on gel in Overlapped mode.

6.1.16.21 Select a non-landmarked spot on the data gel, and click on it to select the spot (marked by a red dot).

6.1.16.22 Press the TAB key to view the reference gel, and select the homologous spot.

6.1.16.23 Press ENTER (or ) to create a landmark for this spot.


6.1.16.24 The second way is quicker and easier: click on a spot on the data gel to landmark, hold down the left mouse button, and drag the mouse slightly over the screen.

The reference gel is shown semi transparently in yellow over the gel (Figure 6-18). Overlapping spots between the gel and the reference gel are green. If a spot on the reference gel is in the close vicinity of the data gel, the program automatically suggests a link with a red connecting line, as shown in Figure 6-18.

NOTE: The quick drag-and-drop method can only be used if the data gel is shown. If this is not the case, press TAB and try again.

6.1.16.25 Drag the mouse until the homologous spot of the reference gel is linked to the selected spot on the data gel, and release the mouse button.

The superimposed mode

6.1.16.26 Show the images in *Superimposed mode* using the menu option **Normalization > Show superimposed images** or by pressing the  button.

The gels are now shown in two colors: the data gel in orange and the reference system in blue. Spots that overlap each other in both gels become black (Figure 6-19).

Similar as in the overlapped mode, there are two ways to create landmarks in this mode. The first way is the same as in the side-by-side mode:


6.1.16.27 Select a non-landmarked spot on the data gel, and click on it to select the spot (marked by a red dot).

6.1.16.28 Click on the *Reference* tab in the bottom of the *Image* panel or press the TAB key to view the reference gel, and select the homologous spot.

6.1.16.29 Press ENTER (or ) to create a landmark for this spot.

6.1.16.30 The second way is both quicker and easier: Click on a spot from the data gel to landmark (orange), hold down the left mouse button, and drag the mouse slightly over the screen to the homologous reference spot (blue).

NOTE: The quick drag-and-drop method can only be used when spots on the data gel (orange) are shown. If this is not the case, press TAB and try again.

6.1.16.31 Select **Normalization > Update normalization** or press  to update the normalized view according to the landmark data.


6.1.16.32 You can view the distortion applied for the matching by selecting the menu option **Normalization > Show distortion maze**. The result looks as in Figure 6-19.

6.1.16.33 To remove an incorrect landmark, you can also select it and press the DEL key.

NOTE: Use the Undo and Redo functions to undo/redo the last actions.

•Linking spots with reference spots

Once the gel is properly aligned to the reference gel, all the spots that have a homologous reference spot should occur very close to that reference spot. It is then relatively easy to allow an automatic matching algorithm to link spots and corresponding reference spots within a certain tolerance, entered in pixels.

6.1.16.34 Select **Normalization > Automatically link spots** or press the  button.

The program asks to **enter the maximum deviation** (in pixels).

6.1.16.35 Leave the default value of 10 and press <OK>.

A linked spot is recognizable by a small green square in the center of the spot (Figure 6-3). Initially, all linked spots are unconfirmed, which is recognizable by a small hole in the center of the square (Figure 6-3). A spot or a selection of spots can be marked as confirmed using **Normalization > Mark selected spot(s) as confirmed** or simply by pressing CTRL+C. Conversely, confirmed spots can be marked as unconfirmed with **Normalization > Mark selected spot(s) as unconfirmed** (CTRL+U).


A linked spot becomes red when the spot is selected (Figure 6-14). When selected, the linked reference spot also becomes selected (visible as a red mask).

Conversely, when you select a reference spot in the *Reference* panel (in side-by-side mode), the linked protein spot in the *Image* panel (if any) also becomes selected.

6.1.16.36 If the program has made a wrong linkage, you can select the spot and choose *Normalization > Unlink selected spot(s)*, or press the DEL.

6.1.16.37 You can link spots manually by dragging the spot node to the node of the corresponding spot on the other gel. Spots linked in this way automatically become landmarks. If you want a spot to be linked without becoming a landmark, hold down the CTRL key while dragging the mouse pointer.

6.1.16.38 This finishes part 2, **Normalization** for this gel.

Press the next step button () to move to the next step, the **Metrics** step, where you may enter some known spot metrics as discussed in 6.1.14.

6.1.16.39 Save the gel **Wthigh** and close the *2D gel processing* window.

The program may ask "**Settings have been changed. Do you want to use the current settings as the new default?**". If you feel the changed settings for the 2D gel type will be useful for future gels as well, answer **<Yes>**. If the changes were only necessary to improve the processing of this individual gel, press **<No>**.

The program may also ask to confirm that the "*configuration has been changed*". This question comes up when the current gel has been changed.

To illustrate the comparison functions in the next section, it is recommended to add a few other gels to the database **Demo2D**. You may want to process gels **Furlow** and **Furhigh**. Alternatively, open the **Demobase 2D** database which contains the four processed gels. **Demobase 2D** is installed with the software.

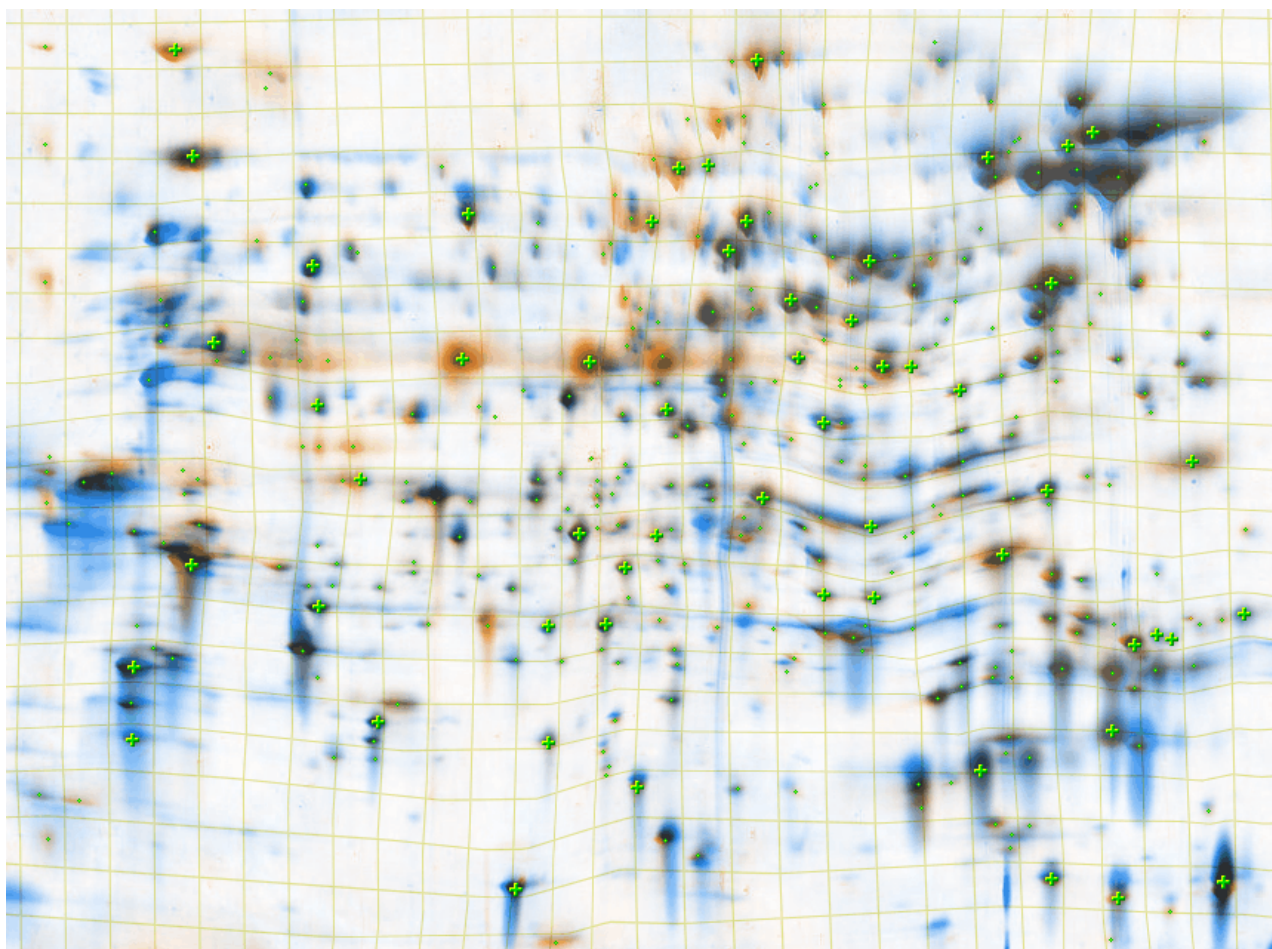


Figure 6-19. Data gel and reference gel in orange and blue, respectively, in Superimposed mode (distortion maze shown).

6.2 Comparing 2D gels 2D

6.2.1 Introduction

One of the main purposes of analyzing 2D gels is to detect proteins that are invariably expressed or differentially expressed in different circumstances. Another application could be to compare patterns of protein expression between different organisms, in the same circumstances. All these applications require that spots representing the same protein are *linked* to each other. This is done by (1) normalizing different gels to a common *reference system* (6.1.13) and (2) by linking spots of the gel to the homologous reference spots (6.1.13.16).

During the normalization procedure, two spots from different gels may be linked to the same reference spot (Figure 6-20). Internally, the software stores a unique identifier for each protein spot on each gel. The spots on the reference system also have an identifier. When a spot is linked to a reference spot, it gets the same identifier as that reference spot, so that the program recognizes it as the same protein. When a spot on another gel is linked to the same reference spot, it also gets the same identifier, so that the program recognizes the spots on both gels as the same. If $A = B$ and $B = C$ then $A = C$.

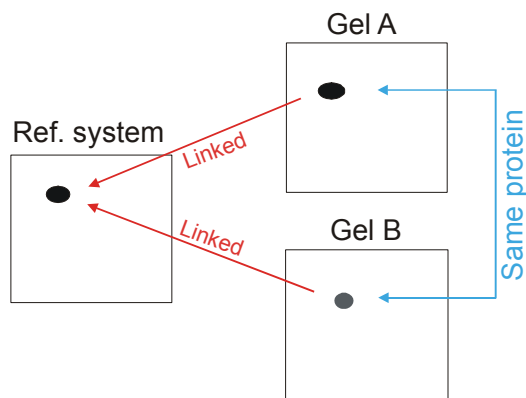


Figure 6-20. Indirect linking of spots via reference spots: two spots linked to the same reference spot are recognized as the same protein.

In the *2D gel matching* window (discussed further in this section) there is also a possibility to link two gels directly to each other, without linking to a reference spot. One can, for example, link spot A from one gel to spot B from another gel, without A or B being linked to a reference spot (Figure 6-21). In this case, both spots A and B will get the same identifier and will be recognized

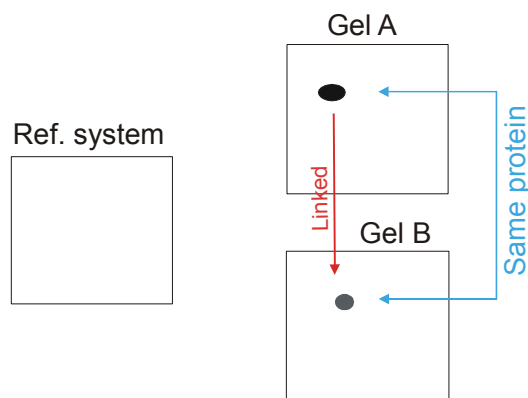


Figure 6-21. Direct linkage of two spots in the 2D gel matching window.

as the same protein, without a reference spot being present for this protein

Each spot in the database represents a certain protein; linked spots represent the same protein. InfoQuest FP stores information fields for each protein. The software stores a *Spot ID*, which is assigned automatically, and a number of additional fields that can be filled in by the user:


- **Accession:** An accession number for the protein so that links to external databases can be made;
- **Description:** A description of the protein such as function, pathway;
- **Gene name(s):** name or code of the gene relating to the protein, and synonyms;
- **Field 1, Field 2, Field 3, Field 4, and Field 5:** free user-definable fields.


'Description', 'Gene name(s)' and the free fields can contain strings of unlimited length.

6.2.2 Matching spots on different gels

The goal of comparing 2D protein gels is to either compare overall protein expression patterns (in different conditions or over different organisms) or to examine the expression level of individual proteins in function of different conditions. Such analyses usually require some kind of clustering or grouping algorithm, but in the first place, require that homologous spots on different gels are properly linked together. The *2D gel matching* window is designed for this purpose.

6.2.2.1 Select the four *Campylobacter jejuni* entries in the database **Demo2D**: Wild type with low Fe concentration, Wild type with high Fe concentration, Fur mutant with low Fe concentration, Fur mutant with high Fe concentration. Use the space bar on the keyboard or click on the entries while holding the SHIFT or CTRL key.

6.2.2.2 Copy the gels to the clipboard by selecting **Edit > Copy selection** or by pressing .


6.2.2.3 In the *Experiments* panel, open 2D gel type **Fur** by double-clicking or pressing the  button.

This opens the 2D gel type window for experiment type **Fur** (Figure 6-22).

The 2D gel type window contains information which is general for the selected 2D gel type, and which is saved along with the experiment type, in this case **Fur**. This information includes the reference system(s) defined within the 2D gel type, the image processing settings, the free field names, the spot queries, and the spot quantification settings. We will deal with queries later (see 6.2.3).

6.2.2.4 The names for the 5 optional *free fields* (see the introduction, 6.2.1) can be entered from the 2D gel type window, by selecting **Settings > Spot label names**. A


dialog box pops up where you can enter a name for each of the free fields.

6.2.2.5 Call the 2D gel matching window by selecting **File > Create matching window** or by pressing .

The 2D gel matching window (Figure 6-23) consists of 3 views, which can be selected by clicking on their respective tabs in the bottom of the window: the *Gel images* view, which is selected by default, the *Query table* view, and the *Scatter plots* view. The latter two views are described in paragraphs 6.2.4 and 6.2.5, respectively.

The 2D gel matching window is intended to display many gels next to each other. Initially, only the active reference system is displayed.

*NOTE: The active reference system is the reference system used for comparisons in the 2D gel matching window. You can change the active reference system in the 2D gel type window by selecting another reference system in the Reference systems panel and choosing **Refsystem > Set as active reference system**.*

6.2.2.6 Paste the gels from the clipboard in the 2D gel matching window by selecting **Edit > Paste entries from clipboard** or pressing .

The 2D gel matching window now displays the reference system and four 2D gels (Figure 6-23).

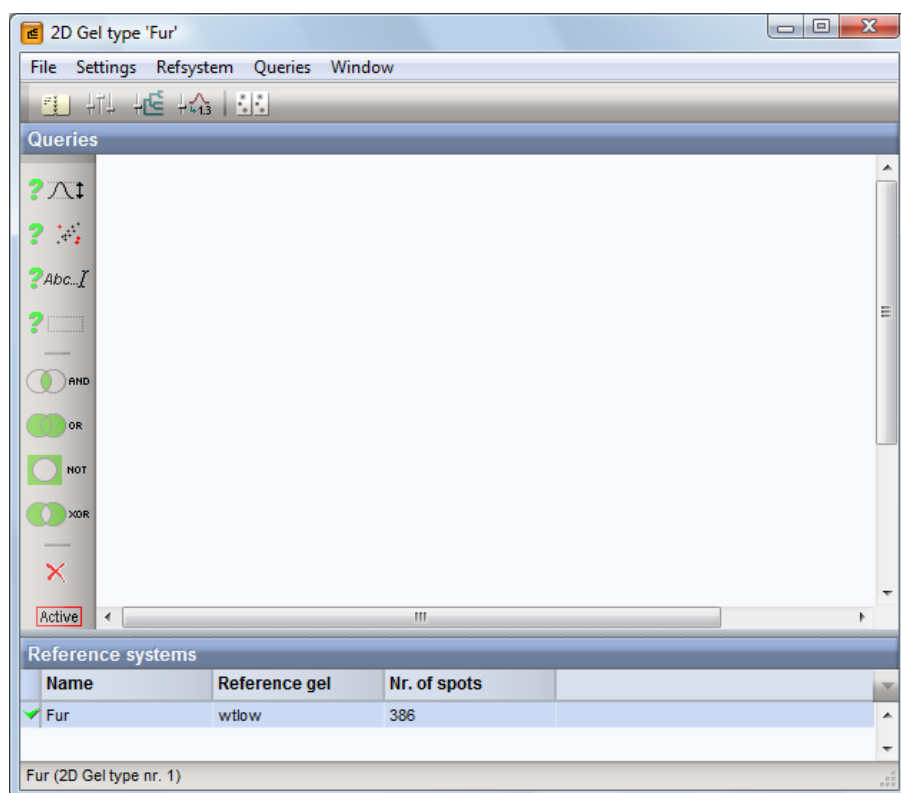


Figure 6-22. The 2D gel type window.

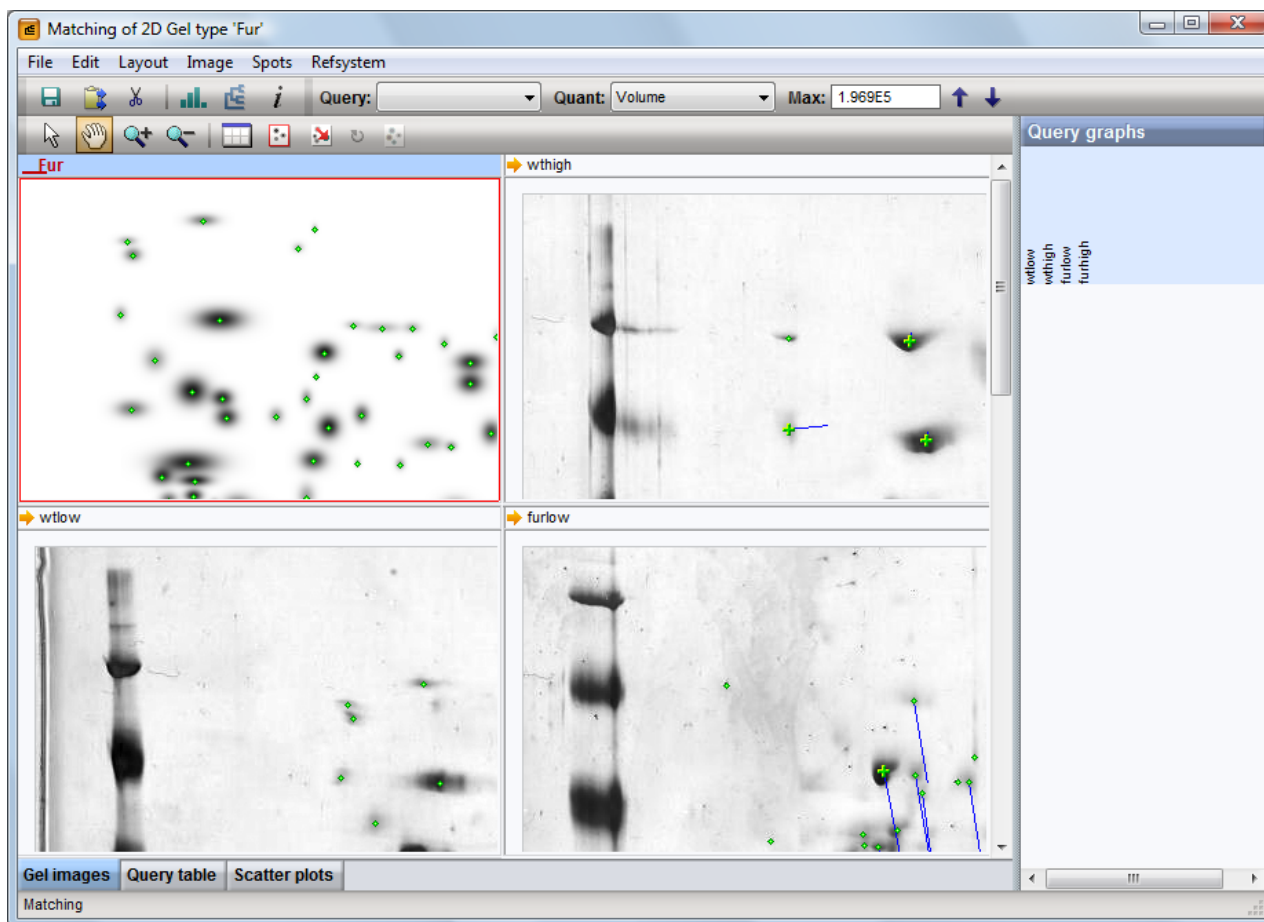



Figure 6-23. The 2D gel matching window.

In 5 separate panels, the four gels and the reference system are displayed. The reference gel is bordered by a red rectangle. For each gel, the landmarks defined in the normalization step are indicated as green thick crosses and the proposed shifts towards the homologous position on the synthetic reference gel are indicated.


6.2.2.7 The position of the gels can be changed in this window by clicking on the name of the gel and dragging it to its new location.

Initially, the gels are not shown in normalized mode.

6.2.2.8 To show all gels in the normalized mode, press the  button or select *Image > Show normalized*.



6.2.2.9 When selecting a protein spot on one of the gels (or on the reference gel), InfoQuest FP 2D will indicate the homologous protein, if present, on all other gels displayed in the 2D gel matching window. A label is shown, pointing to the selected spot, and indicating an quantification value of the spot on each gel. This quantification value can be the *Maximum value*, *Volume*, *Relative volume*, *Quantified value*, and *Area*. The quantification value to be shown can be chosen from the drop-down list in the button bar of the 2D gel matching window.

NOTE: The default settings for the quantification value can also be specified in the Spot quantification settings dialog box (see Figure 6-33).

6.2.2.10 By selecting the image dragging tool () button) you can drag the image to any part of the gel.

6.2.2.11 You can also click and drag using the right mouse button to navigate through the images, without having to select another pointer.

If gels are shown in normalized mode, each displacement of one gel will be automatically followed by the other gel images.

6.2.2.12 You can also use the zoom tools ( and ) to zoom in or out in each gel window. To zoom in you can drag a rectangle on the region of interest.

In the 2D gel matching window, it is possible to improve or correct the normalizations of gels made earlier, by re-linking spots, removing links or adding links.

6.2.2.13 To link a spot on a gel to a reference spot, click on the spot, and hold down the left mouse button while dragging the mouse to the homologous reference spot.

The mouse pointer changes from a prohibition sign into a symbol of two linked spots.

NOTE: In the same way, spots can also be linked between gels directly, without linking to a spot on the reference system. Such spots will also be recognized as the same protein and hence, share the same information fields. They will, however, not be assigned a Spot ID (see further).

6.2.2.14 In the 2D gel matching window, you can also display the gels in the *Superimposed mode* by using the menu option **Image > Show overlay** or by clicking the




button.

The gels are now shown in two colors: the data gels in orange and the reference system in blue. Spots that overlap each other in both gels become black.

6.2.2.15 If you click on a spot on the gel of interest, you can drag it to the homologous spot position of the underlying reference gel. When the correct position on the reference gel has been reached, the mouse pointer changes from a prohibition sign into a symbol of two linked spots. At the same time, the reference spot is bordered by a red square on the reference gel. You can release the mouse button to establish the link.

6.2.2.16 The linked spot is now defined as a landmark position in the original gel.


6.2.2.17 For each spot you can break existing links by selecting the spot on the gel and selecting the menu option **Spots > Break link** or by pressing the **DEL** key.

6.2.2.18 After you created new links you can press the button  or select the menu option **Image > Update normalization** which will re-normalize the modified gel(s) and return the display to the single-gel, normalized mode (not superimposed).

NOTE: For extensive matching and linking work, we recommend to use the 2D gel processing window (see Section 6.1), which contains a multistep undo/redo function.

6.2.2.19 At any time, you can have the program perform a matching again using **Spots > Automatic match**. When doing so, the program will prompt to enter the maximum allowed match distance in pixels.


6.2.2.20 For a spot or a selection of linked spots, you can set the flag to Confirmed with **Spots > Mark spot link as confirmed**. The cross changes into a filled square.


6.2.2.21 You can save the changes by pressing the  button or select **File > Save changes**.

At this stage, it is still possible to change the synthetic reference system by adding or deleting spots.


6.2.2.22 To add spots to the reference system, select one or more spots on a gel and choose **Refsystem > Add selected spot(s)**.

6.2.2.23 To delete one or more spots from the reference system, select a spot on the reference gel and choose **Refsystem > Delete selected spot(s)**.

6.2.2.24 At any time you can add or delete a gel to/from the matching by pasting new gels from the clipboard or by selecting **Edit > Cut selected gel from matching** or by pressing the  button.

6.2.2.25 When doing so it may be useful to change the layout of the 2D gel matching window by clicking the  button or by selecting a grid layout from the **Layout** menu (*1x1 grid, 2x1 grid, 2x2 grid, 3x2 grid, or 5x3 grid*).

In order to match two or more data gels more exactly (not via a reference gel), you can turn any gel into a temporary reference gel. This will allow two non-reference gels to be shown in superimposed mode. For example, you may want to display the gel from which the reference system was derived (**Wtlow**) in superimposed mode rather than the synthetic gel.

6.2.2.26 Click on the gel to use as temporary standard (e.g. **Wtlow**) and select **Refsystem > Use selected gel as temporary standard** or press the  button.

*NOTE: You can remove the synthetic reference gel from the 2D gel matching window (**Edit > Cut selected gel from matching**) and continue to improve the matching between the remaining gels. A reference system can also be added to a 2D gel matching window, by bringing the 2D gel processing window to the front (Figure 6-22), choosing a reference system from the list in the Reference systems panel, and selecting **Refsystem > Add to matching window**.*

As mentioned earlier, spots that are not present in the reference system can also be matched between gels; these spots will have no spot ID.

Based on the 2D gel matching window, the software will be able to assign an ID code to all protein spots present on the reference system. Spots that are linked to the same reference spot in the synthetic reference gel will have the same ID code.

6.2.2.27 You can assign the spot IDs by selecting the option **Refsystem > Assign ID code to spots**.

The ID code assigned to the spots will have a permanent nature if the matching is saved to disk.

6.2.2.28 You can hide the label by selecting **Layout > Label with > No field**.

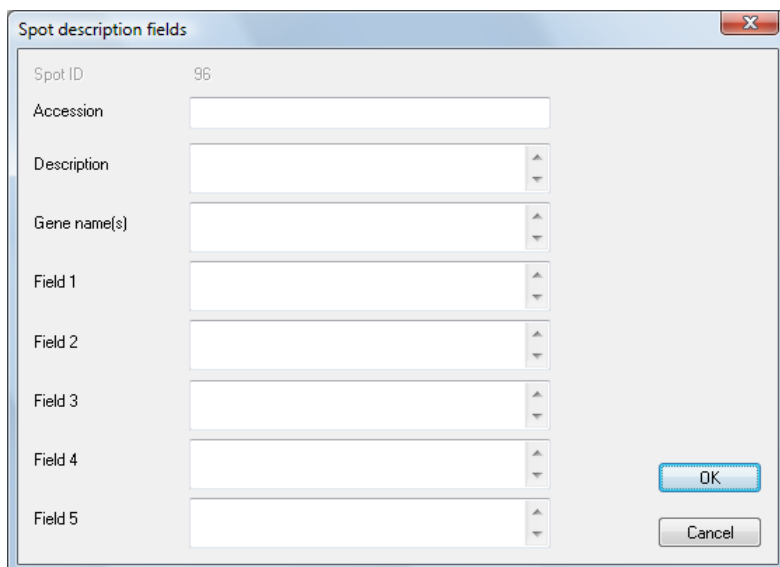


Figure 6-24. The *Spot description fields* dialog box.

The label given by the software is only a primary identifier and can be supplemented by other information.


6.2.2.29 To add additional information, double-click on the spot you like to document, or select *Spots > Change description (current selection)*.

The *Spot description fields* window will open (Figure 6-24) and you can type the information for that spot.

6.2.2.30 The *Spot description fields* window has the three standard fields *Accession code*, *Description* and *Gene name(s)* as well as five optional free fields where you can store other types of information. The information loaded can be typed from the keyboard or can be loaded from public databases using InfoQuest FP scripts.

NOTE: The names for the free fields can be changed in the 2D gel type settings, as described in 6.2.2.4. The information is available for querying only after saving the gel.

6.2.2.31 In case more than one spot is selected and you select *Spots > Change description (current selection)*, the program will prompt that "**There are x spots selected. Do you want to modify all spots simultaneously?**". In case of confirmation with **<Yes>**, a variant of the *Spot description fields* dialog box pops up (Figure 6-25), from which you can choose a field, and enter the string that should be filled in for the selected spots.

6.2.2.32 The information stored can be viewed quickly for each spot by selecting the menu option *Layout > Show spot info* or by pressing the  button.

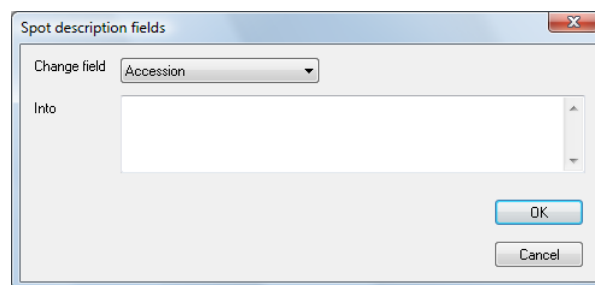


Figure 6-25. *Spot description fields editor* dialog box for multiple selected spots.


6.2.3 Creating 2D spot queries




InfoQuest FP 2D contains a spot querying tool that allows searches to be performed based on spot intensities and spot information fields on gels of selected entries in the database. The resulting set of spots compose a data set that is amenable to further statistical analysis by the InfoQuest FP software, for example, cluster analysis, principal components analysis, discriminant analysis, self-organizing maps, MANOVA, etc.


If you continue from paragraph 6.2.2, you may still have the four database entries in database **Demo2D** selected, and the *2D gel processing* window opened. If not, proceed as in 6.2.3.1 to 6.2.3.2.

6.2.3.1 In the *Database entries* panel of the *InfoQuest FP main* window, select the 2D gels you want to analyze. In this example we will use all four gels **Wtlow**, **Wthigh**, **Furlow**, and **Furhigh** of database **Demo2D**.

6.2.3.2 Double-click on the 2D gel type **Fur** to open the *2D gel type* window (Figure 6-22).

The query tool allows you to create individual *query components*, which can be combined into more complex queries with *logical operators*. The available query types are *Intensity query* , *Significance query*

, *Spot field query* , and *Manual selection* . The available logical operators are **AND**


, **OR** , **NOT** , and **XOR** .

• Intensity queries

With intensity queries, spots can be selected based upon one of the available intensity measures (spot height, volume, quantity).

6.2.3.3 Before creating a new intensity query, make sure the gels you want the query to apply to are selected. This is indicated as a colored arrow left from the gel name in the caption of the images.

6.2.3.4 If the gels are not selected, use CTRL+click in the gel image caption to select them.

6.2.3.5 To prepare a new query click the  button or select *Queries > New intensity query* and enter a name, for example, "**Differential expression**". Press **<OK>**.

6.2.3.6 In the *Intensity query* dialog box (Figure 6-26) select *Volume* as the *Spot intensity measure*.

Other measures to construct intensity queries are *Maximum value* (highest pixel intensity), *Relative volume (in %)* (related to all spots on the gel as 100%) and *Quantity* (according to the Z metric in step 4 of the normalization; see 6.1.14.15).

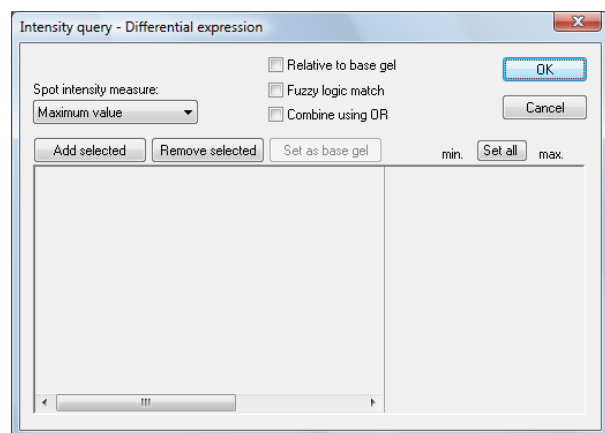


Figure 6-26. *Intensity query* dialog box.

6.2.3.7 Press the button **<Add Selected>** to include the selected gels in the query. Four gels should now be listed. If not, make sure the database entries/gel images are selected (colored arrow).

You can now define the search criteria for the spots. For each gel you can enter a minimum and a maximum intensity value, by clicking on the --- field under *Min.* and *Max.* respectively for the corresponding gel.

6.2.3.8 Under *Max.* for **Wtlow** click on ---. The field changes into an input field where you can enter a maximum volume, for example 10000. Then press ENTER. Repeat the same for **Furlow**.


6.2.3.9 Under *Min.* for **Wthigh** and **Furhigh**, enter 10000.

With this setting, the query will look for all proteins that have a volume of less than 10000 in Wtlow and Furlow, and higher than 10000 in Wthigh and Furhigh. The following additional functions are possible in the *Intensity query* dialog box:

- When clicking the option *Relative to base gel*, the **<Set as base gel>** button will be activated. You can then select a gel from the list that will be regarded as the base gel. For any query that is performed on a non-base gel, the value set will be used as a multiplication factor, which will be multiplied with the spot quantity on the base gel. For example, if you enter 2 as *Min.*, a spot on a gel should be minimum two times as high as the corresponding spot on the base gel.
- You can also select the option *Fuzzy logic match* which will use a weighted approach based upon the different search criteria to determine whether a spot fulfills the criteria or not. One advantage of this method is that not all criteria should be exactly fulfilled: for example if one criterion specifies that a spot should have at least a volume of 20000, a spot with a volume of 19000 may be selected as well if other criteria are matching. Another advantage is that found spots are ranked according to the overall matching of the search criteria imposed.
- The *Intensity query* dialog box also offers the option *Combine using OR*, which, when checked, will combine the *Min.* and *Max.* criteria specified for each gel with **OR**. This means that, when this option is checked, each spot for which at least one gel has its criteria fulfilled, will be selected in the query.

6.2.3.10 Press **<OK>** to finish formatting the intensity query.

The query is now displayed in the *2D gel type* window (*Queries* tab) as a gray box listing the name, the type and the number of spots found. As more queries are generated, all of them will be listed.

6.2.3.11 With the query selected, click the  button or select *Queries > Update*.

The number of spots found of the query is displayed in its box. When the query is bordered by a red rectangle, it is the *active query*.

6.2.3.12 To make a query active, select it and press the



button.

The active query is the query that will be used in all comparison tools of InfoQuest FP. These include the spreadsheet comparisons in the *2D gel matching* window (see 6.2.4), the cluster and grouping analysis tools via a composite data set (see 6.2.6), and the advanced analysis using GeneMaths or GeneMaths XT (6.2.7).

• Significance queries

Using a significance query, spots can be selected that are significantly aberrant from the average expected value, based upon regression between pairs of gels. For a comparison of N gels, each gel is compared to each other gel by mapping all the shared spots into scatter plots and calculating a best fit regression through the plots (see also 6.2.5). This leads to $N(N-1)/2$ regressions, from which a spot is selected if it is significantly different on at least one regression.

6.2.3.13 To create a query by significance, press the



button or select *Queries > New significance*

query. Enter a name, for example, "Outliers" and press <OK>.

The *Spot significance query* dialog box (Figure 6-27) allows the choice between the *Maximum value* (highest pixel intensity), *Volume*, *Relative volume (in %)* (related to all spots on the gel as 100%) and *Quantity* (according to the Z metric in step 4 of the normalization; see 6.1.14.15).

6.2.3.14 Press the button <Add Selected> to include the selected gels in the query. Four gels should now be listed. If not, make sure the database entries are selected (colored arrow).

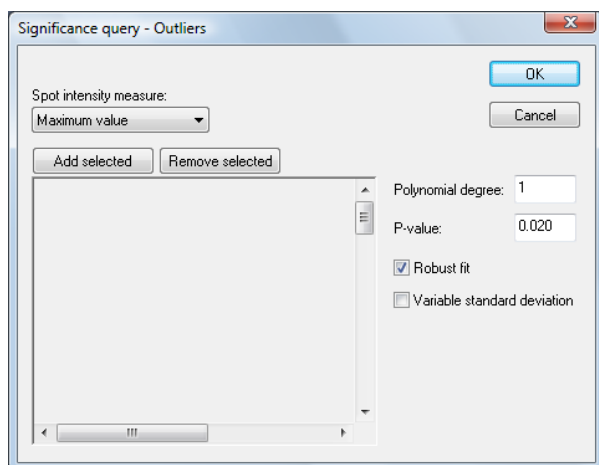


Figure 6-27. *Spot significance query* dialog box.

- With *Polynomial degree* it is possible to enter the degree of the regression; to obtain a linear regression, enter 1.


- With *Robust fit*, an iterative algorithm is applied that assigns less weight to outlier spots, hence obtaining a less distorted regression in case a few strongly outlying spots occur.

- *Variable standard deviation* is an option that calculates the standard deviation in function of the position on the regression curve and not as one single value obtained from the whole regression.

- With *P-value* you can specify the significance for a spot to be considered different. A probability is calculated for each spot to belong to the distribution, based upon the regression curve and its standard deviation limits. The program will select all spots that have a probability (p value) below the value entered in at least one regression. Such spots can be considered to be outliers. With *Variable standard deviation* enabled, the spots identified as outliers may be different from with *Variable standard deviation* disabled.

• Spot field queries

6.2.3.15 To create a query based on a spot information

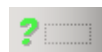
field, press the  button or select *Queries > New spot fields query*. Enter a name and press <OK>.

6.2.3.16 Under *Search in*, you can specify one of the information fields, or <All fields>. A partial search string can be entered using an asterisk (*) as wildcard.

• Manual selections

Manual selections have no search criterion associated and after creation, they are empty boxes. The purpose is to add spots manually, and to store such selections.


6.2.3.17 In case you want to define a set of spots *manually*, without the intervention of a criterion-based query, you can create a manual selection by pressing the





button or with *Queries > New manual selection*.


A manual selection contains zero spots when created. However, in the *2D gel matching* window (see 6.2.4), it is possible to add spots to the active query using the menu option *Spots > Add to active query*. With a manual selection as active query, you can create any set of manually selected spots. Such manual selections of spots are saved along with the 2D gel type.

Individual queries can be assembled into composite queries using one of the *logical operators*. The individual queries should then be considered as query components, which are together part of a composite query, combined by a logical operator.

 **AND**, combines two or more components. All conditions of the combined components should be fulfilled at the same time for a spot to be selected.


 **OR**, combines two or more components. The condition implied by at least one of the combined components should be fulfilled for a spot to be selected.

 **NOT**, operates on exactly one component. This operator inverts the argument (and hence, the selection) of the query component to which it applies.

 **XOR**, combines two or more components. Exactly one condition from the combined components should be fulfilled for a spot to be selected.

NOTE: The buttons for the logical operators contain a helpful Venn diagram icon that clearly explains the function of the operator.


As an example, we will create a composite query containing two components, combined by a logical operator.

6.2.3.18 Prepare a new query by clicking the **<New intensity query>** button () and enter a name, for example, **"Minimal expression"**. Press **<OK>**.


*NOTE: In case a query is created with one or more queries already present, a checkbox **Derive from "QueryName"** is present. "QueryName" is the name of the existing query that is selected when the new query is created. When this option is checked, the new query will be a child query of the existing one, which means that any search conditions specified will apply to the set of spots resulting from the parent query.*

6.2.3.19 Select **Volume**, press **<Add selected>** to add the selected gels, and under **Min.**, enter 20000. Press the **<Set all>** button.

6.2.3.20 By holding down the CTRL key and clicking both queries, you can select them simultaneously.

6.2.3.21 Combine the two selected queries to a more complex query with OR (). Enter a name, for example, **Diff+High**.

A new composite query appears, graphically displayed as a new box combining the two query components with connecting lines (Figure 6-28).

6.2.3.22 To run the composite query, click on its box and press the **<Run selected queries>** button () (or select **Queries > Update**).

A question pops up "This query depends on one or more other queries. Do you want to automatically update these parent queries?".

6.2.3.23 Answer **<Yes>** to update the constituent query components as well as the resulting composite query.

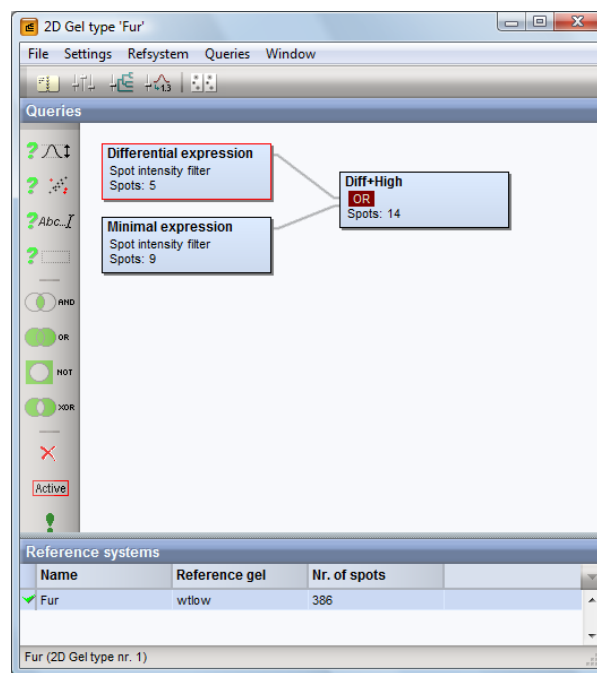




Figure 6-28. Composite spot query in the 2D gel type window.

6.2.3.24 At present, one of the query components is still the active query. To make the composite query active, select it and press the  button or select **Queries > Set as active query**.

The active query is the query that will be used in the **2D gel matching** window (see below) and in all the comparison tools of InfoQuest FP and GeneMaths/GeneMaths XT.

6.2.3.25 If more than one query exists, you can also change the active query directly in the **2D gel matching** window by selecting it in the drop-down list in the button bar (see Figure 6-23).

6.2.3.26 Editable queries (*Intensity* and *Field* queries) can be re-edited by selecting the menu option **Queries > Edit query** (or double-clicking on the query box).

6.2.3.27 Queries or query components can be deleted by selecting the component and choosing the menu option **Queries > Delete query** or pressing the  button.

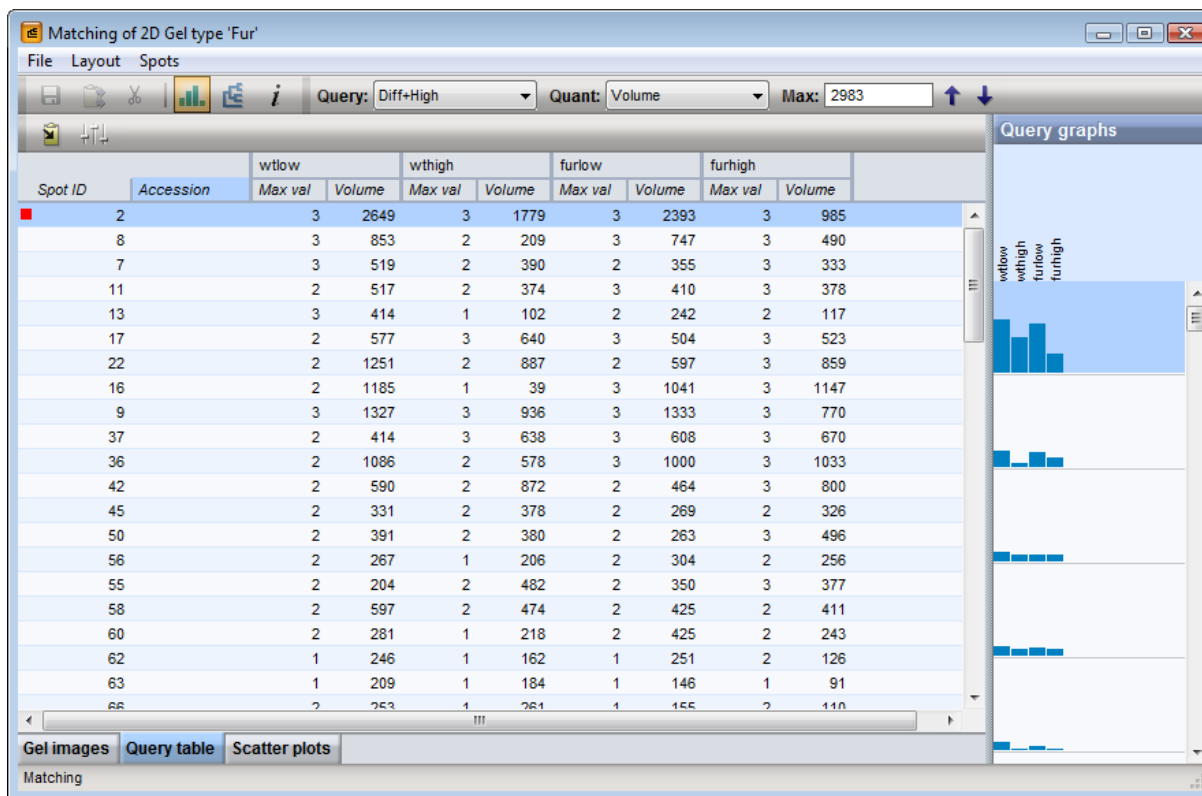




Figure 6-29. The *Query table* view of the 2D gel matching window.

6.2.4 Listing spots in tables


As illustrated above, it is possible to extract a number of proteins from a selection of 2D gels based upon specific criteria using spot queries. Such sets of spots can be retrieved and viewed in the 2D gel matching window (see also 6.2.2). Some of the actions below may already have been carried out (6.2.2.1 to 6.2.2.3).


6.2.4.1 Select the four *Campylobacter jejuni* entries in the database **Demo2D**: Wild type with low Fe concentration, Wild type with high Fe concentration, Fur mutant with low Fe concentration, Fur mutant with high Fe concentration. Use the space bar on the keyboard or click on the entries while holding the SHIFT or CTRL key.

6.2.4.2 Copy the gels to the clipboard by selecting *Edit > Copy selection* or by pressing .


6.2.4.3 In the *Experiments* panel, open 2D gel type **Fur** by double-clicking or pressing the  button.

This opens the 2D gel type window for experiment type **Fur** (Figure 6-22).

6.2.4.4 Call the 2D gel matching window by selecting *File > Create matching window* or by pressing .

6.2.4.5 Paste the gels from the clipboard in the 2D gel matching window by selecting *Edit > Paste entries from clipboard* or pressing .


The 2D gel matching window now displays the reference system and four 2D gels (Figure 6-23).

To show all gels in the normalized mode, press the  button or select *Image > Show normalized*.

6.2.4.6 Press the *Query table* tab in the bottom of the window or select *Layout > Show query table* from the menu.

6.2.4.7 As a result a table like in Figure 6-29 will be displayed.

Figure 6-29 shows a list of all protein spots that have been found. For each spot the maximum value and the volume are displayed on the four gels. These values can be used for further analysis (see below sub 6.2.6 and 6.2.7).

6.2.4.8 The layout of the *Query table* view can be modified by selecting *Layout > Spot table preferences* or press .

This will open the *The spot table preferences* dialog box displayed in Figure 6-30. By default, the Spot ID and Accession number are displayed as spot information

fields. Other fields that can be displayed include the *Description*, *Gene name(s)*, the 5 free *Comment* fields, and the metrics properties (*pI* and *MW*).

As spot quantity measures, you can display the *Maximum intensity* and *Volume* (defaults) as well as the *Area*, *Relative volume*, and *Quantity* defined by the Z metric. With a separate checkbox, the *Standard deviation* can be displayed. Standard deviations are only shown for spots that are averaged from different combined gels (see further, 6.2.8).

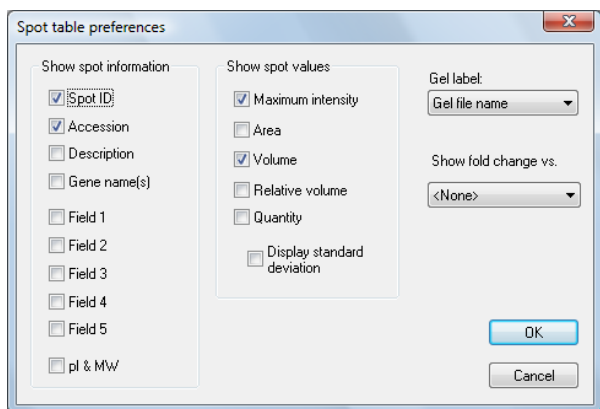



Figure 6-30. The *Spot table preferences* dialog box.

With the option *Show fold change vs.*, you can display the spot values as proportions to another gel from the selection. For example, if one gel represents a wild type and the other gels mutants, you can select the wild type under the *Fold change* option, so that all values will be shown as relative to the wild type. A value bigger than one means overexpression; smaller than one means underexpression.

Finally, you can choose one of the information fields to label the gels with the *Gel label* pull-down listbox (column headers).

NOTE: Each individual column of the table can be made wider or smaller by dragging the header separator lines to the left or to the right.

6.2.4.9 The table can be sorted according to the selected column with *Layout > Sort by column (ascending)* or *Layout > Sort by column (descending)*.

6.2.4.10 The full information stored for each spot can be viewed quickly by selecting the menu option *Layout > Show spot info* or by pressing the  button.

6.2.4.11 You can double-click on a spot to edit its information fields (6.2.2.29).

6.2.4.12 When you have chosen a specific layout of the *Query table* view, you can export the table to the clip-

board with *File > Copy to clipboard* or by pressing the




button.

The table is exported as a tab-delimited text file, which can be easily imported in other software using standard paste functions.

NOTE: Spots that were selected in the Gel images view are indicated with red squares left in the table. Spots can also be selected/unselected in the Query table view using CTRL+click or SHIFT+click.

6.2.4.13 Return to the *Gel images* view by clicking the *Gel images* tab or selecting *Layout > Show images*.

6.2.4.14 Press the  button to display a list of spot histograms right from the gel images.

On the gel images, the selected spots from the query are marked with a red rhomb.

6.2.4.15 The histograms display either the *Maximum value*, *Volume*, *Relative volume*, *Quantified value*, or *Area* of the spots selected. The type of information displayed is determined by the Spot quantification settings in the *2D gel type* window (6.2.6.5), and can also be changed from the pull-down list in the button bar of the *2D gel matching* window.

Absent proteins will be replaced by small red crosses on the histograms, indicating that the spot was not identified on that gel. Amounts that exceed beyond the maximum value that can be displayed are marked with a small horizontal line on top of the bar.

6.2.4.16 The histograms are automatically scaled to the highest value found for the selected quantification parameter. Due to individual excessive values, for example, it may be that the histograms do not fully cover the vertical range of the graphs. In that case, you can change the vertical scale using the **Up** and **Down** arrow buttons right from the *Max* indication in the button bar.

NOTE: The maximum value for the histograms can also be set in the Spot quantification settings dialog box from the 2D gel type window (6.2.6.5).

6.2.4.17 By clicking on a spot histogram in the table or the *Gel images* view, the histogram will become highlighted.

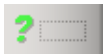
Simultaneously, the spot corresponding to the selected histogram will be marked with a small label on the individual gels where it occurs. The label displays either the maximum value, volume, relative volume, quantified value, or area of the selected spots. As noted above (6.2.2.9 and 6.2.4.15), the type of information displayed is determined by the Spot quantification settings in the *2D gel type* window (6.2.6.5), and can also be changed from the pull-down list in the button bar of the *2D gel matching* window.

Likewise, you can assign a fixed label to each spot as follows:

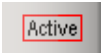
6.2.4.18 From the *Layout > Label with* menu, select the labeling method (by default *No field* is selected). Each of the spot information fields can be selected, including the free fields.


6.2.4.19 With *Layout > Label query members only* you can display labels for the spots present in the query only.

An interesting tool is the possibility to add spots manually to the active query. Likewise, it is possible to remove spots from the active query. Although any query can be edited manually, one should realize that an automatic query, based upon search criteria, will lose the information about manually added or removed spots when it is updated. Therefore, it is recommended to use the *Manual selection* query ("Spot field queries") for that purpose. This can be done as follows.

6.2.4.20 First create a manual selection in the *2D gel type* window (*Queries* tab) with *Queries > New manual selection* or the  button.

The manual selection contains zero spots when created.

6.2.4.21 Make the manual selection active by selecting it and pressing the  button (or *Queries > Set as active query*).

6.2.4.22 In the *2D gel matching* window, select some spots on a gel by dragging the mouse with the *Cursor tool* () selected. Selected spots are displayed with a red dot.

6.2.4.23 Add the selected spots to the active manual selection using the menu option *Spots > Add to active query*.

Such manual selections of spots are saved along with the 2D gel type. A manual selection can be part of a composite query, and when updated, the manual selection is preserved.


6.2.4.24 Likewise, it is possible to remove spots from an active query by selecting them in the *2D gel matching* window and choosing *Spots > Remove from active query*.


NOTE: Adding and deleting spots from the active query works for automatic and composite queries as well as for empty queries. Deletion or addition of spots is saved along with the query. In case of automatic and composite queries, however, any manual work is lost when the query is updated.

6.2.5 Comparing spots in scatter plots


As illustrated in the previous paragraphs, it is possible to extract a number of proteins from a selection of 2D gels based upon specific criteria using spot queries. Such sets of spots can be retrieved and viewed in the *2D gel matching* window (see also 6.2.2). This paragraph describes how selected spots can be compared in gel-to-gel scatter plots between any selection of gels from the database. The actions 6.2.5.1 to 6.2.5.5 below may already have been carried out (6.2.2.1 to 6.2.2.3).


6.2.5.1 Select the four *Campylobacter jejuni* entries in the database **Demo2D**: Wild type with low Fe concentration, Wild type with high Fe concentration, Fur mutant with low Fe concentration, Fur mutant with high Fe concentration. Use the space bar on the keyboard or click on the entries while holding the SHIFT or CTRL key.

6.2.5.2 Copy the gels to the clipboard by selecting *Edit > Copy selection* or by pressing .


6.2.5.3 In the *Experiments* panel, open 2D gel type **Fur** by double-clicking or pressing the  button.

This opens the *2D gel type* window for experiment type **Fur** (Figure 6-22).

6.2.5.4 Call the *2D gel matching* window by selecting *File > Create matching window* or by pressing .

6.2.5.5 Paste the gels from the clipboard in the *2D gel matching* window by selecting *Edit > Paste entries from clipboard* or pressing .

The *2D gel matching* window now displays the reference system and four 2D gels (Figure 6-23).

To show all gels in the normalized mode, press the  button or select *Image > Show normalized*.

6.2.5.6 Press the *Scatter plots* tab or select *Layout > Show scatter plots* from the menu.

As a result, a matrix of scatter plots will be displayed (Figure 6-31). Each scatter plot is the comparison between two gels selected in the *2D gel matching* window. The gel names are displayed in the row header and column header, respectively. The values for the axes are also indicated in the row and column headers.

If a spot is present in one gel and absent in another gel, it will be shown as a black dot on the scatter plot between these gels, having a zero quantity value on the gel where it is absent. If a spot is absent in two gels, it will not be shown on the scatter plot between these gels.

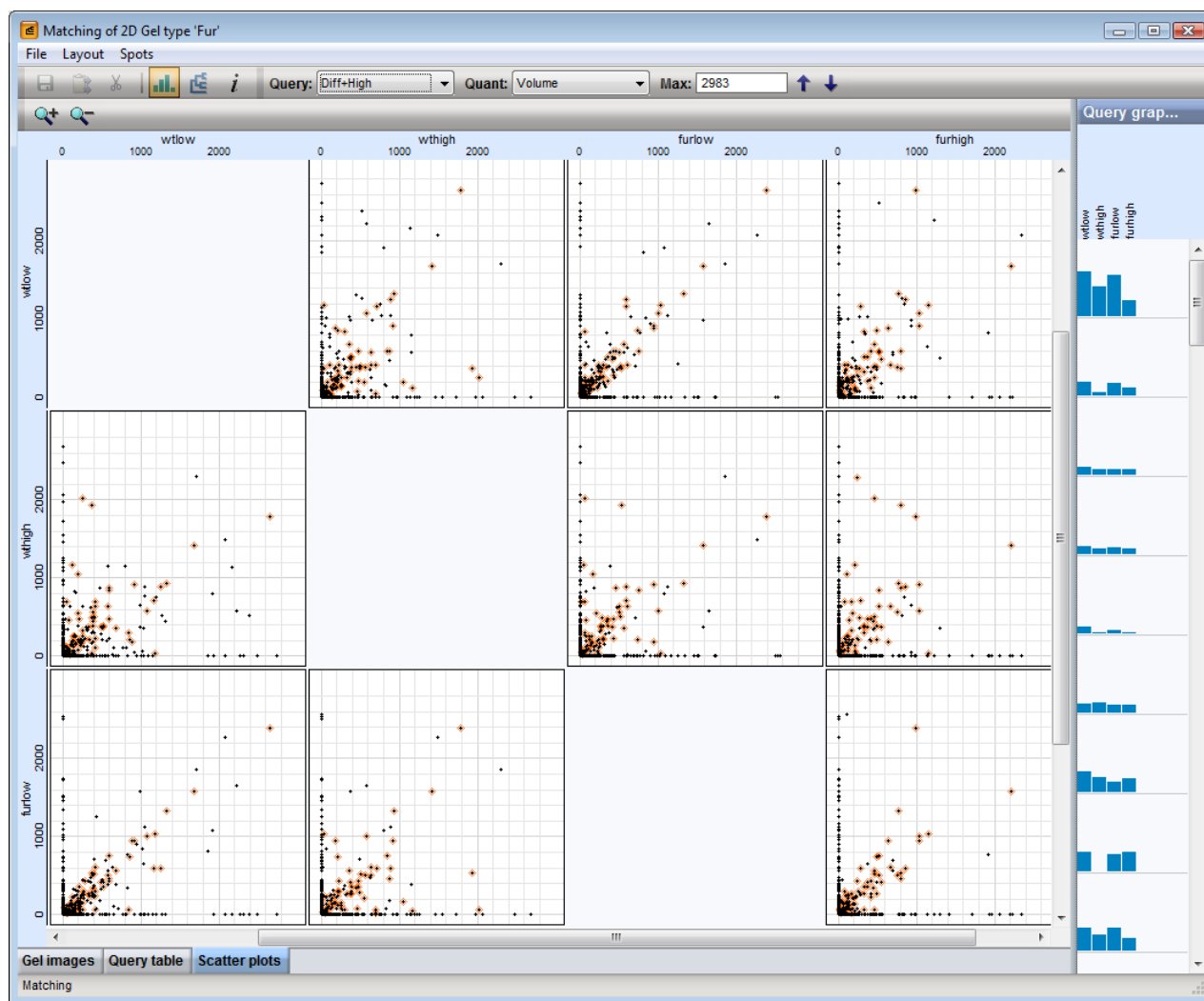


Figure 6-31. The *Scatter plots* view in the 2D gel matching window.

6.2.5.7 The values on the axes are determined by the Spot quantification settings in the 2D gel type window (6.2.6.5), and can be changed from the pull-down list in the button bar of the window.

The scatter plots are automatically scaled to the highest value found in any of the gels for the selected quantification parameter.

6.2.5.8 Due to individual excessive values, for example, it may be that the spots do not fully cover the range of the graphs. In that case, you can change the scale for the quantification parameter used by pressing the **Up** and **Down** arrow buttons right from the *Max* indication in the button bar.

The buttons available in the *Scatter plots* view are the same as those described for the *Gel images* view (6.2.2). It is possible to zoom in or out, to show histograms, to display a spot info box for the selected spot(s), and to launch GeneMaths XT for more sophisticated analysis (6.2.7).

6.2.5.9 When the histograms are shown (6.2.4.14), spots that are part of the active query are marked by a red rhomb surrounding the black dot; non-query member spots are just black dots.

6.2.5.10 To display only the spots from the active query, select *Layout > Show query spots only*.

6.2.5.11 If you click on a spot in one of the scatter plots, it will be pointed to by a red arrow on all scatter plots where the spot occurs in at least one of the two gels.

6.2.5.12 In the *Scatter plots* view, it is also possible to select one or more spots. To select one spot, simply click on it in one of the scatter plots. A selected spot is marked as a red dot, while non-selected spots are marked as black dots.

6.2.5.13 To select additional spots, hold down the CTRL key while clicking on other spots.

6.2.5.14 To select all spots in an area, you can also hold down the SHIFT key and drag a rectangle over the scatter plot.

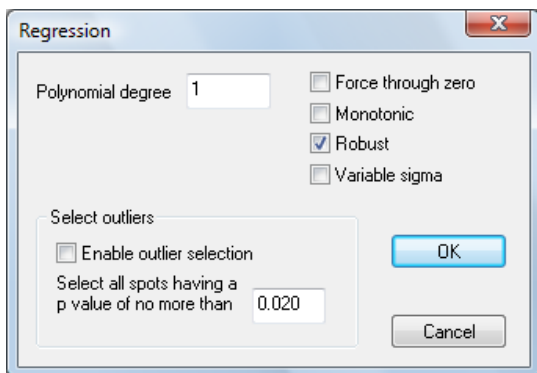


Figure 6-32. Regression dialog box for 2D gel scatter plots.

6.2.5.15 All spots (query and non-query spots) can be selected at once with *Spots > Select all spots*, whereas all spots of the active query can be selected with *Spots > Select all spots in query*.

6.2.5.16 Similar as in the *Gel images* view, you can add spots to, or delete spots from, the active query with *Spots > Add to active query* and *Spots > Remove from active query*, respectively (6.2.4.23 and 6.2.4.24).

6.2.5.17 It is also possible to change the description of a spot by double-clicking on it (6.2.2.29) or selecting *Spots > Change description* to bring up the *Spot description fields* dialog box (Figure 6-24).

6.2.5.18 You can also change a description field for a set of selected spots, as explained in 6.2.2.31.

It is possible to perform a linear or non-linear regression on the scatter plots. The program calculates the regression on the spots that are visualized: if you have chosen to show the spots from the active query only (6.2.5.10), only these spots will be taken into account for the regression calculation.

6.2.5.19 To calculate a regression on the scatter plots, select *Layout > Calculate regression lines*.

This brings up the *Regression* dialog box for 2D gel scatter plots, as shown in Figure 6-32.

- With *Polynomial degree* it is possible to enter the degree of the regression; to obtain a linear regression, enter 1.
- *Force through zero* is an option that forces the regression line to go through the origin of the scatter plot.
- *Monotonic* is an option that will force the regression to continuously increase in both the X and Y direction.
- With *Robust*, an iterative algorithm is applied that assigns less weight to outlier spots, hence obtaining a less distorted regression in case a few strongly outlying spots occur.


- *Variable sigma* is an option that calculates the sigma limits (standard deviation) in function of the position on the regression curve and not as one single value obtained from the whole regression.

- The *Select outliers* function will calculate a probability for each spot to belong to the distribution, based upon the regression curve and its sigma limits. The program will select all spots that have a probability (p value) below a certain threshold which the user can enter. Such spots can be considered to be outliers. Note that the p -values, and hence, the outliers are based upon the average of all scatter plots. Therefore, spots may be identified as "outliers", and yet seem to follow the regression closely in individual scatter plots. With *Variable sigma* enabled, the spots identified as outliers may be different from with *Variable sigma* disabled.


6.2.6 Clustering and statistical analysis of 2D gels in InfoQuest FP

The *Active query* (6.2.3.12) forms the basis for comparative analysis of 2D gel spots in the InfoQuest FP software. The result of a query is a table of spot quantities collected from a number of gels. Such a table can be visualized in the *2D gel matching* window, in the *Query table* view (Figure 6-29), but it can also be treated as a character table to perform cluster analysis, principal components analysis and all derived techniques available in InfoQuest FP.

6.2.6.1 Select the four *Campylobacter jejuni* entries in the database **Demo2D**: Wild type with low Fe concentration, Wild type with high Fe concentration, Fur mutant with low Fe concentration, Fur mutant with high Fe concentration. Use the space bar on the keyboard or click on the entries while holding the SHIFT or CTRL key. Selected entries are marked with a colored arrow.

6.2.6.2 Create a new comparison with *Comparison > Create new comparison ()* or by pressing the  button in the *Comparisons* panel.

The 2D gel type **Fur** is the only experiment type listed in the *Experiments* panel of the *Comparison* window.

6.2.6.3 Show the character table of the spot query by pressing the  button of **Fur** in the *Experiments* panel.

The spot intensities are now displayed as differentially shaded gray blocks. The spot ID numbers are indicated in the column header.


6.2.6.4 Select *Clustering > Calculate > Cluster analysis (similarity matrix)* to calculate a dendrogram.

The *Comparison settings* dialog box allows you to specify the similarity coefficient to calculate the similarity matrix, and the clustering method. Cluster analysis of

2D gel spot query tables is identical as clustering character-based data (see Section 4.3). All the clustering and statistical functions that apply to character data also apply to 2D gel data.

The spot quantities used to construct the character table in the *Comparison* window are those chosen in the *Spot quantification settings* in the *2D gel type* window.

6.2.6.5 To use another quantification measure, close the *Comparison* window, open the *2D gel type* window for

Fur, and press  or select **Settings > Spot quantification settings**.

In the *Spot quantification* dialog box (Figure 6-33) you can choose between the *Maximum value* (in pixel intensity), the *Spot area* (number of pixels included), the *Spot volume* (sum of pixel intensities), the *Relative volume*, and the *Quantified value* (derived from the Z metric).

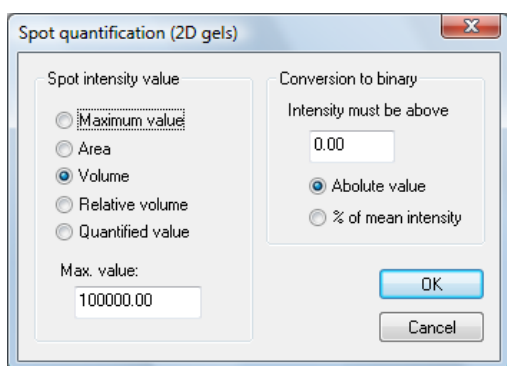


Figure 6-33. The *spot quantification settings* dialog box.

Under *Max. value*, one can enter a value which is important for the bar graph diagrams in the *2D gel matching* window and its *Query table* view: the bar graphs will be scaled to that maximum value entered. For example, in case *Volume* is specified under *Spot intensity value*, the value to enter under *Max. value* depends on the highest volume found in the gels that are being compared. In case *Maximum value* is specified under *Spot intensity value*, the value to enter depends on the OD range of the image, which can for example be 8 bit (256), 12 bit (4,096) or 16 bit (65,536).

NOTE: The *Spot intensity value* as well as the *Max. value* can also be chosen directly from the *2D gel matching window* (*Gel images and Scatter plots tabs*).

Under *Conversion to binary*, you can specify what threshold to use when applying a binary coefficient to the analysis of 2D gel character tables. You can specify a minimum *Absolute value* or a percentage of the *Mean intensity*.

In order for the spot quantification settings to become effective, you will need to re-open the *Comparison* window.

To add some more flexibility to the data set, 2D gel spot tables can also be analyzed as a *composite data set*. The advantage of analyzing 2D spot tables as *composite data sets* is that both the columns and the rows can be clustered (*transversal* clustering or *two-way* clustering) to get a better understanding of the relation experiments versus characters. See Section 4.7.2 to use a 2D gel type in a *composite data set*.

6.2.7 Analyzing 2D gel spot tables with GeneMaths XT

GeneMaths XT (or its predecessor GeneMaths) offers some more advanced statistical tools for the analysis of large data sets and is particularly suited for the analysis of microarrays and gene chips. Since the data generated and the purpose of 2D gels and microarrays is quite similar, GeneMaths XT is also very useful for the analysis of 2D gel spot tables comprising various experiments. Since GeneMaths XT is integrated with the InfoQuest FP software, analysis of 2D gel spot tables in GeneMaths XT is very straightforward.

The *Active query* (6.2.3.12) forms the basis for comparative analysis of 2D gel spots in GeneMaths XT. The result of a query is a table of spot quantities over a number of gels. Such a table can be visualized in the *2D gel matching* window, in the *Query table* view (Figure 6-29), but it can also be directly imported as a character table to perform cluster analysis, principal components analysis and all derived techniques available in GeneMaths XT.

6.2.7.1 Open a table with a large query as explained in 6.2.4.

6.2.7.2 Both from the table and from the *Gel images* view, you can select **File > Statistical analysis** or press the



button.

The GeneMaths XT analysis window will open with the protein spots as rows and the experiments (gels) as columns (Figure 6-34). All the information fields for the spots are displayed, whereas for the experiments, you can choose between the key or one of the information fields defined for the entries in the InfoQuest FP database.

Designed for the exploration and analysis of large data sets such as microarrays, the GeneMaths XT software package is the ideal tool for comparative analysis of sets of 2D protein gels as well. In addition, through its integration with InfoQuest FP, it can be used to compare microarray data with 2D protein gel data. The following main functions can be applied to 2D protein gels:

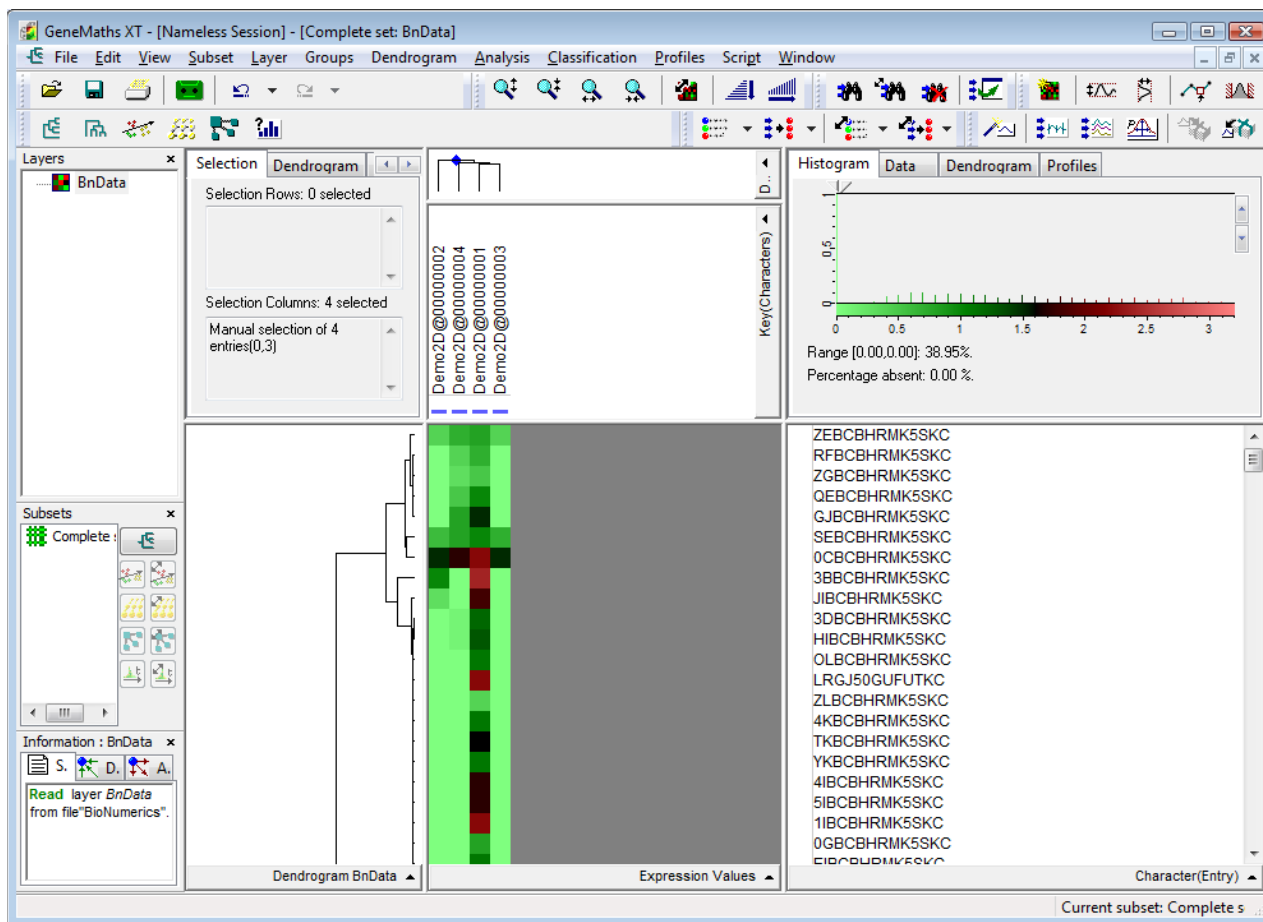


Figure 6-34. Analysis of 2D gel protein spots over different experiments in GeneMaths XT.

- Standardization of data matrix using offset and scaling functions (arithmetic and Median averages, Root Mean Square, Standard Deviation);
- Transformation of data matrix (flipping);
- Transversal cluster analysis of rows and columns using a variety of similarity/distance coefficients, and several pair-group clustering methods, Ward, and Neighbor Joining;
- Cluster significance indication based on bootstrap techniques;
- Special dendrogram layout and visualization tools to facilitate the interpretation of analyses of extreme sizes;
- Pattern matching: search for closest matches with specific profiles, average profiles or theoretical profiles;
- Single, composite or average profile curve or bar graph plotting with indication of standard deviations;
- X-Y plots, time course plots, and scatterplots;
- 2-D and 3-D Principal Component Analysis and Discriminant Analysis with or without variance;
- Self-Organizing maps (Kohonen maps).
- Etc.

6.2.8 Editing reference systems

For a 2D gel type as well as for 1-D fingerprint types, it is possible to create more than one reference system. This possibility is useful to combine gels with different properties in the same experiment type. For example, it is possible to run different 2D gels for the same sample, each having their own pH ranges, and merge such gels with different pH ranges into multiple gels, spanning the full pH range.

A new reference system can be created in the *2D gel processing* window (6.1.13) in the normalization step. Adding *reference spots* to the reference system can also be done during the normalization step. However, for complete editing functionality of the reference system, the *2D gel type* window should be used (6.2.2.3).

6.2.8.1 In database **Demo2D**, open **Fur** in the *Experiments* panel.

The *2D gel type* window (Figure 6-22) contains two tabs: *Reference systems* and *Spot queries*.

6.2.8.2 Select the *Reference systems* tab to display the reference systems present for the 2D gel type **Fur**. Normally only **Fur** should be listed as a reference system.

As soon as more than one reference system exists within a 2D gel type, it should also be possible to select an *active reference system*, i.e. the reference system used for comparisons in the *2D gel matching* window.

6.2.8.3 You can change the active reference system in the *2D gel type* window by selecting the *Reference systems* tab and choosing *Refsystem > Set as active reference system*.

6.2.8.4 To reset a reference system so that all spots are removed from it, you can select *Refsystem > Remove all spots*.


6.2.8.5 To update a reference system according to recent editing work done, select *Refsystem > Refresh spots*.


6.2.8.6 To delete a reference system, select *Refsystem > Delete*.


NOTE: The active reference system cannot be deleted.

Individual spots can also be added or deleted from a reference system in the *2D gel matching* window. To that end, we will need some additional gels to be added to the *2D gel matching* window.


6.2.8.7 Select the four *Campylobacter jejuni* entries in the database **Demo2D**: Wild type with low Fe concentration, Wild type with high Fe concentration, Fur mutant with low Fe concentration, Fur mutant with high Fe concentration. Use the space bar on the keyboard or click on the entries while holding the SHIFT or CTRL key.


6.2.8.8 Copy the gels to the clipboard by selecting *Edit > Copy selection* or by pressing  .

6.2.8.9 In the *2D gel type* window, call the *2D gel matching* window by selecting *File > Create matching window* or by pressing  .

6.2.8.10 Paste the gels from the clipboard in the *2D gel matching* window by selecting *Edit > Paste entries from clipboard* or pressing  .

The *2D gel matching* window now displays the reference system and four 2D gels (Figure 6-23).

6.2.8.11 To show all gels in the normalized mode, press the  button or select *Image > Show normalized*.

6.2.8.12 Select one or more spots on any gel in the window by dragging the mouse pointer over the spot(s) in cursor tool mode ().

6.2.8.13 Select *Refsystem > Add selected spot(s)*. The program asks to confirm to add the selected spots to the reference system.

If you answer **<Yes>**, the selected spots become reference spots in the reference system.


Likewise, you can select spots on the reference system and delete them with *Refsystem > Delete spot(s)*. A confirmation is requested.


6.2.9 Creating synthetic gels


The main purposes of synthetic gels are: (1) averaging repeats of the same experiment to obtain higher accuracy; (2) combining images of the same gel with different exposure times to reveal very weak spots as well as very dark spots (e.g. when autoradiography is used); and (3) combining gels of the same sample with different pH ranges.

Although there is no suitable example in the **Demo2D** database, we can use the available gels to explain this function.


6.2.9.1 Select the two *Campylobacter jejuni* entries in the database **Demo2D**: Wild type with low Fe concentration and Wild type with high Fe concentration. Use the space bar on the keyboard or click on the entries while holding the SHIFT or CTRL key.

6.2.9.2 Copy the gels to the clipboard by selecting *Edit > Copy selection* or by pressing  .

6.2.9.3 In the *2D gel type* window, call the *2D gel matching* window by selecting *File > Create matching window* or by pressing  .

6.2.9.4 Paste the two gels from the clipboard in the *2D gel matching* window by selecting *Edit > Paste entries from clipboard* or pressing  .

The *2D gel matching* window now displays the reference system and two 2D gels (Figure 6-23).

6.2.9.5 Show all gels in the normalized mode by pressing the  button or selecting *Image > Show normalized*.

6.2.9.6 Select *File > Create synthetic gel*.

This will open the *Create synthetic gel* window as displayed in Figure 6-35. The panel on the left shows the shifts towards the averaged spots.

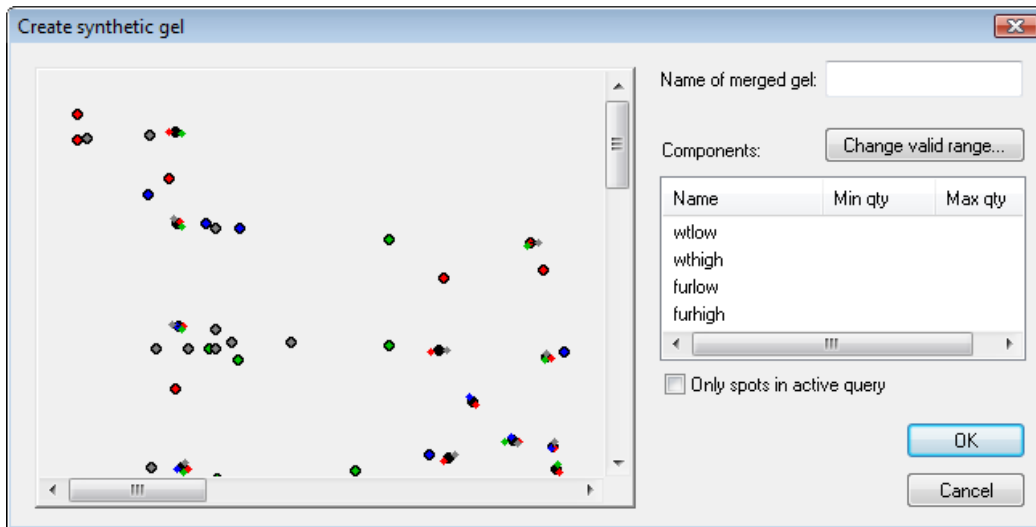


Figure 6-35. The *Create synthetic gel* window.

6.2.9.7 Under *Name of the merged gel*, type a name, for example **Wt**.

In case you are averaging gels with different exposures to obtain a higher dynamic range for the spots, you can then define for each gel the range of spot quantities (Z metric as defined in the Metrics step) to be included in the synthetic gel.

6.2.9.8 To define a range for a gel, select the gel under *Name*, and press **<Change valid range>**.

6.2.9.9 Enter a minimum and/or a maximum spot quantity.

6.2.9.10 With **<Only spots in active query>** checked, the merged gel will only contain the spots of the current active query.

7. APPENDIX

7.1 Connected database table structure

7.1.1 Introduction

In the description below, the structures of the tables required by InfoQuest FP 5.1 in a Connected database are given (see Section 2.3). The tables are indicated with their default names. However, it is possible to use different names for these tables or views in an actual database, which are recorded in the connected database configuration file (.xdb). The names of the columns within the tables, however, are fixed.

The object “CLOB” means a large text field. This may be described differently depending on the database use (e.g., the Access equivalent is “memo”). NULL values should be allowed for all fields.

7.1.2 Table ALIGNPROJ

Contains a record for every alignment project created in the database.

- OBICNAME (VARCHAR(80))

Name of the alignment project.

- OBICDATECREATED (VARCHAR(80))

The date the alignment project was created.

- OBICDATEMODIFIED (VARCHAR(80))

The date of the last modifications made to the alignment project.

- OBICDATA (CLOB)

Alignment data.

Other fields: additional alignment information fields.

7.1.3 Table ATTACHMENTS

Contains a record for every attachment present in the database.

- KEY (VARCHAR(80))

Key of the entry the attachment belongs to.

- IDN (VARCHAR(10))

Identifier attachment.

- CLASS (VARCHAR(20))

The data type of the attachment (1 = Text file, 2 = Bitmap image, 3 = HTML document, 4 = Word document, 5 = Excel document, 6 = PDF document).

- DESCRIPT (VARCHAR(80))

The description of the attachment.

- FILENAME (VARCHAR(250))

The path where the file is stored.

- CONTENT (CLOB)

The content of a text file.

7.1.4 Character Values table

Each character type has its own table holding character value information for the database entries. The default name of this table is the name of the character type, although it is possible to specify any table name (the exact name is contained in the TABLES column of the EXPERIMENTS table). Each record in the table corresponds to a single character value belonging to a single entry in the database.

- KEY (VARCHAR(80))

Key of the entry this character value belongs to.

- CHARACTER (VARCHAR(80))

Key of the character.

- VALUE (FLOAT)

Numerical value.

7.1.5 Character Fields table

Contains information about the additional information fields that can be stored together with characters in a character type. The default table name is the name of the character type, padded with “FIELDS”, but it is possible to specify any other name (the exact name is contained in the TABLES column of the EXPERIMENTS table). Every record in this table corresponds to a single field for a single character.

- CHARACTER (VARCHAR(80))

Name of the character this information field belongs to.

- FIELD (VARCHAR(80))

Name of the field.

- CONTENT (VARCHAR(150))

Content of the field.

7.1.6 Table COMPARISONS

The table contains information about all comparisons saved in the database.

- NAME (VARCHAR(200))

The name of the comparison.

- CMPTPE (VARCHAR(80))

The type of comparison (comparison or library).

- CMPCLS (VARCHAR(200))

The comparison class.

- CMPOWNER (VARCHAR(80))

The comparison 'owner'.

- CMPCREATED (VARCHAR(80))

The date the comparison was created.

- CMPCMODIFIED (VARCHAR(80))

The date of the last modifications made to the comparison.

- CMPCDATA (VARCHAR(CLOB))

Comparison data.

7.1.7 Table DBSCHEMAS

This table holds information about the software and the table structure of some installed plugins.

- NAME (VARCHAR(80))

Name of the software and the installed plugins for which new tables were added to the table structure.

- SCHVERSION (VARCHAR(80))

Version number of the software and the installed plugins for which new tables were added to the table structure.

- SCHDEF (CLOB)

XML information on the installed plugins.

7.1.8 Table DBSETTINGS

This table holds the installed plugins and the active information fields of the *Database* panel.

- NAME (VARCHAR(200))

'ActivePlugins', 'DEFAULTLEVELSETTINGS'.

- CONTENT (CLOB)

The string of the 'ActivePlugins' holds the installed plugins, the 'DEFAULTLEVELSETTINGS' holds the active information fields of the *Database* panel.

7.1.9 Table DECISNTW

The table contains information about all decision networks made in the database.

- OBJCNAME (VARCHAR(80))

The name of the decision network.

- OBJCDATECREATED (VARCHAR(80))

The date the decision network was created.

- OBJCDATEMODIFIED (VARCHAR(80))

The date of the last modifications made to the decision network.

- OBJCDATA (CLOB)

String holding the structure of the decision network.

Other fields: additional decision network information fields.

7.1.10 Table ENLEVELS

Contains information about the levels defined in the database.

- LEVELID (NUMBER)

Holds the levelID of the defined levels.

- LEVELNAME (VARCHAR(80))

Name of the level.

- SETTINGS (CLOB)

String that holds the active fields of each level.

7.1.11 Table ENRELATIONS

This table contains all entries belonging to a relation.

- RLID (NUMBER)

The unique ID for each defined relation in the database.

- RELTYPEID (NUMBER)

The identifier for each defined relation type.

- KEY1 (VARCHAR(80))

The key of the entry belonging to the forward relation.

- KEY2 (VARCHAR(80))

The key of the entry belonging to the reverse relation.

7.1.12 Table ENRELATIONTYPES

Contains information about the defined relation types.

- RELATID (NUMBER)

The unique identifiers for each defined relation type.

- RELATFORWNAME (VARCHAR(200))

The name of the forward relation.

- RELATBACKNAME (VARCHAR(200))

Name of the reverse relation.

- LEVELID1 (NUMBER)

The levelID of the forward relation.

- LEVELID2 (NUMBER)

The levelID of the reverse relation.

- RELTYPE1 (NUMBER)

The type of forward relation: many = 0; one = 1.

- RELTYPE2 (NUMBER)

The type of reverse relation: many = 0; one = 1.

7.1.13 Table ENTRYTABLE

This table contains a record for every entry in the database.

- KEY (VARCHAR(80))

The unique identifier for every entry in the database (e.g. isolate number).

- LEVELID (NUMBER)

The levelID for each entry in the database.

Other fields: additional database information fields.

7.1.14 Table EVENTLOG

This table maintains a history list of events that were generated during the manipulation of the database.

- DATETIME (VARCHAR(80))

Recording date and time of the event.

- LOGIN (VARCHAR(50))

Windows login at the moment the event was generated.

- TYPE (VARCHAR(10))

Event type.

- SUBJECT (VARCHAR(50))

Database component for which this event was generated.

- DESCRIPTION (VARCHAR(500))

Description of the event.

7.1.15 Table EXPERATTACH

This table contains descriptive information for any specific key-experiment combination. For example, the error reports generated in the Spa, MLST and batch sequence assembly plugin are stored in this table.

- EXPRATTACHID (NUMBER)

The unique identifier for each key-experiment combination.

- KEY (VARCHAR(80))

Key of the database entry the information relates to.

- EXPERIMENT (VARCHAR(80))

Name of the experiment the information relates to.

- NAME (VARCHAR(80))

Names assigned to groups of key-experiment combinations.

- CONTENT (CLOB)

Descriptive information specific for each key-experiment combination, e.g. error report.

7.1.16 Table EXPERIMENTS

This table contains a record for every experiment type present in the database.

- EXPERIMENT (VARCHAR(80))

Holds the name of the experiment (should be unique through the whole database).

- TYPE (VARCHAR(80))

Can be "Fingerprint", "Character", "Sequence", "Matrix", "Curve" or "2DGel".

- SETTINGS (CLOB)

XML string that holds the processing, visualization and analysis settings of the experiment type.

- TABLES (VARCHAR(160))

Used for character experiments only: holds the name of the tables that hold character values and additional character fields (separated by a comma).

Other fields: additional experiment type information fields.

7.1.17 Table FPRBNDCLS

This table contains a record for each band class defined in the database.

- CLSID (NUMBER)

The unique identifier for each band class defined in the database.

- CLSEXPER (VARCHAR(80))

Holds the name of the experiment type.

- CLSNAME (VARCHAR(80))

Name of the band class.

- CLSPOSIT (FLOAT)

The position (metrics) of each band class.

7.1.18 Table FPRINT

This table contains a record for every fingerprint that is entered in the database.

- KEY (VARCHAR(80))

The unique identification key of the sample to which this fingerprint belongs.

- EXPERIMENT (VARCHAR(80))

The name of the experiment type to which this fingerprint belongs.

- FILENAME (VARCHAR(80))

The name of the batch to which this fingerprint belongs.

- FILEIDX (NUMBER)

The number of the fingerprint inside the fingerprint file.

- SPLINE (VARCHAR(200))

Holds the exact positioning and size of the gelstrip on the image.

- CURVESPLINE (VARCHAR(200))

Describes what part of the gelstrip is used for calculation of the densitometric curve.

- GELSTRIPINFO (VARCHAR(50))

Contains resolution information about the gelstrip image info.

- GELSTRIP (CLOB)

This field holds the bitmap values of the gelstrip.

- DENSCURVEINFO (VARCHAR(50))

Holds the resolution of the densitometric curve.

- DENSCURVE (CLOB)

Holds the densitometric curve data.

- BANDS (CLOB)

Holds information about the bands assigned on the fingerprint.

- BANDCONC (CLOB)

Holds information about 2D concentration estimates.

- BANDCONCINFO (CLOB)

Holds information about 2D concentration estimates.

- REFPOS (VARCHAR(250))

Contains the reference positions assigned to this fingerprint.

- MAPFORWARD (CLOB)

Contains a forward normalization vector.

- MAPBACK (CLOB)

Contains the reverse normalization vector.

- REFSYSTEM (CLOB)

Holds the reference system of the fingerprint.

- **TONECURVE** (VARCHAR(250))

Contains the tone curve.

- **CHPTRN** (VARCHAR(250)) (only with “Fast band matching” enabled)

Contains cached pattern information on the band positions for a fingerprint type with “Fast band matching” enabled.

Other information fields: additional information fields added in the *Fingerprint information* panel in the *Fingerprint file* window.

7.1.19 Table FPRINTFILES

This table contains a record for every “batch” of fingerprints that is entered in the database. A batch may correspond to fingerprints that should be normalized simultaneously: e.g. they were run on the same electrophoresis gel, or run in the same batch on a sequencer, etc.

- **FILENAME** (VARCHAR(80))

The name of the batch (should be unique for every batch). In case of scanned electrophoresis gels, this corresponds to the name of the TIFF image file.

- **EXPERIMENT** (VARCHAR(80))

Name of the experiment type to which this fingerprint batch belongs.

- **LOCKED** (VARCHAR(10))

Whether or not this batch is locked (Yes or No).

- **INLINELINK** (VARCHAR(80))

If this batch is linked to another batch (for normalization purposes), this specifies the name of the batch that contains normalization info.

- **BOUNDINGBOX** (VARCHAR(200))

Specifies the bounding box of the lanes on a 2D fingerprint image.

- **SETTINGS** (VARCHAR(250))

Data processing settings.

- **TONECURVE** (VARCHAR(200))

Specifies how bitmap pixel values are mapped to grey shades on the screen.

- **REFSYSTEM** (CLOB)

Specifies the reference system that is used to normalize the batch.

- **MARKERS** (VARCHAR(200))

Holds marker points that may be used to align linked fingerprint images to each other.

Other information fields: additional Fingerprint file information fields.

7.1.20 Table MATRIXVALS

Holds pairwise similarity values. Each record in this table represents a single similarity value between two database entries.

- **EXPERIMENT** (VARCHAR(80)).

Name of the experiment type this similarity value belongs to.

- **KEY1** (VARCHAR(80))

Key of the first database entry.

- **KEY2** (VARCHAR(80))

Key of the second database entry.

- **VALUE** (FLOAT)

Similarity value.

7.1.21 Table SEQTRACEFILES

This table holds information about the sequence trace files (four-channel chromatogram files from automated sequencers).

- **KEY** (VARCHAR(80))

For use with the Kodon software.

- **CONTIGFILE** (VARCHAR(80))

Unique ID of the contig that is associated to this sequence trace file.

- **TRACEID** (VARCHAR(80))

Unique ID of the trace file.

- **DATA** (CLOB)

Holds the full trace information including sequence and the chromatogram files in case the trace files are stored in the database. Otherwise, it stores a link to the path of the trace file.

- **INFO** (CLOB)

Contains the full editing information of the sequence trace file.

7.1.22 Table SEQUENCES

This table holds the sequence information stored in the database. Note that the columns designed for contig files have changed with respect to earlier versions of the software.

- KEY (VARCHAR(80))

Key of the database entry this sequence belongs to.

- EXPERIMENT (CHARCHAR(80))

Experiment type of the sequence.

- SEQUENCE (CLOB)

Sequence data.

- SEQUENCEQUAL

Quality coefficient for each base in the sequence.

- CONTIGFILE (VARCHAR(80))

Unique ID of the contig file that is associated to this sequence (if any).

- CONTIG (CLOB)

Holds the contig sequence and its full editing history.

- CONTIGSTATUS (VARCHAR(10))

Contains the status of the contig file, i.e. confirmed or not.

7.1.23 Table SUBSETMEMBERS

This table contains information about the subsets that were defined in the database. Each record specifies the membership of a single entry to a single subset.

- KEY (VARCHAR(80))

The key of the database entry.

- SUBSET (VARCHAR(80))

The name of the subset to which this key belongs.

7.1.24 Table TRENDATA

Holds information about the trend data types.

- KEY (VARCHAR(80))

The key of the database entry.

- EXPERIMENT (VARCHAR(80))

Name of the trend data type.

- CURVE (VARCHAR (80))

Name of the trend curve.

- DATA (CLOB)

XML string that holds the data.

- PARAMS (CLOB)

Lists the parameter(s) defined for the trend data type.

7.1.25 Indices in the database

In order to obtain sufficient speed for larger databases, it is absolutely necessary that a number of indices are present. This section contains a list of advised indices. However, depending on the purpose of the database (emphasis on read or write, database size...), it may be preferable to modify, add or remove indices. For larger databases where speed becomes critical, it is strongly advised to use the tuning tools provided with the database in order to optimize the various settings and indices.

- ENTRYTABLE:

KEY (may be defined as primary key).

- EXPERIMENTS:

EXPERIMENT (may be defined as primary key). This usually won't attribute to the performance, since the number of records in this table is usually very limited.

- FPRINTFILES:

FILENAME (may be defined as primary key).

- FPRINT:

KEY. It should not be unique or primary key, since some lanes on a gel image may not be added to the database and will have an empty key (e.g. reference lanes).

FILENAME. Note that this field should not be required, because some databases may contain fingerprints that are not associated with any batch (file).

FILENAME,FILEIDX.

- Character values table:

CHARACTER.

KEY.

- Character fields table:

CHARACTER,FIELD.

•SEQUENCES:

KEY.

•MATRIXVALS:

EXPERIMENT,KEY1,KEY2.

•SUBSETMEMBERS:

KEY.

SUBSET.

7.2 Regular expressions

A "regular expression" is a pattern that describes a set of strings. Regular expressions are constructed analogously to arithmetic expressions, by using various operators to combine smaller expressions. `grep` understands two different versions of regular expression syntax: "basic" and "extended". In GNU `grep`, there is no difference in available functionality using either syntax. In other implementations, basic regular expressions are less powerful. The following description applies to extended regular expressions; differences for basic regular expressions are summarized afterwards.

The fundamental building blocks are the regular expressions that match a single character. Most characters, including all letters and digits, are regular expressions that match themselves. Any metacharacter with special meaning may be quoted by preceding it with a backslash. A list of characters enclosed by '[' and ']' matches any single character in that list; if the first character of the list is the caret '^', then it matches any character *not* in the list. For example, the regular expression `[0123456789]` matches any single digit. A range of ASCII characters may be specified by giving the first and last characters, separated by a hyphen.

Finally, certain named classes of characters are predefined, as follows. Their interpretation depends on the `LC_CTYPE` locale; the interpretation below is that of the POSIX locale, which is the default if no `LC_CTYPE` locale is specified.

`[:alnum:]`

Any of `[:digit:]` or `[:alpha:]`

`[:alpha:]`

Any letter:

`abcdefghijklmnopqrstuvwxyz,`

`ABCDEFGHIJKLMNOPQRSTUVWXYZ.`

`[:blank:]`

Space or tab.

`[:cntrl:]`

Any character with octal codes 000 through 037, or `DEL` (octal code 177).

`[:digit:]`

Any one of `0123456789`.

`[:graph:]`

Anything that is not a `[:alnum:]` or `[:punct:]`.

`[:lower:]`

Any one of `abcdefghijklmnopqrstuvwxyz`.

`[:print:]`

Any character from the `[:space:]` class, and any character that is *not* in the `[:graph:]` class.

`[:punct:]`

Any one of `!"#$%&'()*+,-./:;<=>@[\] ^ _ { | } ~`.

`[:space:]`

Any one of `\CR FF HT NL VT SPACE`'.

`[:upper:]`

Any one of `ABCDEFGHIJKLMNOPQRSTUVWXYZ VWXYZ`'.

`[:xdigit:]`

Any one of `abcdefghijklmnopqrstuvwxyz0123456789`'.

For example, `[:alnum:]` means `[0-9A-Za-z]`, except the latter form is dependent upon the ASCII character encoding, whereas the former is portable. (Note that the brackets in these class names are part of the symbolic names, and must be included in addition to the brackets delimiting the bracket list.) Most metacharacters lose their special meaning inside lists. To include a literal `]`, place it first in the list. Similarly, to include a literal `^`, place it anywhere but first. Finally, to include a literal `-`, place it last.

The period `.` matches any single character. The symbol `\w` is a synonym for `[:alnum:]` and `\W` is a synonym for `^[[:alnum:]]`.

The caret `^` and the dollar sign `$` are metacharacters that respectively match the empty string at the beginning and end of a line. The symbols `\<` and `\>` respectively match the empty string at the beginning and end of a word. The symbol `\b` matches the empty string at the edge of a word, and `\B` matches the empty string provided it's not at the edge of a word.

A regular expression may be followed by one of several repetition operators:

``?'`

The preceding item is optional and will be matched at most once.

``*'`

The preceding item will be matched zero or more times.

``+'`

The preceding item will be matched one or more times.

``{N}'`

The preceding item is matched exactly N times.

``{N,}'`

The preceding item is matched n or more times.

``{N,M}'`

The preceding item is matched at least N times, but not more than M times.

Two regular expressions may be concatenated; the resulting regular expression matches any string formed by concatenating two substrings that respectively match the concatenated subexpressions.

Two regular expressions may be joined by the infix operator ``|'`; the resulting regular expression matches any string matching either subexpression.

Repetition takes precedence over concatenation, which in turn takes precedence over alternation. A whole subexpression may be enclosed in parentheses to override these precedence rules.

The backreference ``\N'`, where N is a single digit, matches the substring previously matched by the Nth parenthesized subexpression of the regular expression.

Index

Symbols

.dbs file 11, 39

Numerics

2D background subtraction 338
 2D gel image settings dialog box 343
 2D gel types 9, 89, 335, 338
 2D gel types module 11, 17, 335
 2D gels plugin 27
 3D view of a gel area 95, 342

A

Absolute value 372
 Add array of characters 122
 Add client computer (Netkey) 20
 Add new calibration rectangle tool 347
 Add new entries 43
 Add new experiment file 136
 Add new reference system 348
 Add spot tool 345
 Advanced query tool 50
 Align external branch 219
 Align internal branch 219
 Alignment > Bookmarks > Add bookmark 243
 Alignment > Bookmarks > Add new list 243
 Alignment > Bookmarks > Delete selected bookmarks 243
 Alignment > Calculate > Multiple alignment 230
 Alignment > Calculate > Pairwise alignment 233
 Alignment > Consensus > Create from selected entries 233
 Alignment > Consensus > Recalculate 233
 Alignment > Create copy 228
 Alignment > Identity display 234
 Alignment > Load curves 233
 Alignment > Reset 232
 Amino acid sequences 135
 Analysis 25, 27, 29, 36, 43
 Analyze program 9
 Antibiotics susceptibility plugin 27
 Apply to tone curve 347
 Area sensitive (coefficient) 191
 Arithmetic average 100
 Arrange by similarity 301
 Arrange entries by field 49
 Assembler 136–149
 Assign to reference system 348
 Assignment of metrics 339
 Attachments 45–46, 50, 52
 Auto construct tables 60

Automatic spot search 343, 344
 Average (K-means) 181
 Average similarities (jackknife) 182
 Averaging thickness (curves) 99, 116

B

Background color 342
 Background subtraction 91, 97, 100, 338, 342
 Background subtraction (BNIMA) 127
 Backup 85–86
 Ball size 127
 Band class filters 202
 Band classes > Add new band class 200
 Band classes > Auto assign bands to class 201
 Band classes > Center class position 201
 Band classes > Remove band class 200, 201
 Band finding (settings) 105
 Band height 203
 Band matching 197
 Band search filters 105
 Band search, shoulder sensitivity 105
 Band surface 203
 Bandmatching > Auto assign all bands to all classes 204
 Bandmatching > Band class filter 202
 Bandmatching > Comparative Quantification settings 203
 Bandmatching > Export band matching 203
 Bandmatching > Perform band matching 197, 203, 204
 Bandmatching > Polymorphic bands only (for selection list) 204
 Bandmatching > Save band classes to experiment type 201
 Bandmatching > Search band classes 201
 Bands 93, 107, 193
 Bands (assigning) 105
 Bands > Add new band 106
 Bands > Auto search bands 106
 Bands > Delete selected band(s) 107
 Bands > Mark band(s) as certain 107
 Bands > Mark band(s) as uncertain 107
 Batch sequence assembly plugin 28, 137, 225
 Bifurcation (in a decision tree) 329
 Binary coefficient 207, 248, 264
 InfoQuest FP 2D. See 2D gel types module
 InfoQuest FP Help plugin 28
 Bitmap export 185
 BLAST sequence matching 308–312
 BLOSUM matrices 309
 BNIMA (character import tool) 126
 Bookmarks (in an alignment) 243
 Boolean operator (in decision networks) 319, 320
 Bootstrap analysis 179
 Build (connected databases) 60
 Bundles (for data exchange) 12, 81–83

Bypass normalization 110

C

- Calculate > Experiment correlations 187
- Calculate > Similarity plot 188
- Calculation priority settings 175
- Calibration 339, 346
- Calibration > Change calibration value 347
- Calibration > Image calibration 347
- Calibration curve 118
- Canberra metric coefficient 208, 245, 248, 303
- Case sensitive 49
- Categorical coefficient 208, 249, 303
- Cells > Add disk to mask 128
- Cells > Add pixels to mask 128
- Cells > Add selected 128
- Cells > Edit color scale 128
- Cells > Remove pixels from mask 128
- Change access (Netkey) 21
- Change entry key 43
- Change home directory 25
- Change towards end of fingerprint 192, 198
- Change valid range 375
- Changing fingerprint type 113
- Changing sequences in a multiple alignment 217
- Character > Change character range value 122
- Character file, new 124
- Character import, from TIFF 126
- Character types 9, 89, 121, 335
- Character types module 11, 304
- Character value (query) 50
- Characters > Add new character 121, 124
- Characters > Export character table 209
- Characters > Order characters by component 275, 276
- Characters > Show mapping 209
- Characters > Sort by character value 209
- Characters > Use character for comparisons 122
- Charts (as output from decision networks) 321, 327–328
- Check table structure 60
- Clip at max. value 353
- Clip values at extreme points 347
- Clonal complex 266
- Cluster analysis (similarity matrix) 174, 176, 180, 191, 194, 207, 212, 248, 254, 271, 278, 371
- Cluster analysis of composite data sets 247
- Cluster cutoff method 180
- Cluster significance tools 179–180
- Clustering 213
- Clustering > Advanced trees > Create consensus tree 261
- Clustering > Bootstrap analysis 180, 237
- Clustering > Calculate > Clustering (multiple alignment) 235
- Clustering > Calculate > Minimum spanning tree (population modelling) 264
- Clustering > Calculate cophenetic correlations 179, 237
- Clustering > Calculate error flags 179, 237
- Clustering > Collapse/expand branch 176
- Clustering > Congruence of experiments 187
- Clustering > Maximum likelihood cluster analysis 256
- Clustering > Maximum parsimony cluster analysis 254
- Clustering > Rendered tree export 185
- Clustering > Reroot tree 176, 236, 237
- Clustering > Select root 176, 236
- Clustering > Swap branches 176, 236
- Clustering > Tolerance & optimization analysis 195
- Clustering and statistical analysis of 2D gels 371
- Clustering method 212, 213, 245, 253, 371
- Clustering of characters 207–209
- Clustering of fingerprints 191–195
- Clustering of sequences 211–223
- Color codes (for charts) 295, 296, 297
- Color codes (for field states) 47
- Color codes (for sequences) 234
- Color codes (in charts) 282
- Color scale (BNIMA) 127
- Color scheme 29, 30
- Column properties button 33, 34, 48, 65, 83, 85, 170, 228, 302, 303, 304, 313, 315
- Combine using OR 364
- Comparative quantification 197
- Comparing 2D gels 359
- Comparison > Chart / Statistics 291
- Comparison > Compare two entries 169
- Comparison > Create new comparison 170, 301
- Comparison and cluster analysis module 11
- Complete linkage 191
- Component type 273
- Composite > Calculate clustering of characters 251
- Composite > Calculate consensus matrix 250
- Composite > Discriminative characters 161, 206, 250
- Composite > Export character table 205
- Composite > Show quantification (colors) 205, 206, 250
- Composite > Show quantification (values) 205
- Composite > Sort by character 206, 250
- Composite data sets 89, 161–162
- Composite spot query in the 2D gel type window 366
- Concentration 203
- Confidence values 326
- Configuring toolbars 32–33
- Conglomerate spot separation 344
- Congruence between techniques 186
- Connected database 17, 26, 29, 39, 40, 41, 57–70, 71, 85
- Consensus match 214
- Consensus sequence 211, 214
- Consensus sequence (in an alignment) 232–233
- Consensus tree 258
- Consider absent values as zero 121
- Contour palette 341
- Conversion to binary 207, 372
- Cophenetic correlation 179
- Copy content to clipboard 34
- Copy to all characters (color) 122
- Copy to character (color) 122
- Copy to clipboard (log file) 42
- Correct for internal weights 248, 249
- Correction parameters (sequence clustering) 213
- Correction, Jukes and Cantor 218, 232, 236
- Correction, Kimura 2 parameter 218, 236
- Correlation type 187
- Cosine correlation coefficient 191, 208, 245, 248, 303
- Cost table (parsimony) 254
- Create character (ODBC) 79
- Create from database field 177
- Create new 2D gel type 337

Create new fingerprint type 93
Creating 2D spot queries 363
Creating a BLAST database 309
Creating a reference system (2D gels) 348
Creating landmarks for normalization 350, 355
Creating synthetic gels 374
Crop > Add new crop 92
Crop > Delete selected crop 93
Crop > Rotate selected crop 92
Cropped 93
CrvConv program. See Curve Converter
Cubic spline 347
Cup type 123
Curve Converter 114, 115
Curves 93, 116
Curves > Spectral analysis 100

D

Data sources (in decision networks) 320
Database > Add all lanes to database 113
Database > Add lane to database 113
Database > Add new entries 43, 61, 72, 111, 112, 354
Database > Add new information field 44, 354
Database > Change entry key 43
Database > Change fingerprint type of lane 113
Database > Connected databases 59
Database > Link lane 113
Database > ODBC link > Configure external database link 77
Database > ODBC link > Copy from external database 78
Database > ODBC link > Download field from external database 78
Database > ODBC link > Select list from external database 78
Database > Remove all links 113
Database > Remove entry 43
Database > Remove information field 44
Database > Remove link 113
Database > Remove unlinked entries 43
Database > Rename information field 44
Database construction 339
Database descriptor file 11, 39
Database directory 40
Database field (query) 50, 67
Database field range (query) 50, 67, 71–75
Database sharing tools module 12, 63, 305
Database tools 28
Databases 10
Decision networks 319–331
Decision trees 329–331
Decrease zero level 341
Defining a new 2D experiment type 337
Defining metrics 351
Defining reference spots 349
Degeneracy of dendrograms 257
Degree (congruence of techniques) 188
Delay divergent sequences 231
Delete from users list (Netkey) 21
Demo2D 337, 360
Demobase 16, 27, 29, 36, 43
Dendrogram tools 28

Densitometric curves 99, 115
Densitometric values (BNIMA) 126
Details (bundle) 83
Dice coefficient 191, 207, 248, 251
Different bands (coefficient) 191
Differential expression 364, 365
Dimensioning > Multi-dimensional scaling 271
Dimensioning > Principal Components Analysis 272, 276
Dimensioning > Self organizing map 277
Dimensioning and Statistics module 11
Direct linkage (of 2D gel spots) 359
Discard unknown bases 213, 218, 231, 235
Disconnect user (Netkey) 22
Discriminants (with variance) 276
Discriminants (without variance) 276
Divide by variance (PCA) 273, 276
DNA transition weight 231
DNA weight matrix 231
DNS configuration 20
DNS host name 20
Dockable panel 29, 31, 169
Docking guide 31
Double locus variant 263, 268
Drag-and-drop sequence alignment 215
Drawing tool (add pixels) 345
Drawing tool (remove pixels) 346
Duplicate keys 113
Dynamical preview 95, 115

E

Edit > Arrange entries by database field 173
Edit > Arrange entries by field 47, 48
Edit > Arrange entries by field (numerical) 48
Edit > Arrange entries by similarity 301
Edit > Bring selected entries to top 49, 206, 250
Edit > Change brightness & contrast 94, 98, 115
Edit > Clear selection list 49, 54
Edit > Copy selection 55, 172, 230, 313
Edit > Cut selected gel from matching 362
Edit > Cut selected sequences 230
Edit > Cut selection 54, 172, 201, 219, 230, 254
Edit > Delete current (subset) 55
Edit > Delete selection 55
Edit > Edit tone curve 98, 340, 341
Edit > Find > Position 239
Edit > Find > Sequence 240
Edit > Freeze left pane 48, 171
Edit > Load default settings 110, 343
Edit > Move curve down 115
Edit > Move curve up 115
Edit > Paste entries from clipboard 360, 367, 369
Edit > Paste selected sequences 230
Edit > Paste selection 55, 172, 201, 219, 301, 313
Edit > Previous page 183
Edit > Redo 94
Edit > Redo last action 345
Edit > Remove curve 115
Edit > Rename current (subset) 55
Edit > Rescale curves 101, 108
Edit > Save as default settings 110, 343
Edit > Search entries 49, 108

- Edit > Settings 96, 99, 100, 105
 - Edit > Settings (BNIMA) 126, 128, 130
 - Edit > Settings (fingerprints) 101
 - Edit > Show value scale (BNIMA) 127
 - Edit > Spot info 345, 353
 - Edit > Undo 93
 - Edit > Undo last action 345
 - Edit > Zoom in 94, 184, 325
 - Edit > Zoom out 94, 184, 325
 - Edit calibration curve button 347
 - Edit database fields 277
 - Edit image (BNIMA) 126
 - Editing reference systems 373
 - Embossed view 341
 - Enable log files 40, 42
 - Enhance dark bands 341
 - Enhance weak bands 341
 - Enhanced metafile export 185
 - Enter the maximum deviation 351, 356
 - Entries > Add new entries 124, 136
 - Error flags 179
 - Estimate errors 256
 - Estimate relative character importance 278
 - Estimated spot size 344
 - Euclidean distance 156, 162, 208, 245, 303
 - Executing a decision network 328
 - Experiment 29
 - Experiment > Comparison settings 162
 - Experiment > Correct for internal weights 161, 249
 - Experiment > Train neural network 317
 - Experiment > Use for identification 313
 - Experiment > Use in composite data set 161
 - Experiment card 122, 163, 164
 - Experiment presence (query) 50
 - Experiments > Create new 2D gel type 337
 - Experiments > Create new character type 121
 - Experiments > Create new composite data set 161
 - Experiments > Create new fingerprint type 91
 - Experiments > Create new matrix type 159
 - Experiments > Create new sequence type 135, 151
 - Experiments > Edit experiment type 110, 121
 - Export band metrics 163
 - Export normalized band positions 163
 - Export normalized curve 163
 - Extend gap penalty 230, 231
- ## F
- Fast band-based database screening 302-303
 - Fast character-based identification 303
 - Fast sequence-based identification 304
 - Field states 46
 - Fields > Add new field 125
 - Fields > Remove field 125
 - Fields > Rename field 125
 - Fields > Set field content 125
 - Fields > Use as default field 125
 - File > Add experiment file 61
 - File > Add image to database 92
 - File > Add new experiment file 91, 115
 - File > Add new library unit 313
 - File > Add to database 354
 - File > Analyze with GeneMaths 209
 - File > Approved 146
 - File > Clear log file 42
 - File > Convert complexes to groups 269
 - File > Copy correspondence plot to clipboard 280
 - File > Copy discriminants to clipboard 280
 - File > Copy image to clipboard 255, 272, 343
 - File > Copy image to clipboard (characters) 275
 - File > Copy image to clipboard (entries) 275
 - File > Copy page to clipboard 185
 - File > Create matching window 360, 367, 369, 374
 - File > Create new bundle 81
 - File > Create synthetic gel 374
 - File > Delete experiment file 93, 125, 136
 - File > Edit library unit 313
 - File > Exit 43
 - File > Export 303, 304
 - File > Export bands (comparison) 193
 - File > Export character coordinates 275
 - File > Export database fields 179, 181, 255, 272, 274, 301
 - File > Export densitometric curves (comparison) 193
 - File > Export entry coordinates 275
 - File > Export overview 315
 - File > Export report to file 315
 - File > Export sequences 220
 - File > Export similarity matrix 181, 237
 - File > Fill information field 315
 - File > Import experiment file 159
 - File > Import from external database 79
 - File > Load configuration 130
 - File > Load image (BNIMA) 126
 - File > Lock 41
 - File > Open additional database 55
 - File > Open bundle 82
 - File > Open experiment file (data) 93, 115, 116, 124, 135, 136
 - File > Open experiment file (entries) 111, 112, 124, 136
 - File > Open reference gel 116, 117
 - File > Preferences 30, 31, 34, 35, 36, 45, 47, 175
 - File > Print all pages 185, 238
 - File > Print correspondence plot 280
 - File > Print database fields 301
 - File > Print discriminants 280
 - File > Print image 255, 272, 343
 - File > Print image (characters) 275
 - File > Print image (entries) 275
 - File > Print preview 183
 - File > Print selected pages 238
 - File > Print this page 185
 - File > Printer setup 185, 238
 - File > Save changes 362
 - File > Save configuration as 130
 - File > Statistical analysis 372
 - File > Tools > Horizontal mirror of TIFF image 93
 - File > Tools > Vertical mirror of TIFF image 93
 - File > Update linked information 116, 117
 - File > View 3D image 342
 - File > View log file 42
 - Filtering 100, 343
 - Find in table 34
 - Finding a subsequence in multiple alignment 217
 - Fingerprint bands (query) 50
 - Fingerprint data editor 93, 94, 99, 102, 107

Fingerprint image import window 92
Fingerprint processing reports 28
Fingerprint types 9, 89, 91, 335
Fingerprint types module 11
Fixed panel 31
Font type 29, 30, 31
Force through 100% 188
Foreground (calculation priority setting) 175
Foreground color 342
Furthest neighbor (K-means) 181
Fuzzy logic 191, 209, 364
Fuzzy logic match 364
Fuzzy zone 326, 327

G

Gap penalty 213, 218, 231, 235
Gaussian filter 343
Gel image tone curve editor 340
GelCompar version 4.x, import from 119
Gelstrip thickness 116
GeneMaths XT 209
Genescan tables, importing 117
Geographical plugin 28
Global alignment (fingerprints) 103
Global alignment. See Multiple alignment (of sequences)
Gower coefficient 162, 208
Gray zone (bands) 105
Grid > Add new 128
Grid > Delete 128
Grid > Delete selected 128
Grid definition 127
Grid panel 29, 33–35, 45, 46, 47, 170, 226, 228, 240, 242, 268
Group > Create from database field 178
Group > Partitioning of groups 181
Group creation priority setting 177
Group separation statistics 182
Group violations 182, 183
Groups 176–179
Groups > Assign selected to 177
Groups > Assign selected to > None 181
Groups > Create from database field 195
Groups > Group separations 182
Groups > Multivariate Analysis of Variance 278
Groups > Partitioning of groups 182

H

HDA plugin 28
Hidden nodes 317
Home directory 10, 11, 15, 25, 39
Hue only (BNIMA) 127
Hypothetical types 264

I

ID code 40, 41
Identification 299–331
Identification > Create new library 313
Identification > Fast band matching 302

Identification > Identify selected entries 314
Identification > Probabilistic identification 305
Identification against database entries 301–312
Identification module 11, 31
Identification using decision networks 319–331
Identification using libraries 313–317
Idle time background (calculation priority setting) 175
Image > Convert to gray scale > Averaged 92
Image > Convert to gray scale > Blue channel 92
Image > Convert to gray scale > Green channel 92
Image > Convert to gray scale > Red channel 92
Image > Invert 92
Image > Load from original 93
Image > Mirror > Horizontal 92
Image > Mirror > Vertical 92
Image > Rotate > 180° 92
Image > Rotate > 90° left 92
Image > Rotate > 90° right 92
Image > Show normalized 361, 367, 369, 374
Image > Show overlap 362
Image > Update normalization 362
Image coloring 342
Image type (BNIMA) 126
Import Fingerprint files from Automated Sequencers 115
Import plugin 28, 77, 115, 123, 124, 135, 156, 225
Importing 2D gel image files 338
Increase contrast 341
Increase zero level 341
Input operator (in decision network) 319
Inserting and deleting gaps in multiple alignment 215
Install InfoQuest FP 15
Install Netkey server program 19
Installation directory 15
Intensity query 364
Intermediate pen size 346
Internal reference markers 116
Interpolation 347
Inverted values 342
IP address 20, 21
IUPAC code 52, 143, 147, 165, 217, 221, 233, 240, 242, 320

J

Jaccard coefficient 191, 207, 248, 251
Jackknife 182
Jeffrey's X 191
Jukes and Cantor 213

K

Kendall's tau 187
K-means partitioning 181
Kohonen map 276–278
K-tuple size 230, 231

L

Lanes > Add marker point 116
Lanes > Add new lane 97
Lanes > Auto search lanes 96, 116

- Lanes > Copy geometry 117
- Lanes > Delete selected lane 97
- Lanes > Paste geometry 117
- Large pen size 346
- Layout 271
- Layout > Create rooted tree 255
- Layout > Enlarge image size 184
- Layout > Label query members only 369
- Layout > Label with 369
- Layout > Optimize branch spread 255
- Layout > Preserve aspect ratio 274
- Layout > Reduce image size 184
- Layout > Rescale curves 193
- Layout > Show 3D plot 274
- Layout > Show bands 193
- Layout > Show branch lengths 255, 256
- Layout > Show construction lines 272
- Layout > Show curves as images 193
- Layout > Show dendrogram 212, 272
- Layout > Show densitometric curves 193
- Layout > Show distances 176
- Layout > Show gel images 368
- Layout > Show group colors 254, 272, 274
- Layout > Show image 197, 212, 254
- Layout > Show keys 272, 274
- Layout > Show keys or group numbers 254
- Layout > Show matrix 180, 212
- Layout > Show matrix rulers 181
- Layout > Show metric scale 193, 198
- Layout > Show rendered image 272
- Layout > Show similarity matrix 184
- Layout > Show similarity values 181
- Layout > Show space between gelstrips 185, 193
- Layout > Show spot info 363, 368
- Layout > Show table preferences 367
- Layout > Similarity shades 181, 237
- Layout > Stretch (X dir) 198
- Layout > Use colors 184
- Layout > Use component as X axis 274
- Layout > Use component as Y axis 274
- Layout > Use component as Z axis 274
- Layout > Use group numbers as keys 179, 254, 272, 274
- Layout > Zoom in 172, 198
- Layout > Zoom out 172, 198
- Least square filtering 100
- Levels and relations 71, 75
- Library 313
- License settings 20, 23, 25
- License string 16, 19
- Linking spots with reference spots 350, 356
- Load band classes from experiment type 201
- Local alignment (fingerprints) 103
- Local database 26
- Local database, converting to connected database 63
- Lock configuration 32
- Log files 41
- Logarithmic 352
- Logarithmic dependence 112
- Logical operations (in decision networks) 319
- Logical operator 50, 52, 53, 67, 68

M

- Manhattan distance coefficient 264, 303
- MANOVA 278
- Manual selection 364
- Mapping (of characters) 123, 209
- Match against selection only (Jackknife) 182
- Matching 2D gel spots 339
- Matching spots on different gels 359
- Matrix types 9, 89, 335
- Matrix types module 11
- Maximal similarities (jackknife) 182
- Maximum difference 302
- Maximum likelihood clustering 253, 256
- Maximum number of gaps 213, 214
- Maximum parsimony clustering 253–256
- Maximum similarity used 187
- Maximum value 115, 364, 365
- Maximum value (grayscale) 95
- Mean intensity 372
- Median filter 100
- Merge selected spot 346
- Metric > Assign unit 112
- Metrics 357
- Metrics > Add marker 112
- Metrics > Copy markers from reference system 112, 118
- Metrics > Cubic spline fit 112
- Metrics range of fingerprint 118
- Microplate (BNIMA) 126
- Minimal area 106
- Minimal expression 366
- Minimal profiling 105, 106
- Minimum consensus percentage 214
- Minimum match sequence 213, 214
- Minimum profiling 344
- Minimum similarity used 187
- Minimum spanning trees 263–270
- Minimum spot size 344
- Minimum value (grayscale) 95
- MLST plugin 28
- Mode filter 100
- Modules 11–12
- Molecular sizes (defining) 111
- Monotonous fit 188
- Multi-level undo function 345
- Multiple alignment (of sequences) 211, 212, 213, 214, 215, 217, 218, 219, 220, 230, 232, 235, 241, 253
- Multi-state coefficient 248
- Multivariate analysis of variance 278
- Mutation rate 256
- Mutation search 241–243
- Mutations > Jump to next 242
- Mutations > Jump to previous 242
- Mutations > Search 241

N

- Navigator 36, 94
- Nearest neighbor (K-means) 181
- Needleman-Wunch algorithm 231
- Negative search 49, 301
- Neighbor Joining 176, 191, 211, 213, 218, 236

Neighbor match 214
Netkey 19, 20
Network 20
Neural network 315
New character type 121
New database (creating) 26
New fingerprint type 91
New intensity query 366
New matrix type 159
New ODBC 65
New sequence type 135, 151
No. diagonals 230, 231
Nodes 319
Normal priority background 175
Normalization 93, 110, 116, 339
Normalization > Add selected spot(s) to reference system 350
Normalization > Auto assign (bands) 103, 116
Normalization > Auto link spots 351, 356
Normalization > Automatically find landmarks 350
Normalization > Delete all assignments 103
Normalization > Show distortion bars 104
Normalization > Show distortion maze 356
Normalization > Show image side by side 355
Normalization > Show normalized view 101, 103, 104, 105, 350, 355
Normalization > Show overlapped images 355
Normalization > Show reference gel 350, 354
Normalization > Show superimposed images 355, 356
Normalization > Show synthetic reference system 350, 354
Normalization > Unlink selected spot(s) 351, 357
Normalization > Update normalization 104, 350, 355, 356
Normalization > Use spot as landmark 355
Normalization of other 2D gels 354
Normalized view 116
Nucleic acid sequences 135
Number of bootstrap simulations 254
Number of columns (character type) 121
Number of groups 181
Number of nodes 116
Number of rows (character type) 121
Numerical coefficient 208, 248
Numerical values 121

O

Ochiai 191
ODBC connection string 60
ODBC, import 77, 123
One dimension quantification 203
Only spots in active query 375
Open entry 44
Open gap penalty 212, 213, 230, 231
Operators (in decision networks) 320–321
Optimization 192, 194, 195, 198
Optimize positions (MDS) 271
Optimize topology 254
Output actions (of decision networks) 319, 320

P

Pairwise alignment 211
Pairwise alignment settings 212
PAM matrices 309
Panels, display options 31–32
Parsimony 253
Paste data from clipboard 130
Peak detection parameters 103
Pearson correlation coefficient 110, 187, 191, 208, 245, 248, 251, 273, 303
Pearson correlation test 287, 295
Performing a BLAST search 310
Plate (characters) 122
Plot > Use discriminant as X axis 280
Plot > Use discriminant as Y axis 280
Plugin installation toolbox 27, 28, 29, 115, 124, 156, 225
Plugin tools 27–28
Point-biserial correlation 180
Polymorphism analysis 197
Polynomial 347
Polynomial degree (BNIMA) 130
Port number 20
Position tolerance 191, 192, 195, 198, 302
Position tolerance, find best 194
Position-based search results (in alignments) 228, 229, 236
Preview (band search) 106
Principal component analysis (PCA) 272–276
Priority rules 265
Probabilistic identification 304–308
Processed 92
Processing 2D gel images 338
Properties 20

Q

Quantification > Add cells to character set 129
Quantification > Band quantification 109
Quantification > Calculate concentrations 109
Quantification > Define calibration point 130
Quantification > Export to clipboard (BNIMA) 130
Quantification > Search all surfaces 109
Quantification > Search surface of band 109
Quantification units 105
Quantification, comparative 197
Quantified value 372
Quantity 364, 365
Queries 49
Queries > Edit query 366
Queries > New manual selection 365, 369
Queries > New spot fields query 365
Queries > Set as active query 369
Queries > Update 364, 366
Query > Delete query 366
Query > Set as active query 366

R

Radius 343
Rainbow palette 95, 341

- Rank correlation coefficient 208, 248
 - Raw data 96
 - Recall saved configuration 32
 - Reduce contrast 341
 - Reference > Use as reference lane 101
 - Reference lane 116
 - Reference system 101, 348, 374
 - References > Add external reference position 101
 - References > Add internal reference position 104
 - References > Copy normalization 117
 - References > Paste normalization 117
 - References > Use all lanes as reference lanes 116
 - Refresh (connected databases) 60
 - Refsystem > Add selected spot(s) 362, 374
 - Refsystem > Add to matching window 362
 - Refsystem > Assign ID code to spots 362
 - Refsystem > Delete 374
 - Refsystem > Delete selected spot(s) 362
 - Refsystem > Delete spot(s) 374
 - Refsystem > Refresh spots 374
 - Refsystem > Remove all spots 374
 - Refsystem > Set as active reference system 360, 374
 - Refsystem > Use selected gel as temporary standard 362
 - Registry 9
 - Regression (congruence of techniques) 188
 - Regression curve 111, 117
 - Regular expressions 35, 320
 - Relations in a database. See Levels and relations 71–75
 - Relative band surface 203
 - Relative to base gel 364
 - Relative to max. value (bands) 106
 - Relative usage (Netkey) 22
 - Relative volume 203, 364, 365, 372
 - Remove spot tool 345
 - Removing common gaps in a multiple alignment 217
 - Rename (bundles) 82, 83
 - Rendered tree 185
 - Represent as list 122
 - Represent as plate 122
 - Resolution of normalized tracks 110
 - Restore default configuration 32, 33, 72
 - Restore default settings 104, 232
 - Restrict content to states 47
 - Restricting query 60, 66
 - Result set 302
 - Rotated 352
 - Rotated & curved 352
 - Rotation 353
 - Run selected queries 366
- S**
- Save as default calibration 347
 - Save current configuration 32
 - Save settings 225
 - Script language 12
 - Scripts > Browse Internet 12, 118, 123, 135, 194
 - Scripts > Edit script 12
 - Scripts > Run script from file 12
 - Search in list 49, 301
 - Security driver 19
 - Security key 19
 - Select branch into list 176, 236
 - Self-Organizing Map (SOM) 276–278
 - Send message (Netkey) 22
 - Sequence > Align external branch 219
 - Sequence > Align internal branch 219
 - Sequence > Calculate global cluster analysis 218
 - Sequence > Change saved sequence 217
 - Sequence > Consensus blocks 214
 - Sequence > Consensus difference 214
 - Sequence > Create consensus of branch 214, 223
 - Sequence > Create locked group 216
 - Sequence > Edit 136
 - Sequence > Find sequence pattern 217
 - Sequence > Lock / unlock dendrogram branch 216
 - Sequence > Multiple alignment 213, 254
 - Sequence > Neighbor blocks 214
 - Sequence > Paste from clipboard 136
 - Sequence > Reload sequence from database 217
 - Sequence > Show global cluster analysis 219
 - Sequence > Unlock group 217
 - Sequence operator (in decision networks) 319, 320
 - Sequence translation 239
 - Sequence translation tools 28
 - Sequence types 9, 89, 135, 335
 - Sequence types module 11
 - Server computer name 20
 - Set as base gel 364
 - Settings 25, 42
 - Settings (Netkey) 22
 - Settings > Binary conversion settings 123
 - Settings > Brightness & contrast 110
 - Settings > Comparative quantification 110
 - Settings > Edit reference system 111, 118
 - Settings > Enable fast band matching 302
 - Settings > Exclude active region 223
 - Settings > General settings 110, 122, 131
 - Settings > Include active region 223
 - Settings > New reference system (curve) 118
 - Settings > New reference system (positions) 118
 - Settings > Set as active reference system 118, 119
 - Settings > Spot quantification settings 372
 - Settings > Statistics 183
 - Shape/darkness 345
 - Shoulder sensitivity 105, 106
 - Show > Show less matches 315
 - Show > Show more matches 315
 - Show bands 193
 - Show dendrogram 177, 179, 186, 219
 - Show matrix 186, 271
 - Show quantification (colors) 248
 - Show spot info 353
 - Similarity 248
 - Similarity calculation 212, 213
 - Simple matching coefficient 207, 248, 251
 - Single linkage 191
 - Single locus variant 263, 266
 - Small pen size 346
 - SNP analysis. See Mutation search 241–243
 - Source file location 60, 61, 65
 - Spa Typing plugin 28
 - Spearman rank-order correlation test 295
 - Spike removal 338, 343
 - Split selected spot 346

Spot area 372
Spot contrast enhancement 344
Spot detection 338, 343
Spot field query 364
Spot information box 353
Spot information pop up window 345
Spot intensity measure 364
Spot quantification 353
Spot removal (2D image) 97
Spot volume 372
Spots > Add to active query 365, 369
Spots > Automatic search 343
Spots > Break link 362
Spots > Delete selected spots 345
Spots > Merge selected spot 346
Spots > Remove from active query 369
Spots > Split selected spot 346
SQL query 302
Standard deviation 179
Standardized characters 248
Start service (Netkey) 20
Startup program 25
Statistics (Netkey) 22
Status (Netkey) 23
Stop service (Netkey) 21
Stored trees dialog box 261
Streak removal 338
String operator (in decision networks) 319, 320
Strips 93, 116, 117
Strips > Increase number of nodes 97
Strips > Make larger 97
Strips > Make smaller 97
Subdivide existing groups 178
Subsequence (query) 50
Subsequence search 217
Subset 48, 49, 53–55, 66, 67, 68, 69, 173
Subset (of an alignment) 228
Subtract average (PCA) 273
Synthetic gels 374

T

T test for mean value 294
Take from experiments 248, 249
TCP/IP 20
The 2D gel data editor window 340
The 2D gel type window 360
The create synthetic gel window 375
The spot description fields window 363
Thickness (image strips) 97
Threshold (for peak detection) 103
Tie handling 182
Tolerance & optimization statistics 194
Tolerance. See Position tolerance 302
Tone curve 98
Toolbar 13, 32–33
Transversal clustering 250
Trend data parameter 50
Trend data types 9, 89
Trend data types module 11
TrendData > Create trend data window 246

TrendData > Export character table 246
TrendData > Order by curve 246
TrendData > Order by parameter 246
TrendData > Show parameter colors 245
TrendData > Show parameter values 245
TrendData > Show parameter values & colors 245
TrendData > Sort entries by character 246
Truecount operator 327
Two dimensions (quantification) 203

U

Uncertain bands 105, 192, 198
Unit gap penalty 212, 213
UPGMA 182, 191, 213, 248
Use active zones only 218
Use as default database 60
Use ClustalW tree 232
Use conversion cost 213, 231
Use existing pairwise clustering 231, 232
Use fast algorithm 213, 214, 231
Use quantitative values (PCA) 272
Use square root 208, 248
Used range 302

V

Validation samples 317
Valley depth (peak separation) 103
Value operator (in decision networks) 319, 320
Vertical only 352
View > Show spot outlines 342
View calibration curve (BNIMA) 130
VNTR plugin 28
Volume 203

W

Ward 191, 213
Wilbur-Lipman algorithm 231
Wilcoxon signed ranks test 295
Window > Show / Hide toolbar 32
Window size 231
Windows Vista 9, 19, 20, 25, 32
Write to field operator 324

X

XML files (for data exchange) 12, 28, 63, 81, 83, 305, 312
XML Tools plugin 28, 57, 63, 64, 65, 77, 83, 305

Z

Zero value 353
Zoom slider 29, 35, 94, 109, 138, 143, 172, 184, 198, 228, 235, 238, 266, 271, 275, 291, 325



**Bio-Rad
Laboratories, Inc.**

*Life Science
Group*

Web site www.bio-rad.com **USA** 800 4BIORAD **Australia** 61 02 9914 2800 **Austria** 01 877 89 01 **Belgium** 09 385 55 11 **Brazil** 55 21 3237 9400
Canada 905 712 2771 **China** 86 21 6426 0808 **Czech Republic** 420 241 430 532 **Denmark** 44 52 10 00 **Finland** 09 804 22 00 **France** 01 47 95 69 65
Germany 089 318 84 0 **Greece** 30 210 777 4396 **Hong Kong** 852 2789 3300 **Hungary** 36 1 455 8800 **India** 91 124 4029300 **Israel** 03 963 6050
Italy 39 02 216091 **Japan** 03 5811 6270 **Korea** 82 2 3473 4460 **Mexico** 52 555 488 7670 **The Netherlands** 0318 540666 **New Zealand** 0508 805 500
Norway 23 38 41 30 **Poland** 48 22 331 99 99 **Portugal** 351 21 472 7700 **Russia** 7 495 721 14 04 **Singapore** 65 6415 3188 **South Africa** 27 861 246 723
Spain 34 91 590 5200 **Sweden** 08 555 12700 **Switzerland** 061 717 95 55 **Taiwan** 886 2 2578 7189 **United Kingdom** 020 8328 2000
