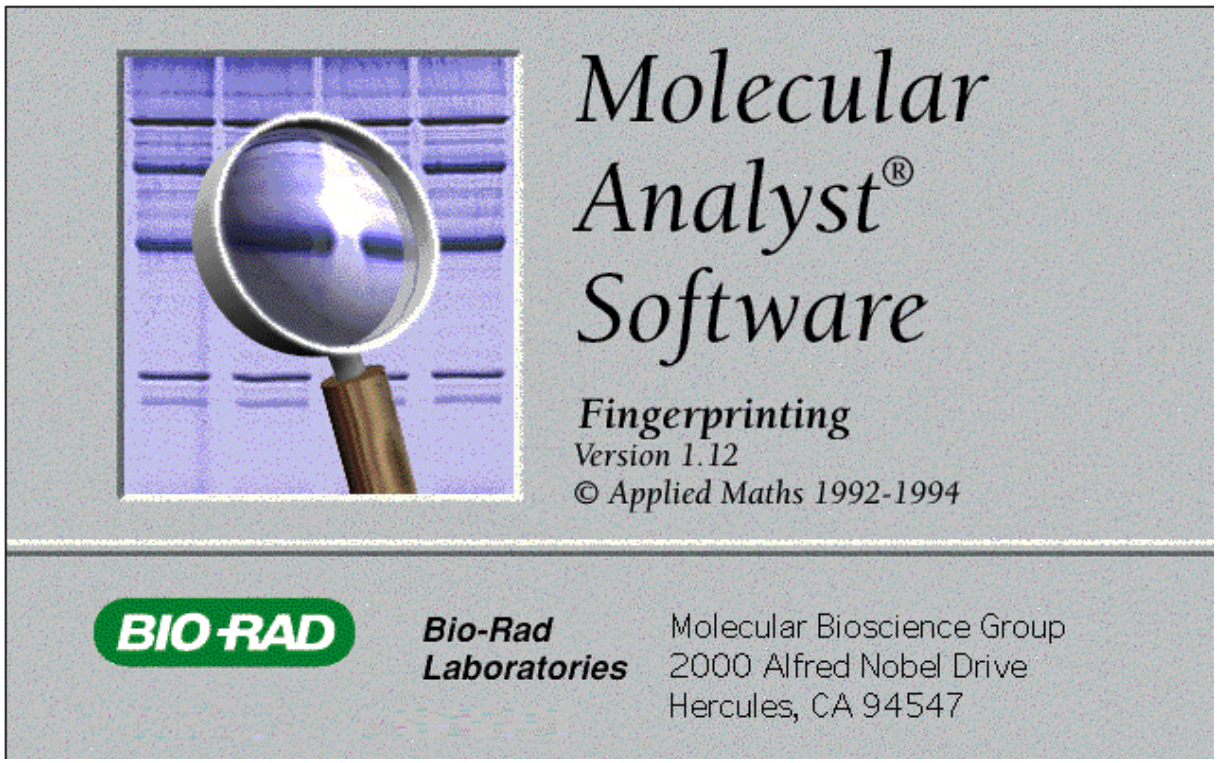



# Molecular Analyst Fingerprinting, Fingerprinting Plus, and Fingerprinting DST Software

## Software Instruction Manual

**Catalog Numbers**

**170-7561, 170-7562, 170-7901**

For Technical Service
Call Your Local Bio-Rad Office or
in the US Call **1-800-4BIORAD**
**(1-800-424-6723)**

P / N 4000076-02 Rev A

**SUPPORT BY BIO-RAD**

The Molecular Analyst® Fingerprinting software and this manual are released by Bio-Rad Laboratories. 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. Molecular Analyst Fingerprinting 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 Molecular Analyst Fingerprinting software, or suggestions for improvement, refinement or extension of the software to your specific applications. If we are unavailable to answer your e-mail, phone call, or fax message immediately, we will contact you as soon as possible.

Upgrades of the licensed Molecular Analyst Fingerprinting software modules will be provided to registered customers in exchange for a nominal charge which will cover only the expenses made for recording media, manuals and shipping. In order to have entitlement to full support and upgrades, one signed copy of the license agreement should be returned to Bio-Rad.

**LIMITATIONS ON USE**

The Molecular Analyst Fingerprinting software and this accompanying manual 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 manual may be reproduced by any means without prior written permission of the authors.

# TABLE OF CONTENTS

# 1. Introduction

Different separation techniques such as electrophoresis, gas chromatography and HPLC are routinely applied for the separation of cellular proteins, fatty acids, polyamines, lipopolysaccharides and DNA fragments or chromosomes in order to obtain fingerprints to classify, identify or subtype organisms.

Among separation techniques used for fingerprinting, typing and identification, electrophoresis is certainly the most successful one. An electrophoretic banding pattern of DNA fragments or cellular proteins can be used as a characteristic fingerprint of an organism under study. Various extraction, fragmentation, separation and detection techniques make it possible to differentiate at virtually all levels of relationship. The possibility to store and manage large quantities of data on personal computers has made electrophoresis even more attractive than ever for typing of micro-organisms, plants and animals.

Basically, the application of electrophoresis to generate fingerprints involves three important steps.

(1) A sample of macromolecules (protein or DNA) is electrophoresed to produce a characteristic and reproducible banding pattern. The choice of a specific electrophoresis technique depends on the purpose of the identification system.

(2) The patterns are recorded by a scanner, video camera or densitometer linked to a computer. A densitometer measures the optical density along the patterns whereas a flatbed scanner or video camera rasters the complete gel into a matrix of densitometric values called a bitmap. The bitmap of digitized optical density values is transferred to a computer and stored as image file, often in Tagged Image File Format (TIFF). In DNA sequencers, steps (1) and (2) are integrated in one single system, where the optical detector is located at a fixed position on the gel, and the bands are recorded directly during electrophoresis while passing the detector.

(3) The tracks are further processed by appropriate software. This process, which will be discussed in more detail, involves (i) documentation of gels, which involves delineating and naming patterns (ii) normalization of patterns; (iii) generation of databases and (iv) grouping or identification of patterns by quantification of their resemblance.

With the availability of sophisticated electrophoresis apparatus, new types of scanners, densitometers and video cameras and powerful personal computers, identification and characterization by electrophoresis has become possible for laboratories which are involved in practical diagnosis rather than in fundamental research. The availability of specialized and reliable software to process and compare patterns is determinative for the success of any technique

used. Below, we describe the principles of the general approaches for normalization, grouping and identification of electrophoresis patterns as realized in the PC-directed software package Molecular Analyst Fingerprinting software.

## *Normalization of patterns*

When protein or DNA samples are electrophoresed to obtain fingerprints for identification, it is obvious that all the experimental procedures need to be standardized and reproducible in order to warrant the reliability of the technique. However, even in the most reproducible conditions, discrepancies are observed between and even within gels. Visually, it is often easy to compensate for distortions between similar patterns. The human brain is known to have an intelligent pattern recognition system, but when less similar patterns are to be compared, particularly on different gels, alignment by eye becomes ambiguous and may result in subjective or faulty interpretations.

Therefore, more objective methods have been designed, based on direct alignment of patterns via known reference bands within the patterns or on indirect alignment via dedicated reference patterns.

**a) Alignment by external reference patterns.** The "reference sample" can be a complex protein or DNA sample, or a set of molecular size markers. This sample is applied once or at regular intervals on each gel. After densitometric recording of the gel, the reference tracks are stretched or compressed to match each other. The alignment of reference patterns is done by aligning their corresponding bands and by subsequent interpolation of the intermediate values. The other (non-reference) tracks are aligned gradually according to their closest neighbouring reference tracks. We will call this process the *normalization* of a gel. By defining one reference track as "standard" pattern and further aligning the bands of all reference tracks from every gel to the corresponding bands of that single pattern, all gels become compatible with one another, which makes it possible to generate databases, to compare and cluster patterns from different gels and to identify unknown organisms.

**b) Alignment by internal reference bands.** Instead of, or in addition to normalizing gels by separate reference patterns as described above, it is possible to correct for distortions in the gel by combining any sample with any other using one or more bands which are known to be the same in different patterns. Molecular Analyst Fingerprinting software provides the possibility to define and store sets of reference positions to perform such alignments. This allows the user to apply one or more reference patterns on each gel plus one or more internal reference bands in non-reference pattern, to achieve stronger internal normalization of gels.

***Normalizing 2D image files***

All comparative methods in Molecular Analyst Fingerprinting software can make use of densitometric curves. These can be derived from the gel patterns in case of electrophoresis scans. This also makes it possible to process and analyze gas chromatographic, HPLC, capillary electrophoresis, or spectrophotometric records, as well as any other fingerprinting data recorded as densitometric profiles. Realistic reconstructions of patterns, derived from the densitometric curves may be displayed as images. In case of electrophoresis tracks, spots or artefacts on the original image may be wrongly be recorded as bands on densitometric traces. Therefore, Molecular Analyst Fingerprinting software offers a unique feature to normalize the two-dimensional TIFF (or other bitmap) files in addition to the densitometric curves, and display patterns as ***normalized gelstrips***. The possibility to display normalized 2D images of any selection of patterns or in clusterings combines all the power of the Molecular Analyst Fingerprinting software normalization methods with all the information of original 2D scannings. The normalized images are a performant tool for publication and reporting.

## *Generating databases*

In order to generate databases, efficient file management of normalized gels, allowing editing and modification of information, visualization of curves and gelstrips, etc., is necessary. For taxonomy as well as identification purposes, one should be able to compare any selected group of tracks. Molecular Analyst Fingerprinting software contains extensive edit, search and selection possibilities, and allows unlimited databases to be managed with maximal speed and efficiency.

Modern fingerprinting techniques based on PCR and primer technology require sometimes that more than one gel is run for each pattern in order to enhance the resolving power of the system. Molecular Analyst Fingerprinting software allows different runs for each organism to be combined in order to create composed fingerprints, which behave like normal complex patterns and may be shown as images, or clustered, identified, etc.

## *Cluster analysis*

Numerical comparison provides a way to objectively compare large sets of characters of between many organisms. It allows the homogeneity of populations to be established, unknown organisms to be located into known taxa, misnamed or atypical strains to be determined, etc. The most common grouping method is to calculate a matrix of similarities between every pair of organisms, and to deduce a dendrogram from the matrix by the UPGMA clustering technique (unweighted pair group method using arithmetic averages). Many coefficients have been described to express similarity between sets of characters, each having their specific properties (advantages and disadvantages) and applications. In case of complex banding patterns such as protein patterns or RE fingerprints with large numbers of bands, the Pearson

correlation coefficient has proved to be the most objective and reliable measure of similarity. This coefficient is by far the most robust one, for the following reasons:

(1) It is independent of relative concentrations of patterns;
(2) is largely insensitive to differences in background.
(3) Unlike most other coefficients, it does not suffer from subjective band-detection and band-matching criteria, since it compares entire curves rather than band characteristics.

However, applications based on fingerprints often require that only the positional correspondence of bands is compared, not the relative area under the peaks. Molecular Analyst Fingerprinting software offers some other similarity coefficients based on bands instead of curves: the Jaccard coefficient and the related Dice coefficient. In addition, a more sophisticated coefficient, taking into account the peak area as well as occurrence, is available.

Hierarchic clustering based on similarities is one of the standard methods for grouping in the Molecular Analyst Fingerprinting software system. Three approved clustering algorithms, i.e. UPGMA, Ward's method and the Neighbour Joining method are available for any type of similarity coefficient. An impressive variety of advanced edit functions allow hierarchic dendrogram structures and differentially shaded and sorted similarity matrices to be visualized and printed in the most diverse ways. These functions are particularly of interest for taxonomic and epidemiological studies, but are also useful for the delineation of homogeneous groups and for selection of representative entries to generate libraries for identification (see below).

Principal Components Analysis (PCA) can be used as an interesting alternative to the hierarchical methods. Starting directly from the densitometric data, PCA allows a three-dimensional representation of the taxonomic groups to be produced as clouds of dots in spatial conformation. PCA overcomes some important disadvantages of conventional hierarchical grouping methods but introduces other simplifications. One of the peculiarities of PCA is that it is not suited to represent systems in which many different groups exist. The dendrogram-based clustering techniques are incontestably better to represent large numbers of groups. From a mathematical point of view however, PCA is the most genuine grouping method, but it requires millions of precise floating point operations to compare many patterns with one another. The ideal case for PCA is the separation of two or three populations, which can be hardly separated by any other clustering technique. The PCA module in Molecular Analyst Fingerprinting software is made attractive by its unique three-dimensional representations and on-screen animation and its possibility to perform statistics and significance tests on predefined or user-defined populations.

## *Identification*

A great interest of electrophoresis fingerprinting lies in its power as rapid identification technique. As outlined above however, reliable identification requires in the first place standardized patterns, but obviously also a reliable identification library that consists of representative entries. There are two ways to identify unknown patterns in Molecular Analyst Fingerprinting software. The first way is the simplest, and is based on comparison of a batch of unknown patterns against any selection of database patterns, using Pearson correlation, band matching or area-sensitive coefficients. The result typically is a list of the closest database patterns in decreasing order of likeliness for each of the unknowns.

The second method is more sophisticated and is based on the creation of specific identification libraries by the user. Since Molecular Analyst Fingerprinting software allows large numbers of patterns to be compared and grouped, it is possible to define representative patterns accurately within these populations. Such patterns can be used to create libraries which will serve for identification of new and unknown patterns. A Molecular Analyst Fingerprinting software library consists of user-defined units, each representing a homogeneous entity that contains one or more representative patterns. Multiple libraries can be generated. Identification is based on one of the three described correlation or band-matching coefficients using the units of a library as test cases. Libraries for identification are highly specialized objects allowing as accurate identification as possible in the given electrophoresis system. Specific bands or zones on the profiles can be defined for each taxon or group and when multiple patterns are included in a library unit, the Molecular Analyst Fingerprinting software identification module will calculate the statistical significance of an identification case. Detailed reports include display of variable and stable zones between unknown and library entries, zone-dependent standard deviations, and indication of confidence of identification cases.

## *Quantification and polymorphism analysis*

A somewhat different application of Molecular Analyst Fingerprinting software is its extensive module for "comparative quantification". Starting point is a high-quality database of normalized patterns. Each pattern is decomposed mathematically into Gaussian curves which represent the peaks and shoulders as closely as possible. This approach allows independent areas of overlapping peaks to be reproduced, and the exact position and area of a shoulder to be reconstructed iteratively. The process can be performed automatically, but can be overridden or manipulated by the user at any time, which allows correction for artifacts whenever necessary.

When two-dimensional data are available as TIFF scans, the concentration of bands is determined by delineating the contour of bands or spots on the 2D scan and calculating the volume. The extensive quantification tools in Molecular Analyst Fingerprinting software can make use of both the Gaussian

desintegration and the 2D-volume determination. Once bands have been defined for a pattern, this information is stored together with the densitometric data and can be used for cluster analysis or quantification. Database searches can now be performed not only on the information fields of the tracks, but also on the presence of a band or a combination of bands with given specifications.

The most evident result of quantification is to display or print detailed reports of all bands of a pattern, including for each band its position, absolute or relative concentration, RF-value, molecular weight, length in kb, isoelectric point or whatever metric relevant to the system. However, the Molecular Analyst Fingerprinting software system is designed to compare patterns, and in this light, the specialization of the quantification module is that it allows comparison and quantification of many selected patterns from one of the normalized databases, compiled in one surveyable report. The comprehensive on-screen anatomy and comparison of up to 1000 patterns is a unique feature with great power for diagnostic and analytical applications.

The quantification module generates tables of all bands found, with their occurrence and concentration on each of the patterns, links corresponding bands on different patterns, shows matrices of correspondence and allows printing of detailed and analytical reports of comparisons between any pair of patterns. Powerful tools for polymorphism analysis allow comparative tables in various formats to be exported to mathematical or genetic mapping software.

## *Exchange of databases and information*

The success of epidemiological research and control relies strongly on the availability of channels through which information can be exchanged among different research units. Molecular Analyst Fingerprinting software offers a module, the ***Database Sharing Tools***, which defines an open standard for the exchange of fingerprint information among different research sites. Databases or parts thereof can be condensed into "*Bundles*", which are structured information packets ready to transmit over the Internet. One can create a list of any length of database patterns and convert this to a *bundle*. The user can specify which type of information the bundle will contain: densitometric curves, 2D-gelstrips, band positions, molecular sizes, and any combination of information fields available for the database. All this information is merged into the bundle and can be compressed into small packets to send over the Internet. The recipient can decompress the received data using Molecular Analyst Fingerprinting software, and compare the patterns with his own database.

The Bundles concept addresses both the issues of a common and flexible file format and the comparison of differently standardized gels. Because different research laboratories often use different electrophoresis parameters and/or reference markers, molecular weights of bands are the sole common source of information which can be shared among different sites. The Database Sharing

Tools therefore include a *remapping* technique which makes it possible to convert the full fingerprint information from one reference system to another, including *densitometric* and *2D-image* information. This remapping is based either on the *molecular weight calibration curves* in both systems, or on a dedicated *remapping gel*, which combines both reference patterns in one run.

# 2. Installation of the programs

The Molecular Analyst Fingerprinting software programs run on any IBM compatible personal computer with Microsoft Windows 3.1 Windows 95 or Windows NT installed. The minimal hardware requirement is a 486DX with 8 MB random access memory (RAM), although 16 MB RAM is highly recommended. Accelerated VGA graphics in 256 colours or better is required. High resolution modes (e.g. 800x600 or 1024x768 pixels) will make editing and interpretation of large images, dendrograms, quantifications etc. easier. Therefore it is recommended to have a 17" or larger monitor to run Molecular Analyst Fingerprinting software comfortably.

**You should have your Windows system in 256 colours mode or better**. When higher colour definitions are used (e.g. 65 thousand or 16 million colours), dynamic adjustment (animation) of colour palettes (i.e. instant adjustment of screen colours while dragging brightness and contrast scroll bars) will not possible. In such modes, the palette changes will only take effect after closing the palette edit window, or pressing the ***Preview*** button, so that the image windows are updated. On the other hand, all colours needed by Molecular Analyst Fingerprinting software will always remain available, even when other software requesting custom colour palettes is run simultaneously. This is not the case in 256 colours mode. Since graphics in Molecular Analyst Fingerprinting software is quite complex, a powerful graphics accelerator is recommended to speed up graphical output, particularly when high resolution modes are used. Special (non standard Windows) fonts can be tried but are not warranted to give satisfactory results in all windows and dialog boxes. Do not use large fonts.

Printer outputs are possible on all printers supported by Windows. True RGB-colour printing is fully supported for all reports and images.

*NOTE: Utility and desktop manager programs which change the interface of Windows, may cause conflicts with Molecular Analyst Fingerprinting software. Do not install any of such interfaces unless you have strong reasons to do so.*

## 2.1 Installing the single user software

Install Molecular Analyst Fingerprinting software as follows:

1. Switch the computer's and printer's power OFF and insert the protection key (dongle) in the LPT1 port of the computer. Then connect the printer to the key and switch both the printer and computer ON.

2. Insert the Molecular Analyst Fingerprinting software installation disk 1 in the disk drive.

3. Start Windows and choose the command "**Run...**" from the "**File**" option in the Program Manager's menu. In Windows 95, Press "**Start**" in the Taskbar and select "**Run...**"

4. Type "A:INSTALL" if the drive is A: and press <**ENTER**>.

5. The installation program is now started. An input box shows the directory where Molecular Analyst Fingerprinting software will be installed. The default path name **MA-F** may be changed if desired.

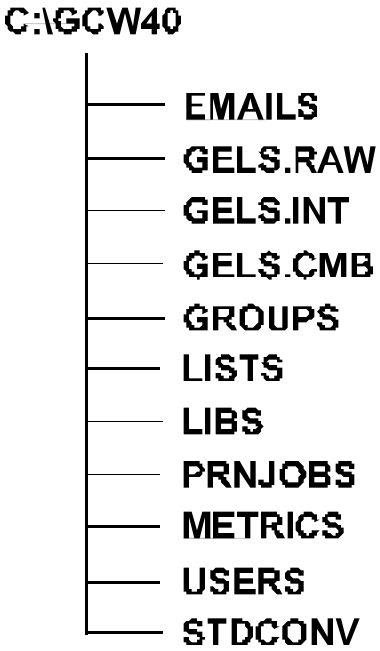
6. Press the <***Start***> button. The program now starts the installation.

7. After completion, remove the installation disk from the drive.

The installation program has created a program group called "Molecular Analyst Fingerprinting software 1.12" containing these icons:

The "Molecular Analyst Fingerprinting software" icon loads the Startup program, from where all Molecular Analyst Fingerprinting software applications can be run. The "About version 1.12" icon contains important last minute information in Microsoft Write™ format.

The installation program has created the following directories on the hard disk (for the default Molecular Analyst Fingerprinting software directory C:\MA-F):



The subdirectories GELS.RAW and GELS.INT are intended to contain the "raw" (not normalized) and "interpolated" (normalized) patterns, respectively. It is not necessary to store the gels in these directories; if you wish to do so, you can make your own ones. The same applies to GELS.CMB, intended for combined gels, LISTS for selection lists and LIBS to store libraries.

The subdirectories PRNJOBS, METRICS, USERS, EMAILS, and STDCONV are fixed and should not be renamed or removed.

## 2.2 Installing the Molecular Analyst Fingerprinting Network Software

### 2.2.1 Introduction

The network version of Molecular Analyst Fingerprinting software supports several network transport protocols which include Netbios, IPX/SPX, TCP/IP, and Named Pipes. The system consists of three components: the *security server*, the *security key* (dongle) and the *client software*.

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

The **security driver** is a program that is available on every computer where the Molecular Analyst Fingerprinting network software is installed. This program manages multiple licensing over the network. It should be loaded on the server computer in the network, i.e. where the security key is attached.

The **client software** is a Molecular Analyst Fingerprinting 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 a the same time.

### 2.2.2 Setting up the network software

First, select a computer in the network as server. In case the network is very long, or contains many computers, it is recommended to choose a centrally located device. Further, it is important to have a stable computer in terms of hardware and software, that is permanently working and available over the network by other computers.

#### 2.2.2.1 Installing the programs

Once the server computer is chosen and the network key is connected to the

installed on the server first, and can then be installed on any other computer connected to the network. To install Molecular Analyst Fingerprinting software on the server, follow the guidelines as explained in 2.1 for the single user software. After completion of the installation, the same program group is created as described for the single user version, except that it contains one additional icon: the *Network Key Server* program.

#### 2.2.2.2 The server program

Once the security key is attached to the parallel port of the server computer, the program "Network Key Server" in the Molecular Analyst Fingerprinting software group (program NSRVGX.EXE) should be loaded on the server. It is perhaps a good idea to have the server program NSRVGX.EXE loaded automatically when the operating system starts on the server computer. This can be done easy using the Windows Notepad program. In Windows, click "***Start | Run...***" or "***File | Run...***" and type "notepad \windows\win.ini" <ENTER>. Then, add the line "Run=C:\MA-F\NSRVGX.EXE" and save WIN.INI.

While loading, this program searches the network for a while and automatically determines the available transport protocol(s). Once this has happened, it minimizes as icon and starts managing the Molecular Analyst Fingerprinting software licensing over the network. It can be popped up to show a table with the available network protocol and the securtity key(s) detected. It also shows the maximum number of sessions allowed, the current number of sessions in use, as well as the peak number registered.

> ***NOTE****: If the server table doesn't show the network detected in the header in "Server names", or doesn't show a key detected in "Key1", this means that the protection key was not found. This is the case if the network is not recognized, not functioning properly, or if the key is not connected.*

You should not close the network key server program, because in this event, the active sessions, i.e. Molecular Analyst Fingerprinting software programs in use, will loose their license permit and will terminate. When you try to close the program or shut down Windows, it will warn when there are still users active.

#### 2.2.2.3 Running Molecular Analyst Fingerprinting software

Once the server is running properly as describe above, you can install and run Molecular Analyst Fingerprinting software on any computer, including the server computer, that is connected to the network using a compatible protocol. When the license limit number is reached, it will not be possible to open

another Molecular Analyst Fingerprinting software session until one of the active users terminates his Molecular Analyst Fingerprinting software session and thus a license comes free.

If a Molecular Analyst Fingerprinting software session is terminated in an illegal way, e.g. by rebooting the computer, restarting Windows or by a shutdown, the license of that session will not terminate on the server. Inactive licenses are made available again by the server program if a *Strict Time-out* (ST) command is specified while loading the server: **NSRVGX /ST**. By default, Molecular Analyst Fingerprinting software sets a time-out of 30 minutes on the server.

While running, the Molecular Analyst Fingerprinting software network programs check for the presence of the network dongle at regular times when operations are done. In this way the programs verify if the license granted is still valid, and avoid ending of the license by the time-out implied by the server. When the network is temporarily overloaded or unavailable, the Molecular Analyst Fingerprinting software programs will produce a warning message “Security key not found - Check network status”. At this moment, it is recommended to save your work, e.g. a list in the Analyze program. You will be able to work further until the program checks for the security key or the license a next time. If the license is still not recognized, the program halts with a fatal error.

# 3. The basic principles of Molecular Analyst Fingerprinting software for Windows

Before going into a detailed description of the possibilities of each program or module, this section gives a brief overview of some important principles.

## 3.1 The programs

The software consists of 5 programs:

1. The **Startup program.** This program is run by double-clicking or pressing **ENTER** on the "Molecular Analyst Fingerprinting software icon" in the Molecular Analyst Fingerprinting software 1.12 program group. This program shows the intro screen with version number and license site information. It allows you to run the Molecular Analyst Fingerprinting software server applications listed below, to setup and select multiple users and to customize various settings (colours, directories, labels, etc.) for each user. Use the "***Exit***" button when you are finished running the Molecular Analyst Fingerprinting software applications.

2. The **Conversion program** converts gel images from a particular scanner to the standard track format recognized by the Normalization program. This program is loaded using the "***Convert***" button in the Startup program.

3. The **Normalization program.** This program is used for normalization of raw track data coming from the conversion program. The resulting files (which have the extension ".INT") can be saved into a database and processed by the main Molecular Analyst Fingerprinting software program. This program is loaded using the "***Normalize***" button in the Startup program.

4. The **Main program.** This is the most extensive program, including all database construction and management functions, search, editing and imaging tools, the printer driver and the optional modules Cluster Analysis, PCA & Statistics, Identification, Quantification, and Database Sharing. This program is loaded using the "***Analyze***" button in the Startup program.

5. The **Diagnostics program.** This program analyzes all the system and hardware features of your computer and detects and reports any errors. It also checks the Molecular Analyst Fingerprinting software drives and directories of the current user and inspects the validity of all the files contained therein. This program is loaded using the "***Diagnose***" button in the Startup program.

## 3.2 Multi-user system

In order to facilitate the use of the Molecular Analyst Fingerprinting software system by more than one person and/or for more than one application, it is possible to set up a list of users, having their own settings, and defining their own directories, databases and libraries. In this way, there is now interference between information and settings of different users. **The settings of different users will only be unique when their IMAGE and DATABASE directories are different!** If a new user is defined in the User Setup (Startup program), Molecular Analyst Fingerprinting software automatically assigns unique directories to it, if the question "*Automatically create user directories*?" is answered with yes.

## 3.3 Storage of normalized gels

All tracks of a particular normalized gel are stored in one file with an extension "**.INT**". Each track is described by 7 information fields. The labels of these fields can be changed in the **Startup** program. In addition, Molecular Analyst Fingerprinting software stores all band information of each gel in files with the same name as the gels, but with the extension "**.PKS**". The band data can be used for automatic band searching, for clustering and for quantification. The normalized gelstrips for each gel are stored as "**.I2**" files. If a two-dimensional quantification of a gel is done, this is stored in a "**.QNT**" file

*NOTE: Do not create subdirectories with the extension .INT in your database directory.*

## 3.4 Database management

A set of gel files (".INT" files), normalized against the same "standard reference" pattern and stored in the same directory is called a Molecular Analyst Fingerprinting software **database**. The default database directory after installation is "**MA-F\GELS.INT**".

Most of the Molecular Analyst Fingerprinting software functions, such as grouping analysis, identification, polymorphism analysis etc. apply to so-called **lists**. Lists are selections of gel tracks from one or more gels or databases. They are like *queries* in a database program. Molecular Analyst Fingerprinting software offers powerful search routines to simplify the construction of these lists. Lists can be stored for use at another time.

## 3.5 Printing results

Molecular Analyst Fingerprinting software allows colour, halftone or black-and-white hard copies to be created of virtually every result by simply selecting the "PRINT" option in the window menu. However, to make printing even more flexible, this action causes the result not to be printed out immediately, but creates a print job file instead. This preserves you from having to wait each time a graphical report is generated. A print job can be printed at any time and on different printers using the **Molecular Analyst Fingerprinting Printer Manager**. This unit lists the print job queue and allows you to show a preview of each particular job, to print it out on one of the many printers supported by Windows or to export it to other Windows applications (such as word processors, spreadsheets, DTP programs...). A very useful feature is the possibility to print all jobs from the job list with one single command, which can be done after a Molecular Analyst Fingerprinting software session. The Molecular Analyst Fingerprinting software Printer Manager can be easily called by pressing function key F8, or by selecting "***Molecular Analyst Fingerprinting software Printer Manager***" from the item "***System***" in the main menu. See section 9 for detailed description of the Molecular Analyst Fingerprinting software Printer Manager.

## 3.6 Graphical user interface

This manual supposes you are familiar with the Windows user's interface. Refer to the Windows manual for more information about resizing windows, menu selection, the various input possibilities of dialog boxes, the use of scroll bars to scroll through partially displayed images, etc. Since Molecular Analyst Fingerprinting software takes full advantage of the multi-windowing environment, the user is allowed to display many windows simultaneously, even windows of the same kind. This offers a great flexibility, for instance to show, edit and compare multiple gel windows, images or dendrograms at the same time and cut and paste information from one window into another. However, it requires at the same time that the user knows how to manage complex applications and overlapping windows. To facilitate working with multiple windows, "child" windows (descendant from another window) will always remain on top of their parent, whereas the parent window will not be disabled, except for dialog boxes.

Molecular Analyst Fingerprinting software makes use of the available multitasking features in Windows 3.1 and later, by running time-consuming calculations in background. For example, during an extensive grouping analysis, which may take several minutes, Molecular Analyst Fingerprinting software would accept other commands, as long as they do not interfere with the running operations. For example, if a dendrogram is calculated of a list of 800 patterns, Molecular Analyst Fingerprinting software would not allow you

to modify that list as long as the calculations are going on. It is also possible to switch to another application while Molecular Analyst Fingerprinting software is calculating.

**In order to run Molecular Analyst Fingerprinting software with success it is an absolute requirement to have at least some understanding of the DOS directory system, file naming and management. For example, you should know that DOS filenames can have maximally 8 characters, without reserved characters such as spaces, question marks, or periods. In addition, the user should be familiar with the windows user interface, program manager, multi-application control etc. to take full advantage of the Molecular Analyst Fingerprinting software features and possibilities.**

*NOTES:*

*All images such as patterns, dendrograms, curves etc. are rescalable in Molecular Analyst Fingerprinting software. If you are using a large screen (17" or 20") in a high resolution mode, you will benefit from this feature to reveal more details and/or have a better overview.*

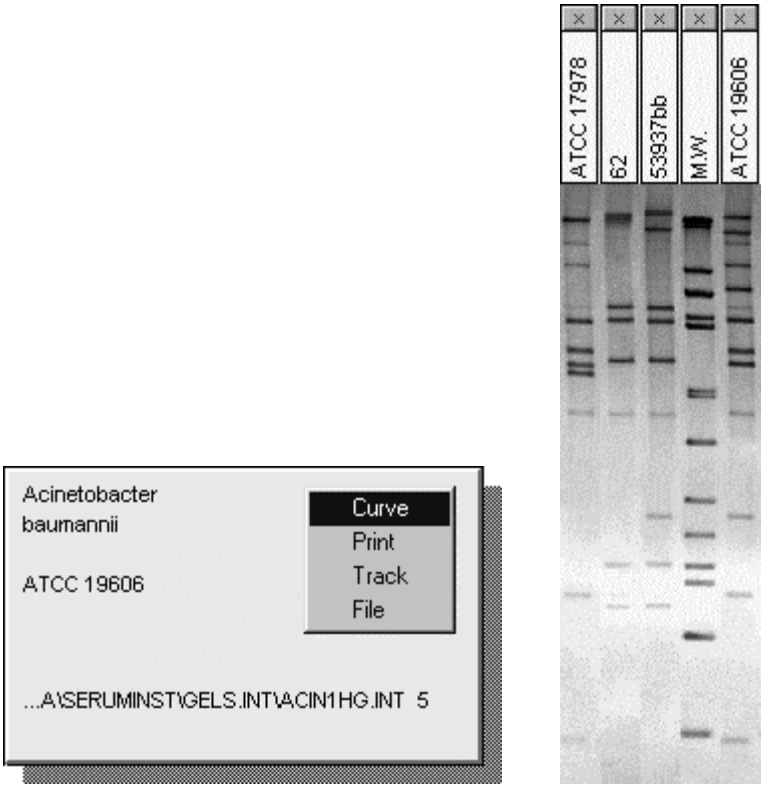
*Although most Molecular Analyst Fingerprinting software functions are accessible from the keyboard, a mouse is necessary for some advanced features.*

In the following sections, menu options will be written in ***bold-italic***, with a vertical separation line between the successive items. For instance, ***File|Load*** stands for the ***Load*** item which can be found in the ***File*** pull down menu. Buttons also will be noted in italics, but between inequality characters. For example, **<*OK*>** means the OK button. Each time a new object from the Molecular Analyst Fingerprinting software program is explained, the name will be written in *italic*. For instance, when the window showing the track information is described for the first time, its name is written in italic: "*Information card*".

## 3.7 Instant detailed information of tracks in all applications

In many windows which display gel tracks, pointing a track name with the mouse arrow and pressing the left mouse button while the CTRL key is held down pops up an information box listing all important information of the selected track. In addition, a pop-up menu allows you to show the densitometric curve, to show a "***Gel strip***", to create a print-out of the track information, or to call the gel window with the track being selected. The information box disappears by any action outside the pop-up menu or by pressing ESC. Further on in the manual, this option is called "***info-box***".

The **Gel strips**, which can be called from the info-box, are small windows, each of which displays one particular pattern. They can be created in the same applications that allow you to show info-boxes by simply pressing SHIFT + clicking left mouse button on a track. The Gel strips behave like separate windows and can be moved and rearranged by moving them over the screen. They can be removed by pushing the small button on top of the strip. Several functions are available to edit the gel strips using the menu item ***Gelstrips***. The colour palette can be adjusted/changed with ***Gelstrips|Palette*** (see 6.3.3.4 and 8.2.3). The labels of the strips can be changed between the gel name, organism name and entry code with ***Gelstrips|Labels*** (see also 8.2.4). All gelstrips can be removed with ***Gelstrips|Remove all***. ***Gelstrips|Rearrange*** causes all strips to fit next to each other, respecting the order in which they were rearranged. With ***Gelstrips|Add to list***, all strips shown are added to the selection list (see 6.3.2). ***Gelstrips|Show bands*** will mark or unmark the bands that were defined on the patterns. With ***Gelstrips|Size*** you can customize the length of the strips in vertical direction according to your screen resolution or original pattern length. An input box allows you to enter the number of pixels in vertical direction.




***Information box*** with floating menu (left) and five aligned ***Gelstrips*** (right). The information box is popped up with CTRL + left mouse click, a gelstrip using SHIFT + left mouse click. Gelstrips can be dragged and moved individually and remain always on top of any window.

## 3.8 Hot keys

Some functions which you will often need running Molecular Analyst Fingerprinting software are directly accessible using a hot key, which is one of the function keys F1 to F8.

**F1** pops up a context-sensitive help. The help gives an overview of all menu functions of the active window and a list of related topics that describe the functions.

**F2** to save the current selection list of tracks on disk (see 6.3.3.1).

**Ctrl+F2** to save the selection list of tracks on disk (see 6.3.3.1) calling the "Save as" dialog box..

**F3** lets you load a list from disk.

**Ctrl+F3** lets you append a list from disk to the current list.

**F4** clears the current list in the computer's memory.

**F5** calls the search window (see 6.3.3.1).

**F6** allows the colour palette to be changed when pattern images are displayed.

**F7** lets you change the band settings (see 6.6.1.2 and 6.11.1).

**F8** calls the Molecular Analyst Fingerprinting software Printer Manager (see also 3.5 and 9).

**F9** calls the Entry description dialog box to customize names in images, dendrograms etc. (see 8.3).

**Ctrl+C**: Run cluster analysis of the current list based on Pearson Correlation.

**Ctrl+B**: Run cluster analysis of the current list based on Band matching.

**Ctrl+A**: Calculate a Principal Components Analysis of the current list.

**Ctrl+I**: Identify the current list.

**Ctrl+Q**: Compose a comparative Quantification of the current list.

Special hot keys are available in the image conversion program, which are described in section 4.6.

# 4. Conversion of gelscans into Molecular Analyst Fingerprinting software format

## 4.1 Introduction

When a gel is scanned using a densitometer, flatbed scanner or video-camera, a TIFF image file is usually created, containing the rastered image. However, Molecular Analyst Fingerprinting software uses one-dimensional densitometric arrays of each separate pattern for comparative purposes. The conversion program is designed to mark the different tracks on the image, calculate the densitometric curves, extract the "gel strip" for each pattern and provide the patterns with descriptive information. Each track on the gel is defined by using a "spline" which can be moved through the gel image and distorted in order to optimally follow the shape of a particular track.

In case you are running Molecular Analyst Fingerprinting software with a scanner producing bitmaps as **TIFF (Tagged Image File Format) files**, such as flatbed scanners and video cameras, your Molecular Analyst Fingerprinting software package can edit the TIFF files and define tracks and gel strips on the gel image. The operation of this program is explained in section 4.2.

In case your scanner produces **one-dimensional line-tracks** as densitometric records, you should read section 4.3, explaining how you can convert the scanning files into Molecular Analyst Fingerprinting software format. You will not be able to display gel strips in the main program, since this option requires a TIFF or bitmap file as input.

## 4.2 Using the Image Conversion program

### 4.2.1 Loading an image file

Press the <***Convert***> button in the Molecular Analyst Fingerprinting software Startup screen to load the Conversion program. The *conversion main window* shows a list of tracks (which is initially empty) and the information input fields. Selecting ***File|Load image*** opens a file dialog box, showing the directories on the hard disk and the files in the current directory. The current directory is the default image directory C:\MA-F\IMAGES at installation, or the one you have specified for gel images in the User Setup (see 8.2.1). The *Image size* check box lets you choose between ***Normal*** and ***Fit in window***. When the option Normal is checked, large gels will not fit within the screen and you will be able to scroll through the image. When the Fit in window

option is checked, all gels will be rescaled to fit within the window, at least when the window is maximized to cover the entire screen. When the option ***Negative*** is checked, the image will be inverted. Select the desired TIFF file and press **<OK>** to load the file (since bitmap images are usually large files, this may take some time). An *image window* appears, showing the image of the gel as it is provided by the scanner. The horizontal (Xres) and vertical (Yres) resolution of the image are written in the status bar.

*IMPORTANT NOTES:*

*(1) In order to obtain a realistic image of the gel, the graphics adaptor should operate in 256 colours mode.*

*(2) The gel should always be scanned so that the patterns appear vertically.*

*(3) Only **non-compressed** TIFF files in IBM-PC format will be recognized by the program.*

*(4) The gels should always be displayed with dark bands on a light background. If this is not the case (e.g. for ethydium bromide stained gels), the image should be inverted (negative) and reloaded.*

*(5) If you want to define tracks using the automatic search systems, choose preferably the Normal mode (not Fit in window).*

### 4.2.2 Delineating the gel

*NOTE: When you have enabled the "Doublegel Normalization" option in the User Setup (8.2) the first menu will look differently (see further in this section, paragraph 4.5).*

A green rectangle is shown on the image window, containing 8 nodes: one in each corner and one in the middle of each side. There are two classes of mouse actions to border the gel: (1) pressing the left mouse button only to resize the rectangle, and (2) pressing SHIFT and the left mouse button together to distort the rectangle into any shape.

**a) Resizing the rectangle.** By moving the mouse pointer to a node, the pointer's shape will change to show which action the node serves for. You can drag the nodes while pressing the left mouse button. By dragging the upper left node you can displace the rectangle as a whole. With the bottom right node you change both the horizontal and vertical dimension of the rectangle. The upper and bottom middle nodes will only change the vertical dimension of the rectangle whereas the left and right middle nodes will only change its horizontal dimension. The upper right and bottom left nodes have no resizing function.

**b) Distorting the rectangle.** When the SHIFT key is held down, the four corner nodes can be used to distort the rectangle into any tetragonal shape. With the SHIFT key held down, the nodes in the middle of the sides can be dragged to distort the sides into curves. To even better follow the contours of the gel, the nodes can be dragged away from the middle of the side which causes the curve to become asymmetric.

The easiest way of working is to first border the gel by the corner nodes, either by resizing the rectangle or by changing the tetragon shape. The next step, if necessary, is to distort the sides to border curvy gels.

In the menu of the image window you can select ***Limits|whole image*** to resize the rectangle to cover the whole image or ***Limits|Auto search*** to let the program automatically search the contour of the gel. The latter function will only work successfully with clear-cut gels.

***NOTES****:*

*(1) The success of normalization (see further, section 5) depends on the consistency of the delineation of the gels. You should try to find easily recognizable markers at the top and the end of the reference patterns and set the upper and bottom borderlines based on the positions of these markers, so that all bands are still included. If the gel contains an easy marker at its top (for instance, a high molecular weight band on the reference patterns) you will benefit from always including this marker by placing the upper borderline of the rectangle a few millimeters higher.*

*(2) The shape of the bordering rectangle or tetragon is saved upon exiting the program. This saves work when a standardized electrophoresis and scanning procedure is established.*

### 4.2.3 Optimizing the image

It often happens that the 256 gray levels are not optimally used in the TIFF file to represent the gel. This may result in a (dark) gray background, and/or weak-looking bands. Molecular Analyst Fingerprinting software offers a function to recalculate the TIFF image so that the full range of 256 gray levels is used by the TIFF file. The result is that the clearest background region of the gel will be white, whereas the darkest band center of the gel will be black. Optimization has the advantage that different gels are better standardized for brightness and contrast. In other words, if gelstrips of patterns from several gels are shown together, they will look more similar.

The optimization is based on the lightest and darkest OD values within the green bordering rectangle. Thus, before optimizing the gel, select the region of interest by setting the green bordering rectangle as explained before (4.2.2). Then, select the menu item ***Edit|Optimize image***. The Optimization menu item

is indicated as active when it is marked with ✓. To undo optimization just select ***Edit|✓Optimize image*** again, and the original brightness and contrast will be restored.

### 4.2.4 Track scanning settings

Before explaining how to define tracks on the image we will describe the different settings. With ***Edit|Palette*** you can optimize the brightness and contrast of the image and choose one of the predefined Molecular Analyst Fingerprinting software palettes or custom palettes (see 8.2.3). When ***Edit|Show zoom box*** is enabled, a square box showing a zoomed part of the gel is popped up each time the left mouse button is pressed.

***Edit|Settings*** displays the *Track scanning settings box*, which contains four parameters for scanning tracks:

(1) The ***Number of tracks*** relates only to the automatic lane-finding algorithms and should reflect approximately the number of lanes on the gel. This number has no importance when tracks are defined manually.

(2) The ***Track resolution*** determines the number of points each densitometric track will consist of. This number may be less than or equal to the Y-resolution of the gel (indicated at the bottom of the image: Yres) but should not exceed this value! Otherwise the program would introduce data points between the original values derived from the image. Although this is not a critical error, it is better to allow the program to reduce the original data than to introduce data by interpolation. It is necessary to keep this resolution value always constant, i.e. the same as the resolution of the standard reference pattern of your database (see further, section 5).

(3) The ***Spline thickness*** determines the number of points in lateral direction of a tracking spline (see 4.2.5) that will be averaged in order to obtain a more stable profile. The value indicates the number of averaging points at either side of the centre of the spline. When a tracking spline has a certain thickness, it is drawn as two vertical lines indicating the borders of the averaging zone. Obviously, the spline thickness should never exceed the thickness of the patterns.

> *NOTE: When the patterns are distorted or suffer from some "smiling effect" we recommend to define a rather small spline. However, it is not advised to define a zero thickness!*

(4) The ***Gelstrip thickness*** determines the number of points in lateral direction of the gel strips that will be cut out from the image. The value indicates the number of points at either side of the centre of the spline. Two blue vertical lines indicate the borders of the gel strips. The thickness should be chosen so

that the blue splines cover the complete patterns but not the neighbouring patterns. Ideally, the borders of neighbouring splines should fall togehter.

(5) The ***Curve smoothing*** factor is only useful to smoothen gel images with high background noise, and to convert images scanned in very high resolution when the defined track resolution (see item 2) is much lower. **In normal cases, this factor is be set to zero**, since the Normalization program performs curve smoothing which is recommended to use rather than this one.

(6) The ***Number of nodes*** determines the default number of distortion nodes which will be assigned to each spline when tracks are added either manually or automatically. This number should not be taken larger than strictly necessary. For example, with a spline consisting of four nodes, the automatic search algorithms will be able to follow the curvature of most gels perfectly. When the scroll bar of this parameter is moved completely to the left, the indication *Straight* appears instead of the number of nodes. This means that the automatic search algorithms will not apply any distortion for the nodes. For good-looking and well-scanned gels, this setting can be convenient. If one or more splines should be distorted, it is possible to add nodes manually for each particular spline.

(7) The ***Rescaling*** algorithm can be disabled by checking *None*. *Whole gel* means that all tracks will be rescaled identically according to the highest and the lowest densitometric value found among all tracks of the gel. With *Each track* checked, the program will rescale each track separately according to the highest and lowest densitometric value within the track.

(8) As automatic ***Track search algorithm*****,** two systems are available: *System I* and *System 2,* each having their own specialization. System I searches mainly for the most stable zone within each track, making it possible to correctly find the patterns on gels with very few or no space between the lanes (for instance, sequencing gels). System II is rather based on the overall optical density of the lanes and as such is better suited for gels with more space between the lanes. For normal gels system II will provide the most satisfactory results.

### 4.2.5 Marking tracks on the image

There are three possibilities to define tracks on the image: (1) by adding and positioning the tracks manually, track by track; (2) by adding a group of tracks, and (3) by automatic lane-finding.

**a) Manual addition of tracks.** Select ***Track|Add*** from the menu to add a spline to the image. Position the mouse pointer on one of the nodes and press the left mouse button to drag the spline to the desired gel track. Press SHIFT + left mouse button to distort the spline in one of the nodes.

**b) Adding a group of tracks.** Choose ***Track|Add group*** from the menu and enter the exact number of lanes on the gel in the input box which appears. The program will equally divide the defined gel contour into the number you have entered. In case the sides of the bordering rectangle were distorted, the spline shapes will fluently turn from the curvature of the left side to the curvature of the right side. This algorithm is very powerful provided that the contour of the gel is defined exactly. If the splines do not match the lanes exactly, you can correct this by moving or distorting the (green) borders of the gel until the matching is optimal.

**c) Automatic lane finding.** The program will automatically search for tracks when option ***Track|Auto search*** is called. The tracks are searched using the algorithm specified in the track scanning settings (see 4.2.4). The number of tracks specified in the settings is also important to a certain extent. For instance, if you have specified 10 tracks in the settings and want to scan a gel containing 20 tracks, the automatic lane finding may fail using either system I or system II. As explained in the settings, each search system has its own specialization, but it may require some trial before you establish the best suited algorithm for your type of gels. For large gel images, the automatic search systems will work better when the image is loaded in Normal mode (not Fit in window).

### 4.2.6 Editing splines

After automatically or manually adding tracking splines you may want to delete, add or modify some splines. The spline drawn in red is currently selected. Selecting another spline happens by clicking the mouse on one of the nodes of a spline. With ***Track|Add***, you add a new spline at the right side of the selected spline. Using ***Track|Remove***, the selected spline will be removed. With ***Track|Remove all***, all tracking splines are removed from the image. Press the left mouse button on one of the nodes to position the spline correctly on the gel track. Press SHIFT + left mouse button to distort the spline in one of the nodes. ***Track|Add node*** (or Page Up key) divides the selected track into more nodes, to allow finer distortions, whereas ***Track|Remove node*** (or Page Down key) does the reverse.

You can call a preview of the densitometric curve of the selected track by ***Track|Preview*** or by pressing ENTER on the selected tracking spline. The gray curve shows the non-optimized densitogram (corresponds to the "Rescaling - None" option in the scanning settings (see 4.2.4), the red curve is the optimized densitogram ("Rescaling - Each track"). Press a mouse button to remove this window.

While defining tracking splines on a quite large or complicated gel image, it is recommended to save the positions of the tracking splines that are already defined at regular times, using ***Track|Save track positions***. Later, when the

same image is loaded, the last defined and saved spline positions can be reloaded using ***Track*|*Load track positions***.

### 4.2.7 Scanning the image

When all splines are positioned correctly, you can select ***Scan*** from the image window. The program now derives the densitometric arrays from the image, using the selected borders, splines and track scanning settings. This scanning process requires some time, depending on the resolution of the image, the averaging factor, the speed of the hard disk etc. Each time a densitometric curve is "scanned", a new lane number is added to the list in the conversion main window. The same label is now shown on the tracks of the image. When finished, you can minimize the image window by pressing the ⇩ button in the upper right corner of the window and add descriptive information to the tracks. You can restore the image window to view the tracks by double-clicking on the icon that remains in front in the left bottom corner.

## 4.3 Using the track conversion program

### 4.3.1 Track conversion settings

(2) The *Track resolution* determines the number of points each track will consist of, after conversion to Molecular Analyst Fingerprinting software format. This number may be chosen less than the resolution of the tracks as they come from the densitometer BUT SHOULD NOT EXCEED THIS VALUE! Otherwise the program would introduce data points between the original values derived from the scan. This is not a critical error, but it is better to to delete than to introduce data points. It is necessary to keep the resolution value the same as the resolution of the standard reference pattern of your database (see further, section 5.2.2, item 2).

The **Rescaling** algorithm can be disabled by checking *None*. *Whole gel* means that all tracks will be rescaled identically according to the highest and the lowest densitometric value found among all tracks of the gel. With *Each track* checked, the program will rescale each track separately according to the highest and lowest densitometric value within the track.

The settings are stored on disk.

### 4.3.2 Loading track files

Running the data conversion program (**<*Convert*>** button in the Molecular Analyst Fingerprinting software Startup screen) creates the *conversion window*, showing a list of tracks (which is initially empty) and the information input

fields. Selecting ***File|Load*** opens a file dialog box, showing the directories on the hard disk and the files in the current directory. The current directory is the one that was specified for the current user in the User Setup menu (see 8.2.1). Load a gel by double-clicking on the file name of one of the geltrack files or by pressing the **<*OK*>** button.

## 4.4 Assigning descriptive information to the tracks

All tracks are initially marked with a red question mark. This means that they are not yet edited. If the automatic rescaling function was enabled (see 4.2.4, item 7), the tracks are marked with "S" Select a particular track by using the mouse or ⇧⇩ cursor keys and fill in the information fields. It is recommended to enter the most important identity label of the sample in the fourth field (the first entry code field, see section 8.2.4). This field may contain up to 12 characters and is together with the field which combines gel name and track number the most important track identifier in all Molecular Analyst Fingerprinting software reports. Check the *Reference* checkbox if the track is going to serve as reference for normalization.

When the information is entered correctly, press ENTER or double-click on the selected track. When the track was rescaled, it is marked with "S" and in case *Reference* was checked, it is labeled with "R". The question mark disappears which means that the track has become edited. When you now select another not yet edited track (marked with red question mark), it will adapt the information fields from the previously edited track. This function lets you fill in the descriptive information of similar tracks very quickly. The obvious reason for this option is that one can efficiently edit tracks in batches of those having similar information fields.

After you have pressed ENTER for a given track to assign the information written in the fields at the top of the window (red question mark disappeared), it is still possible to undo this assignment with the menu command ***Edit|Initialize track*** or by pressing the HOME key. The information fields of the track become all empty, the question mark reappears and you can adapt information from any other track, if desired, or enter new information.

It is possible to paste all information from a previously created gel file by using the ***Edit|Paste Gel info*** menu option. The only condition is that the number of tracks from the gel copied into the Molecular Analyst Fingerprinting software clipboard and from the current gel in the conversion program are the same. Similarly, it is possible to paste the information from one particular track into the currently selected track with ***Edit|Paste Track info***. For detailed explanations on how to copy gel and track info into the Molecular Analyst Fingerprinting software clipboard, see section 6.3.1.

***NOTE**: The information of the last selected track is adapted for unedited tracks. In case you are editing tracks in batches, it is recommended to move the selection bar from one track to another using the mouse, NOT the arrow keys, since in the latter case you may have to scroll through edited tracks having other information filled in.*

### 4.4.1 Manual rescaling of tracks

Call ***Rescale*** from the menu to rescale a track in the *Rescale* window. This window displays the densitometric curve of the selected track. Depending on the track scanning settings (see 4.2.4 or 4.3.1), the ***Rescale*** item in the menu will be enabled or not. When the automatic rescaling function was disabled, the tracks in the main window are not marked with "S" and the ***Rescale*** menu item will be enabled. An upper and bottom horizontal line is shown on the image. Drag the lines to the highest and lowest relevant densitometric points, respectively, and press ***Rescale***. If a track is rescaled either automatically during scanning, or manually in the rescale window ("S" shown in main window), select ***Initial*** from the menu if you want to restore the original scaling of the track. Press ***Exit*** to close the rescale window.

### 4.4.2 Saving the converted gel

***File|Save gel*** from the main window allows the defined tracks to be saved into an unnormalized (raw) gel file. A dialog box appears, showing an input field for the filename and a directory list box. The default directory is the path specified in the User Setup menu for a given user (see 8.2.1). **Do not select another directory unless you have special reasons**: selecting another directory would prevent the normalization program from processing the TIFF file for preparing gelstrips. Enter a DOS-filename as desired (without extension!) and press **<*Ok*>** to save.

## 4.5 Doublegel conversion

In some electrophoresis systems, it is possible to mix a set of markers with each sample and visualize these using a different probe or staining dye. Typically, this will result in two scanning files of the same gel, one of which showing the patterns, and the other the internal marker bands. As an extension to the conventional gel conversion and normalization functions, Molecular Analyst Fingerprinting software allows gels to be normalized through such internal reference marker bands scanned into a separate TIFF file. This special method of normalizing is called here the "Doublegel" conversion and normalization. The feature can be enabled by selecting the checkbox "***Doublegel normalization***" in the User Setup menu for the user which is active (see 8.2.5).

## 4.5.1 Coinciding conversion of tracks and internal references on separate scanning files

The following criteria should be fulfilled in order to make the "Doublegel conversion" possible:

1) At least two marks should occur at the same position on both scans. By preference, these marks should be made at the outermost edges of the gels. They will allow the program to position both scans onto each other.

2) The complementary gel plates should be scanned using the same scanning resolution.

3) The marks on both scans should be fine and sharp so that their position can be accurately determined.

The scanning file containing the reference lanes will be called the ***primary file***. The scan containing the patterns to be analyzed will be called the ***secondary file***.

### 4.5.1.1 Loading primary file

In a first step, the scan containing the internal reference markers, i.e. the primary file, is loaded in the usual way using the ***File|Load image*** command. The image window displays a ***Marker*** menu and asks you to "mark control spots on gel image". To define a marker spot, click and hold down the left mouse button in the close proximity of the marker spot on the image. A popup window shows a zoomed detail of the selected area. Still holding down the left mouse button, you can exactly define the center of the spot by moving the mouse pointer and releasing the button at the correct position. A red encircled cross points the marked spot.

Instead of releasing the mouse button inside the zoomed window, you can move the arrow outside the zoomed area and then release the button to undo selection of the spot.

The menu option ***Marker|Delete last*** allows the last defined spot to be deleted. Choose ***Marker|Ok*** to confirm the marker spot setting. The program enters now in the normal conversion mode as explained in 4.2.5 and further.

### 4.5.1.2 Converting the primary file

After scanning tracks on the image, some minimal information should be entered in the conversion main window (track listing). Select every fifth pattern as "Reference" (or depending on the internal distortion of the gel, less or more). Mark only good-looking patterns as reference.

Save the primary file as explained earlier. In saving the file, try to give a name which allows you to easily recognize it as a primary gel (containing only reference patterns). For example, if the gel name is PF1608, one could add "P" to the name to indicate that it is a primary file. After saving the gel, **DO NOT EXIT THE PROGRAM.**

#### 4.5.1.3 Converting the secondary file

Now select ***File|Load image (secondary)*** to load the secondary file. The ***Marker*** menu again allows you to define the corresponding marker spots on the secondary gel, as explained for the primary file. When the marker positions are confirmed using ***Marker|Ok***, the program enters the normal mode (see 4.2.5). Exactly the same bordering rectangle and track splines are now shown as defined for the primary gel, except that they are rotated and shifted to compensate for the positional differences of the marker spots on both files. When the secondary gel is rotated with respect to the primary gel, the tracks will be rotated accordingly, thanking this correction to the alignment of the spots. **DO NOT RESIZE OR DISTORT THE BORDERING RECTANGLE OF THE SECONDARY GEL!** The track positions can be slightly corrected if necessary.

*NOTE: It is now clear that the accuracy of the positioning of both files will depend on the sharpness and definition of the marker spots, and also on the number of such spots applied. The latter is particularly true because Molecular Analyst Fingerprinting software automatically calculates thet average of all spot positions defined in case the positions on both scans differ slightly.*

Scan the tracks and enter the track information as explained previously. We recommend that a fixed pattern be run in one lane of each gel for quality and reproducibility control. Check that track as ***Reference*** and save the secondary gel. Similar as for the primary gel, try to give a name which allows you to easily recognize it as a secondary gel (containing the patterns under study). For example, if the gel name is PF1608, one could add "S" to the name to indicate that it is a secondary (or study) file.

## 4.6 Hot keys in the Conversion program

### 4.6.1 Main window

| | |
|---|---|
| F1 | Call help |
| F3 | Load image from disk |
| F2 | Save gel |
| ⇧⇩ keys | Select another track to edit |

| | |
|---|---|
| ENTER | Apply the current information to the selected track |
| HOME | Remove all information for selected track (initialize) |

### 4.6.2 Gel image window (Image Conversion program only)

**a) Settings.**

| | |
|---|---|
| F6 | Change colour palette |
| F7 | Change track scanning settings |

**b) Gel border commands.**

| | |
|---|---|
| F8 | Border whole image |
| F9 | Find gel borders automatically |

**c) Tracking spline commands.**

| | |
|---|---|
| SPACE | Add tracking spline right from selected one |
| DEL | Remove selected spline |
| HOME | Remove all splines |
| F4 | Add group of splines between the borderlines |
| F5 | Search tracks automatically |
| ⇦⇨ keys | Select previous/next spline |
| PgUp | Add node to selected track |
| PgDn: | Remove node from selected track |
| | |
| F1 | Call help |

# 5. Normalization of patterns

## 5.1 Principles

Proper normalization of gels is of primary importance to achieve reliable databases of patterns. The Molecular Analyst Fingerprinting software normalization program is a powerful tool for alignment of patterns, but much of the success depends on the quality of the gels, and the care and strategy of the user. The normalization program contains a number of user-defined parameters which can be adjusted to work optimally for a given electrophoresis system.

**It is recommended to read the introduction of this manual for a general explanation of the normalization principles and further to read this chapter carefully and spend some time while determining the best configuration settings with your electrophoresis system**.

### 5.1.1 Normalization by reference tracks

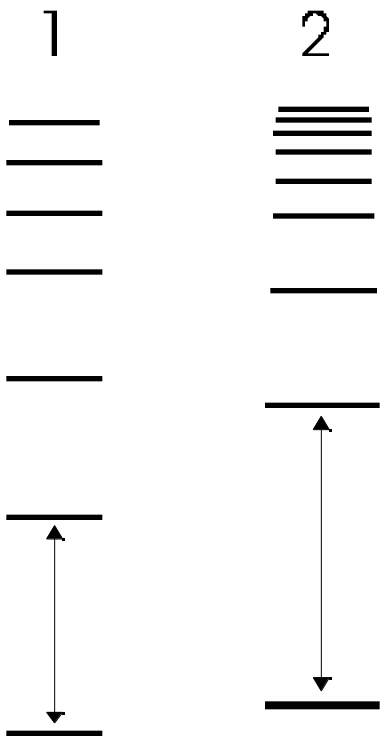
A basic requirement for proper normalization by Molecular Analyst Fingerprinting software is that a *reference sample* be loaded on each gel. The choice of a suitable reference sample is of crucial importance. The reference sample should provide a banding pattern, which meets the following conditions:

(1) The pattern should consist of clear and sharp bands;

(2) The bands on the reference pattern should cover as much as possible the entire pattern area, or at least the important range of the patterns.

(3) The reference pattern should be stable and reproducible; a sample should be chosen of which the continuity and the stability in function of the time is guaranteed.

(4) In general, the more bands the reference pattern contains, the more accurate normalization will be achieved. At least should the bands be regularly spread over the pattern. It is actually the distance between the two most distant subsequent bands that counts. In the figure below, pattern (1) is to be preferred although it contains only 7 bands whereas pattern (2) contains 9 bands. The bands in pattern (1) are better spread.




Below we give an example of how the reference sample can be applied to a gel with 20 wells.

(R: reference track; N: non-reference track)

N R N N N N N N R N N N N N N R N

As distortion is usually most prevalent in the outer wells, we have loaded the references in the second wells. In case distortion in the outer wells is very high, they can be left free, e.g. as follows

(-: free well)

- R N N N N N R N N N N N R N N N N N R -

Obviously, the number of references to be loaded on a gel depends on the reproducibility and distortion of the electrophoresis system used and on the number of tracks on each gel. In the second example, four reference samples have been applied. When only one reference patterns is available on a gel, all the patterns will be aligned according to that single reference pattern, and no compensation for internal distortion of the gel will be possible.

The next step is to choose one suitable reference pattern from a representative gel that will serve as *standard reference*. As explained before, once a pattern of the reference sample is defined as standard reference (hereafter named *standard*), all other reference patterns are aligned to that single pattern. Clearly, the standard should be a proper and representative reference pattern.

Normalization of a gel is achieved by aligning the bands of all reference patterns on that gel to the corresponding ones of the standard. Non-reference tracks are interpolated gradually according to both surrounding references. In case only at one side a reference is present, the non-reference tracks are interpolated fully according to that reference.

#### 5.1.1.1 Normalization by associating reference bands

***NOTES:***

*(1) Always try to scan the gels in a similar manner, i.e. try to find an easily recognizable marker peak in the high molecular weight zone and to start scanning a few millimeters before that marker peak. This will facilitate the alignments and will save you work.*

*(2) To facilitate the normalization process and to enhance the reproducibility, it is recommended to keep the electrophoresis conditions, the running time and the length of the gels as well as the scanning area as constant as possible.*

*(3) The Conversion settings, and in particular the resolution specified in the Conversion settings (see 4.2.4 or 4.3.1) should not be changed.*

The normalization program allows alignment of gels to be achieved by associating bands with each other, which are known to be the same on different patterns (usually *reference* patterns). By entering and storing fixed positions for each of the reference bands and aligning all the reference bands with these positions, it becomes possible to align different gels. The standardized positions (hereafter called *reference positions*) for these reference bands (hereafter called *reference peaks*) can be defined and saved together with the other normalization settings. The program is then able to align each track combined with one or more of the reference positions in order to match the positions of the internal reference peaks with these standardized positions. Tracks that do NOT contain any of the reference peaks are interpolated gradually according to the alignment information provided by the closest neighbouring patterns containing any of the reference peaks.

Alignments can be achieved in various ways: it is in the first place possible to align the reference patterns using predefined reference positions for their bands, and in the second place also non-reference patterns if these have bands in common. A reference position can be defined for each such band and the alignment by external and internal bands can be combined.

First of all, the positions of the bands from the standard reference pattern will be defined as reference positions. Once this is done, the standard reference and its reference positions are saved to disk, and the corresponding bands of all reference patterns of any gel can be aligned with these positions. In some cases

however, one or more bands are known to be common for all or most of the tracks of the database. These bands also may be suitable as reference markers for alignment of the tracks but one should be very careful when relying on such bands since attempts to align non-identical bands from different patterns would cause wrong alignments!

### 5.1.2 Background subtraction

After normalization, a non-linear background can be subtracted from the patterns. The two available background subtraction algorithms are based on curve fitting and the rolling disk principle, respectively. Automatic background subtraction as available in Molecular Analyst Fingerprinting software can significantly enhance the resolving power of the technique and is reliable provided that the settings chosen initially are not changed.

The curve fitting algorithm is based on the decomposition of the curve into a broad trend and a pattern of narrow peaks. The broad trend is calculated as a set of *Lorentzian* curves. By increasing or decreasing the number of Lorentzian curves, the strength of the background subtraction can be varied.

The principle of the rolling disk mechanism is that a disk is rolled on the inside across the curve. Every area of the curve below the imaginary trace left by the disk will be subtracted as background. The strength of the background subtraction is varied by increasing or decreasing the diameter of the disk. This algorithm warrants very stable and reliable background subtraction.

When both methods are compared, the rolling disk method is capable of subtracting more of the background and appears to be more reliable whereas the curve fitting method is much faster.

## 5.2 Using the program

### 5.2.1 Loading a gel

Press the **<*Normalize*>** button in the Molecular Analyst Fingerprinting software Startup screen to display the *normalization main window*. Use ***File*|*Load*** to load a raw gel from the hard disk. A dialog box appears, showing the different directories at the left side, and the files of a directory at the right side (the default directory for raw gels can be specified in the User Setup menu, see 8.2.1). Select a converted gel file and press **<*OK*>** to load the gel. An image of the "gelstrips" of all patterns is now shown in the main window (the default palette for staining of the gel can be changed using the User Setup menu**,** see 8.2.3, or can be changed at run-time using the F6 key).

*NOTE: When the program cannot find the TIFF file from the loaded gel, an error message appears "Error. TIFF file not present". This can have the following reasons: (1) The TIFF file has been removed from disk or moved to another directory; (2) in the conversion program, the gel was not saved to the default directory specified for "raw gels" in the User Setup. In this event, you will have to move or restore the files to the correct directories, or reconvert the gel.*

The reference tracks on the gel are marked with "R". A selection bar, which can be moved with the mouse or the arrow keys, shows the currently selected track. The subwindow at the right side shows a densitometric curve of the currently selected pattern.

***IMPORTANT NOTE:** In order to obtain a realistic image of the tracks, it is necessary to have your display in 256 colours mode.*

### 5.2.2 Normalization settings

Menu option ***Edit|Settings*** (F7 key) allows the various normalization settings to be changed in a *Normalization settings dialog box*.

*NOTE: All normalization settings are saved in the file **SETTINGS.NOR** which occurs in the database directory specified in the User Setup menu (see 8.2.1). It is a good idea to keep records of all your database settings by creating the User Information report described in 11.4 and/or by backing up the file SETTINGS.NOR on a separate disk. In this way, the settings can be restored in case the original file is corrupted, removed or changed.*

(1) Selection of a new standard.

When a reference track is selected in the main window, ***<Use currently selected>*** lets you select this reference as the new standard. This button is not present when no reference pattern is selected. In order to define the standard reference pattern, first load the gel from which you want to choose it, select the reference pattern which will serve as standard, call the Settings dialog box as explained and press ***<Use currently selected>***.

(2) Specification of the resolution.

The resolution means the number of points each track has after its interpolation. The value can be changed by dragging the corresponding scroll bar. SETTING A TOO HIGH RESOLUTION WOULD ONLY BE TIME CONSUMING FOR FURTHER CALCULATIONS AND WOULD NOT IMPROVE THE RESULTS! It is worth looking for a compromise between reduction of the number of points and a satisfactory resolution. As a general rule, a reduction to 400 - 600 points is sufficient for a 100 mm (4 inch) scan. However, this number depends on the

complexity and resolution of the patterns and the length of the tracks. In particular cases, e.g. high resolution isoelectric focusing gels of complex protein mixtures, a higher resolution (up to 1000 points/100 mm) may be significant. For long sequencing gels, even higher resolutions, e.g. 1500 points may be warranted. The maximal resolution possible for normalized single gels is 2000 points. Some scanners allow scanning resolutions that exceed their real optical resolution, by "resolution enhancement" software. It makes no sense to apply such features.

Example: When the optical resolution of the scanner is 300 dpi (dots per inch), the maximal scanning size would be ±1200 points for a 100 mm (4 inch) scan. We recommend to set the resolution in the Normalization settings to max. 1000 points and apply a smoothing factor of 1 to 3 points.

***IMPORTANT NOTES:***

*(1) In no event should the resolution of the normalization be set to a higher value than the resolution of the scanned and converted file (see Track scanning settings, 4.2.4 or 4.3.1)!*

*(2) It should be emphasized that for most 8 bit scanners (256 gray levels), the OD range is a much more restricting factor than the resolution and scanning in too high resolutions would not at all improve the final results. It is unwarranted to select the resolution as high as possible in Molecular Analyst Fingerprinting software, since this would only slow down calculation speed and unnecessarily use disk space!*

(3) Selection of a smoothing factor.

It is possible to smoothen the densitometric curve of the normalized tracks by averaging over a few points. The amount of averaging points can be changed between 1 (no smoothing) and 21. A value of 3 means that one point at either side of a data point will be averaged with the data point. The value to be chosen depends on the resolution of the patterns. For a 100 mm scan set at 500 points after normalization, we recommend 1 (no smoothing) to 3 averaging points. Higher values may be tried for simple patterns such as restriction endonuclease patterns containing few bands.

*NOTE: One should be careful with selecting too high smoothing values, since this may flatten out sharp peaks.*

(4) Background subtraction.

Select whether background subtraction should be performed or not with the corresponding checkbox. Further, choose between the ***Curve fitting*** and the ***Rolling disk*** mechanism with the corresponding radio buttons. The scroll bar next to it allows the intensity of the background subtraction to be varied by increasing or decreasing the number. If only a broad trend of the pattern needs

to be subtracted, a small value should be selected. If really all of the background is to be subtracted, increase this parameter. Typical settings are 8 to 12 for SDS-PAGE protein patterns, 12 for DNA restriction profiles, and up to 20 for IEF patterns. The rolling disk mechanism is better suited for drastic and exact background subtraction, but takes more time.

***NOTES**: (1) For normal purposes it is unwarranted to select the maximal background subtraction, since this would only hollow out and flatten broad diffuse peak!*

*(2) This parameter has no influence on the gelstrips which are always shown as on the original scanned image.*

(5) Showing 2D-gelstrips.

Enabling this feature will allow you to display the real gelstrips in the normalization program, cut-out from the TIFF file for each pattern, together composing the entire gel. This feature warrants more reliable normalization, especially since the alignments are real-time performed on the gelstrips. However, extracting the 2D gelstrips directly from the TIFF file may take some time for large TIFF files (e.g. 4-5 MB or more) and therefore, this feature is made optional.

If a line-scanner producing densitometric curves is being used, this option should be disabled.

(6) Saving 2D-gelstrips.

When this feature is enabled, the normalized TIFF strips will be saved to disk so that they can be shown in the Main program in gel images, dendrograms, gelstrips etc. The saved gelstrips use quite some disk space (on the average 200-300 KB for a 20-lanes gel), and therefore this feature is made optional.

*NOTES:*

*(1) Both features (5) and (6) can be enabled or disabled separately. It is for example possible to save 2D-gelstrips without loading them in the Normalization program.*

*(2) Once the 2D-gelstrips are extracted from the original TIFF file, normalized and saved to disk, loading them in the Main program is extremely fast.*

*(3) Extracting gelstrips from the TIFF file can be up to 100 times faster when 32 bits file access is enabled. The cache size should be chosen as large as or larger than the size of the largest TIFF files loaded. Refer to the Windows manual for installing 32 bits file access.*

(7) Space between tracks.

The scroll bar allows you to set the distance between the tracks on the screen, in pixels.

(8) Automatic rescaling.

This option will enable you to have all the tracks of the gel rescaled in the Y-direction to have the same apparent intensity. For some quantification purposes absolute differences in concentrations between tracks may be important and would be neutralized by the rescale option. Fore such applications the automatic rescale function should be disabled.

> *NOTE: This parameter has no influence on the gelstrips which are always shown with the intensity as on the original image.*

(9) Interval for correlation.

The **<*Mark correlation interval*>** button shows the densitometric curve of the database standard pattern, where the user can define a region within which the correlation between the references of a gel and the standard is calculated. This value is shown in the Normalization curve window (5.2.4.6) and is an indication of the reproducibility of the gels after normalization. The same value is also adopted for the quality control tools in the Main program, i.e. again for the calculation of the correlation between the reference patterns within a gel and the database standard (see 6.4.1).

Two vertical lines on the curve indicate the start point and the end point of the correlation region. To change the region, drag the lines with the mouse pointer (left button) to the desired position. The positional value on the curve is indicated (the resolution is that of the raw gel). The correlation region values in percent can also be seen in the Diagnostics report created in the Diagnose program (see 11.4).

Press **<*OK*>** to validate the changes made. Before saving new settings, you should realize that by changing the standard or the resolution, by toggling the background on or off, by changing the background intensity value, newly normalized gels may become incompatible or less compatible with previous ones. In general, change these settings only when you want to start up with a new database. **MOLECULAR ANALYST FINGERPRINTING SOFTWARE APPLICATIONS WILL ONLY CHECK FOR THE COMPATIBILITY OF TWO GELS IN THEIR RESOLUTION AND STANDARD SPECIFICATIONS, BUT NOT FOR THE OTHER SETTINGS**!!! The only way to verify the compatibility of two gels for their complete normalization settings is by comparing the information windows of the normalized gels (see 6.3.1).

When the database standard has been defined or changed, the new standard pattern becomes visible at the left hand side of the normalization window.

## 5.2.3 Selecting colour palettes

Use ***Edit|Palette*** (F6 key) to change the colour palette (default is the user-defined palette: see User Setup, 8.2.3; other palettes are predefined by Molecular Analyst Fingerprinting software) and to edit the contrast and brightness of the image.

## 5.2.4 Normalizing gels in practice

A small subwindow at the left hand side of the gel image shows the **reference positions** which are currently used (initially, there are no reference positions present; see below, 5.2.4.1 for defining the reference positions). A horizontal line, the *selector line*, indicates the currently selected position on a track. This line can be moved by clicking the left mouse button within the image window. ***Edit|Snap to peaks*** causes the selector line always to snap to the closest peak on the currently selected track. By default this option is enabled; disable it to undo snapping to peaks. Snapping to peaks is also disabled while holding down the SHIFT key.

### 5.2.4.1 Defining reference positions

If you are going to align bands of **reference** patterns, you should select the database standard for defining the positions. First make sure that a database standard has been selected (see 5.2.2) and that the gel containing the database standard is loaded. When defined, the database standard pattern is shown at the left side of the gel window after a gel is loaded.

To define the position of the reference peaks for the current database, select ***Edit|Reference positions***. This brings the program in the *reference position editing mode*, which changes the menu of the window. Select the first band of the standard reference pattern in the gel window. ***Peak|Add*** adds a reference position at the place of the selector line, which is indicated by a red arrow left from the gel window field; the *reference position marker*. A blue arrow, which points to the red reference position marker, indicates that this position is currently selected: the *reference position selector*. Move the selector line to the next marker peak on the track and select ***Peak|Add*** again to add the current position to the table of reference positions. Continue this until all marker bands are defined as reference positions.

*NOTE: It is not necessary to follow the sequence of bands on the pattern while defining the reference positions. For example, a few major bands can be defined first, and smaller bands may be selected afterwards.*

The reference position selector can be moved to another reference position by pressing the left mouse button on the reference position marker (the red arrow). The currently selected position is removed from the database using ***Peak|Delete***

In case a previous set of reference positions is to be restored at exactly the same positions (e.g., to warrant compatibility of previous and new database patterns), the positional value of each reference position can be exactly entered by selecting it (blue reference position selector) and using ***Peak|Position*** to enter the known positional value of the current reference position. If all reference positions are defined, ***Ok*** updates the normalization settings and quits the reference peak editing mode. For a given database, this work normally has to be done only once.

*IMPORTANT NOTE:*

*ONCE A DATABASE IS IN USE (CONTAINING NORMALIZED GEL FILES) EXISTING REFERENCE POSITIONS SHOULD NEVER BE MODIFIED! However, it is still possible to remove or add reference positions afterwards, if based upon the same database standard pattern. We recommend to do so with care and only if really necessary. If you are not sure about the compatibility after changing the reference positions, it is perhaps the best idea to renormalize all gels of the database, or otherwise, to check the compatibility carefully, by comparing identical patterns on new and previous normalized gels (in the Main program).*

#### 5.2.4.2 Manual association of bands with reference positions

The next step is the *association* of the bands on the patterns that correspond to the reference positions. The association of a peak on a given pattern with one of the reference positions works as follows:

(1) Select a peak on a given pattern to correspond with a reference position by clicking the left mouse button on the band of the pattern. Depending on the choice "Snap to peaks", the selector line will stick to the peak top or not (press and hold the SHIFT key while dragging the mouse to disable peak-snapping).

(2) Select the reference position you want to associate with by clicking on the red reference position marker.

(3) Now you can either press the RIGHT MOUSE BUTTON, the SPACE bar, or choose ***Associate|Peak*** from the menu. The association is now made and marked with red lines connecting the reference peak position with the peak position on the pattern. The currently selected associated peak on one of the patterns is marked with a green rectangle, the *association selector*.

***NOTES:***

*(1) If the gel is not too distorted, one can directly press the RIGHT mouse button on a reference band of a pattern. The <u>closest</u> reference position*

*with that position. If the reference position the peak is to be associated with, is NOT the closest, this fast method would associate peaks with the wrong reference position. In this case, first select the correct reference position, and press and HOLD the CTRL key while pressing the right mouse button to associate any reference peak with the reference position.*

*(2) It may sometimes be necessary to disable the "Snap to peaks" option in a particular case. Instead of disabling this option in the menu, it is possible to press and HOLD the SHIFT key while selecting the position. The "Snap to peaks" option is disabled as long as the SHIFT key is held down.*

*(3) The SHIFT and CTRL key controls can be combined.*

#### 5.2.4.3 Automatic association of peaks with the closest reference positions

***Associate|All reference positions*** causes the program to automatically associate all reference positions with the closest peak on each reference track, if any such peak is present. Similarly, ***Associate|Reference position*** causes the program to associate only the currently selected reference position with the closest peak on each reference pattern, if any. The menu item ***Associate|Selection*** pops up a submenu, allowing you to perform this automatic association on ***All patterns***, ***Reference patterns only*** or ***Non-reference patterns only***. The current active selection is marked with ✓. The default selection at startup is "Reference patterns only". Select "Non-reference patterns only" if you want to align internal bands on the non-reference patterns and the reference patterns do not contain such a band (see further, 5.2.5).

*NOTE: The selection between reference patterns, non-reference patterns or all patterns applies only to the automatic association. Manually, one can associate peaks from any pattern with any reference position.*

Of course, the automatic associations are sometimes incorrect and they can be edited manually. The green association selector marks the association which is currently selected. This association selector can be moved to another associated peak (on the same or another pattern) by clicking on the peak. ***Associate|Delete peak*** or simply pressing DEL deletes the currently selected associated peak. Press the RIGHT mouse button on another peak to associate that peak with the selected reference position. If the "Snap to peaks" option is enabled, the peak top will automatically be selected. ***Associate|Delete reference position*** removes all associations with the currently selected reference position. ***Associate|Delete all reference positions*** removes all associations that were present on the gel.

#### 5.2.4.4 Automatic association of all reference patterns by pattern recognition

In the alignment option ***Associate|By pattern recognition***, a powerful pattern recognition algorithm is used to align the reference patterns with the database standard. The association of bands is NOT based on peak detection, but on the alignment of pattern contours, which has a number of advantages in terms of reliability. This option is the obvious choice to use. Very often, it will find the correct associations, even for "aberrant" gels. The choice of a representative (not aberrant) database standard is important in this respect.

The associations made by the pattern recognition method can always be corrected manually, before the bands are aligned.

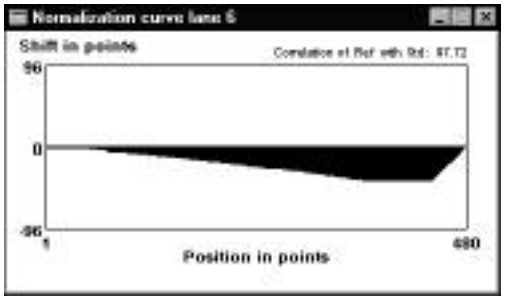
#### 5.2.4.5 Aligning the associated peaks

The last step, i.e. the alignment of the associated peaks is perfomed by ***Alignment|Align associated peaks***. All associated peaks are now aligned with their corresponding reference positions and the zones between the positions are interpolated linearly. In addition, if a particular track has no association with any of the reference peaks (e.g. if the peaks are too faint or not present), the position is interpolated linearly between those of the closest neighbours having associated peaks. ***Alignment|Undo alignment*** restores the peak alignment as before the reference peak normalization.

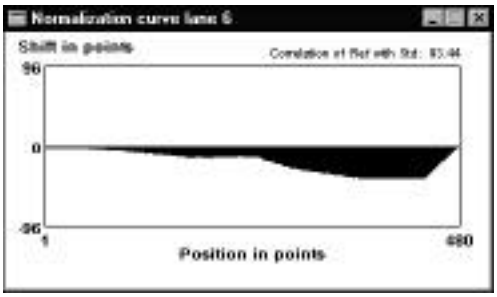
#### 5.2.4.6 Checking the correctness of the alignment

Alignments can be checked visually by comparing the reference patterns with the database standard shown at the left side of the gel. It is possible to scroll through the gel, to move any reference next to the standard pattern.

For each track, and in particular for the reference tracks it is possible to display the "Normalization curve", i.e. the shift in each data point that has been applied to obtain the current alignment. This curve can be shown by selecting ***Alignment|Normalization curve*** in the Normalization main window. If no correction is made, a flat horizontal line is seen. In normal circumstances, a red graph will be seen after association AND alignment. The further the graph deviates from the middle line at a given location, the larger the correction (shift) being applied at that location. This graph should be smooth and should not contain any discontinuities, except before the first and after the last reference position (example 1). If unnatural dents, shoulders or depths are detected for a reference, the alignment should be undone and one should check whether the correct bands have been aligned (example 2).




Example 1: Extreme though acceptable normalization curve.




Example 2: unacceptable normalization curve; one or more bands are faulty associated.

Upper right in the normalization curve window, the correlation between the current pattern (if the pattern is a reference) and the database standard is given in percent. This value can be an interesting indication of the normalization quality in systems where complex reference patterns are used.

*NOTE: A low correlation value does not necessarily imply that the alignment is wrong; it simply gives an indication of the reproducibility of the electrophoresis system used. The value is also dependent on the correlation interval as set in the Settings menu (5.2.2).*

#### 5.2.4.7 Stepwise alignment to reduce work with aberrant gels

A very useful feature of the "associating" alignment system is that the alignment can be executed stepwise to simplify the whole process for very aberrant or distorted gels. In case the automatic association does not align the peaks with the correct reference positions, in a first step, some peaks on one or a few patterns can be associated manually. If the distance between the reference positions and the peaks to be aligned is very large, first select a reference position and hold down the CTRL key to keep the selected reference position and then associate the corresponding peak(s) using the right mouse button. The peaks associated are then aligned, and in a second step, the pattern recognition or automatic association can be executed to align all patterns. More steps can be executed if necessary. Successive alignments cause no degeneration of the data. All association method can be executed in any order.

#### 5.2.4.8 Initializing the gel

After several trials, it can sometimes be necessary to undo any alignments and restart from the initial gel. ***Alignment|Initialize*** will remove all associations and restore the original, unaligned gel. This function has the same effect as reloading the gel.

## 5.2.5 Combining internal and external reference bands

### 5.2.5.1 Defining internal reference positions

If non-reference patterns contain internal reference bands or bands which are known to be the same in most or all of the non-reference patterns, these bands can be used for individual alignment of each pattern.

First, align the gel using the external reference pattern(s) and the defined reference positions. Then select a non-reference pattern that contains the band(s) to align and that occurs AS CLOSE AS POSSIBLE to an aligned reference pattern. Now add the band position(s) to the reference position table using the ***Edit|Reference positions*** option.

Initialize the gel by ***Alignment|Initialize***.

### 5.2.5.2 Aligning the internal reference positions

The internal reference positions are to be aligned before the external reference patterns in order to have effect. First, select the reference position with which the common bands should be associated. If the bands are distant from the reference position (or simply to make sure that the correct associations are made), press and hold down the CTRL key. While holding down the CTRL key, associate all corresponding bands with the reference position by pressing the right mouse button on each band. Repeat the same for a next reference position if available, and so on. If the distortion on the gel is not too large, one can also try ***Alignment|Selection|Non-reference patterns*** (or ***All patterns*** if all patterns contain the internal band), select the internal reference position and ***Associate|Reference position***.

If the correct bands are associated with the reference position(s), choose ***Alignment|Align associated peaks***. Then remove the associations with ***Associate|Delete all reference positions*** to subsequently align the external reference patterns as explained earlier. This can be done manually or by using the ***Associate|By pattern recognition*** tool, or with ***Associate|Selection|Reference patterns only*** and ***Associate|All reference positions***.

## 5.2.6 Saving the normalized gel

When the alignment of the gel is complete, ***File|Save*** calls a save dialog box. At the left side is a list of the directories shown, at the right side the filename input field. The current path is the default directory which has been specified as database in the User Setup menu see 8.2.1), but you can select another directory. Normalized gels are always saved as ".INT" files. This means that you can specify a file name of 1 to 8 characters (NEVER USE SPACES IN DOS FILENAMES) without extension. When background subtraction is

appears, showing the background subtraction process for all successive tracks. When this process is finished, you can load and normalize another gel, or quit the normalization program to return to the Molecular Analyst Fingerprinting software Startup screen.

## 5.3 Doublegel normalization.

In some electrophoresis systems, it is possible to mix a set of markers with each sample and visualize these using a different probe or staining dye. Typically, this results in two scanning files of the same gel, one of which showing the patterns, and the other the internal marker bands. As an extension to the conventional gel conversion and normalization functions, Molecular Analyst Fingerprinting software allows normalization of gels through such internal reference marker bands scanned into a separate TIFF file. This special method of normalizing is called the "Doublegel" conversion and normalization. The feature can be enabled by selecting the checkbox "***Doublegel normalization***" in the User Setup menu for the user in question (see 8.2.5). The conversion of such complementary files is discussed in section 4.5.

This special normalization method involves two steps:

1) Normalization of the primary gel containing the internal reference tracks;

2) Superimposition of the alignment vectors of each track of the primary gel to each corresponding track of the secondary gel.

In practice, the primary gel (containing the reference patterns) and the secondary gel are loaded together and the primary gel can be normalized in the usual way. The secondary gel automatically undergoes the same normalization as the primary file and can be saved to the database.

### 5.3.1 The choice of a standard

One good-looking and not too aberrant pattern from a primary gel (i.e. an internal reference pattern) is first defined as **database standard** (see 5.2.2) and the bands of that pattern are defined as **reference positions** (see 5.2.4.1).

### 5.3.2 How to normalize

In the Normalization program's main menu, select ***File|Load*** to load the *primary* gel file. Then select ***File|Load secondary gel*** and load the secondary gel. The primary gel is now shown in the window. Depending on the shift between the database standard and the current gel, automatic associations using the ***Associate|Pattern recognition algorithm*** can first be performed on the patterns marked as references, or a first rough manual association can be done

(see section 5.2.4). Before and after alignment, you can toggle the view between both gel images with ***Edit|Show secondary gel***. Whatever gel shown, it will only be possible to normalize tracks from the primary gel.

After alignment the primary gel can be saved with ***File|Save*** (not necessary!) and the secondary gel with ***File|Save secondary gel***.

# 6. The Main program

All analysis functions after normalization of gels, which include displaying patterns, drawing curves, displaying and changing descriptive information, searching for entries through databases using a range of information fields, composing lists of tracks for comparative analysis, quantifying molecular sizes, etc. are provided by the **GCMAIN** program. The GCMAIN program is loaded by pressing the <***Analyze***> button in the Molecular Analyst Fingerprinting software Startup screen.

## 6.1 Short menus and extended menus

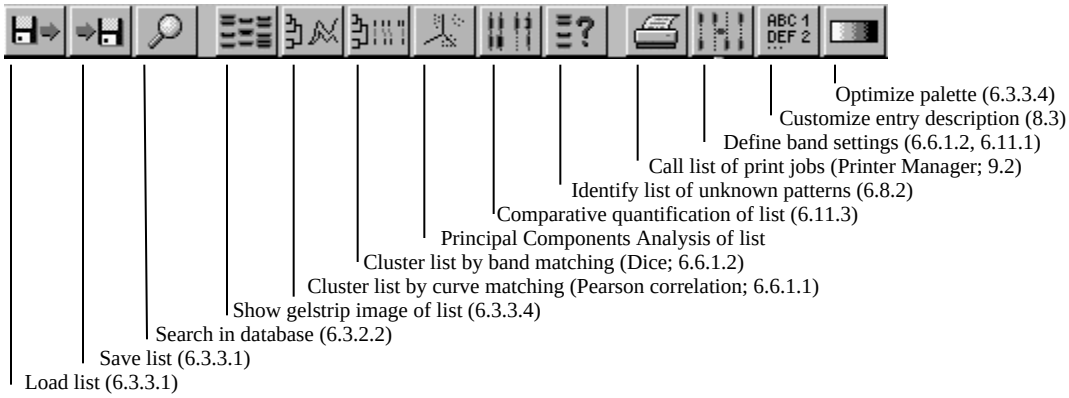
The Molecular Analyst Fingerprinting software Main program can run in two modes: with short, simple menus and extended, complete menus. After installation Molecular Analyst Fingerprinting software starts up with short menus. This allows you to quickly perform some of the most common analysis tools such as cluster analysis, identification etc, without having to bother about numerous settings, parameters, options and choices. In this mode, the program automatically selects the most generally applicable combination of settings and parameters for you and chooses the most common clustering and identification methods.

In the “short menus” mode, the program can be switched to the “extended menus” mode using the menu command ***System|Extended menus***. Conversely, the extended menus are turned into short menus using the command ***System|Short menus***.

All descriptions in this manual apply to the extended menus, because the short menus are part of the extended menus. There are only few exceptions where the short menus are different: ***Edit|Gel image*** corresponds to ***Edit|2D-Gel image*** in the extended menu, whereas ***Comparison|Identify*** corresponds to ***Comparison|Identify with list*** in the extended menu.

## 6.2 The Analyze Toolbar

The most frequently used functions in the Analyze program are directly available on the toolbar right under the menu bar. The toolbar contains 13 buttons which are described in the picture below. The numbers refer to the relevant sections in the manual.




If no list of patterns is selected, all the toolbar buttons for functions related to lists are greyed.

## 6.3 Database management and construction of lists

The main Molecular Analyst Fingerprinting software window is divided into three parts:

(1) The *file table*, showing a survey of the normalized gel files in the default directory (i.e. a database). The default start directory can be specified in the User Setup menu in the Startup screen (see 8.2.1). Subdirectories of this directory are marked in bold, and the parent directory of the current directory is marked with "<==". At the bottom of the window, the *status bar* shows the full name of the current directory, which is the defined database directory at startup. A selection bar can be moved through the files by using the mouse or the arrow keys. The current directory can be changed by positioning the selection on a subdirectory name and double-clicking on it or pressing ENTER. Exit the current directory by clicking on the "<=" field. The menu command ***Database|Go to database*** lets you return to the default database for the current user. Other available drives are represented in **bold**, wityh the corresponding drive letter between square brackets.

(2) The *list table*, showing the current selection list (which is, of course, initially empty). Info boxes are supported for the list table.

(3) The list and ClusterBases maintenance subwindows, allowing previously stored selection lists, dendrograms, or clusterbases to be loaded (green arrow buttons) or deleted (red cross buttons).

The subwindows are separated by separator splines, which can be dragged to the left or to the right (or up and down) to display more information either in the file window, in the list table or in the maintenance window. Any file, list or list entry can be pointed and selected by clicking with the left mouse button,

causing the selection bar to be placed on that item. Double-clicking will cause a gel to be opened, a list to be loaded, or a densitometric curve of a list entry to be shown, respectively. Alternatively, files, lists or entries can be selected using the arrow keys. One can use the TAB key to switch from the file table, list table or entry table. In the active subwindow, the selection bar is marked with red borders. The ENTER key has the same effect as double-clicking with the mouse.

### 6.3.1 Showing and editing gel information.

To edit a particular gel file, double-click or press ENTER on the corresponding file name in the file table. A *gel window* (info-box supported) appears, listing the tracks in the file with abbreviated information fields.

*NOTE: More details and fields will be displayed when the gel window is enlarged or maximized.*

A selection bar can be moved through the entries of the gel. At the bottom of the window, the status bar shows the full filename of the gel (including its path), the name of the standard used for normalization and the resolution of the tracks (i.e. the number of points). With ***File|Information*** you can show all the detailed information about the TIFF file and the conversion settings (first set of data) and the normalization settings, files, their directories and dates of creation (second set of data). ***File|Print*** generates a print-out of the complete gel information. ***Image*** shows the original TIFF file of the gel (if present!) with the splines and gelstrip cut-outs indicated as used.

With the option ***Gel|Copy info***, it is possible to copy all information (the information fields for all tracks) into the Molecular Analyst Fingerprinting software clipboard. Similarly, all information from another gel can be pasted into the current gel using ***Gel|Paste info***. The only restriction is that the number of tracks in both gels be the same; otherwise the function would not work. The same function is available in the GC-Conv program, which is very useful when a gel is being rescanned for some reason. This information is stored in a text file "GELCLP.TXT" to be found in the Molecular Analyst Fingerprinting software home directory. The file can be easily edited in the Windows Notepad program. It has the following format:

**HEADER:**

```
LINE 1: full path gel name
LINE 2: Number of tracks in the gel (e.g. 'n')
```

**TRACK DESCRIPTION (n times repeated):**

```
LINE 1: Lane number; plus "R" if reference track
LINE 2: First name field
LINE 3: Second name field
LINE 4: Third name field
```

```
LINE 5: First entry code field
LINE 6: Second entry code field
LINE 7: First comment line
LINE 8: Second comment line
```

If you want to paste information from a gel containing 20 tracks into a gel containing 19 tracks, load GELCLP.TXT in the Notepad program, change line 2 of the header into 19 and remove the 8 information fields of the track which does not occur in the gel with 19 tracks.

***Gel|2D-Gel image*** displays the normalized gelstrips of the gel in a g*el display window*, whereas ***Gel|Reconstructed gel image*** shows a reconstructed image of the whole gel after normalization and background subtraction, if done. See section 6.3.3.4 for more information about this window. ***Gel|Bands*** calls the *band edit window*, which allows you to mark bands for all of the tracks of the gel. This window will be discussed in detail in section 6.11.1. ***Gel|Bands quantification*** calls the *Quantification edit window,* a tool which is described in 6.11.2.

***Track|Densitometric curve*** calls a *curve window* showing the densitometric curve together with the gelstrip of the selected track. If this command is executed a second time on another track, the new curve and gelstrip are added in the same window, the curve drawn in a different colour (up to 3 curves in different colours can be shown simultaneously). See sections 6.3.3.5 and 6.3.4 for more information about the use of the curve window.

***Track|Information*** or double clicking (ENTER key) on a particular track pops up an *information window* of fixed size, containing input fields for the descriptive information of the selected track. **Print** sends a print out of this information to the print queue. With ***Edit|Copy track info*** and ***Edit|Paste track info*** you can copy and paste all information fields from one track into another, even from different gels. Similar to the gel copy function described above, the information is stored in a text file "TRACKCLP.TXT" in the Molecular Analyst Fingerprinting software home directory, which can be edited in the Notepad program. The first two lines are the gel name the full path respectively, and the next 8 lines are identical to the track description as for the GELCLP.TXT file. After changing contents of these fields, ***OK*** or the ENTER key is used to update the gel window, whereas ***Cancel*** is to exit without changing (Esc key). All gel and track modifications are not actually saved until ***File|Save*** is called in the gel window.

## 6.3.2 Construction of lists of tracks

Most of the features (such as cluster and principal components analysis, identification, quantification etc.) use a selection list as input. Molecular Analyst Fingerprinting software offers sophisticated tools for constructing, handling and storing such lists. The list table at the right side of the screen

shows the abbreviated information of the currently selected tracks. By dragging the separator between the file table and the list to the left, the entry names will be shown in full length, as well as other information fields, depending on the resolution of the screen.

#### 6.3.2.1 Manual selection of tracks

After calling a gel window from the file table, it is possible to select a track from this gel into the list by clicking the right mouse button on its name. A green arrow appears at the left hand side of the track, indicating that this track is currently selected, and the track's name appears in the list table. A selected track can be removed from the list by pressing the right mouse button a second time on its name; the green arrow disappears and the track name is removed from the list table. A selected track can also be removed by clicking with the right mouse button on its name in the list table. If one or more tracks from a particular gel are selected, the gel window can be closed without loosing the selection information. By reopening the gel at a later time, the tracks selected in the list will be indicated again.

***NOTE:*** *Do not try to select tracks from gels with a different number of points (resolution) or from gels normalized according to different standards since these tracks cannot be compared (see Normalization Settings, section 5.2.2). In such events an error message is shown. You can also compare the normalization settings for each gel by displaying its Information window (6.3.1).*

#### 6.3.2.2 Automatic topic search

Molecular Analyst Fingerprinting software allows database search for tracks that fulfill certain conditions in their information fields (see also 8.2.4). ***Search|Topic*** from the main window is used to call the *topic search dialog box*. This box can be called using the F5 hot key as well (see 0). This dialog box contains input fields for all seven information fields. If a string is entered in one of these fields, tracks that have the same string in the corresponding information field will be selected (e.g. if the second field stands for the genus name and "ENT" is entered in this field, all species of the genera containing "ENT..." as three first characters will be selected). The "***Case sensitive***" checkbox allows you to determine whether the case of the characters is checked during the search (if this box is not checked, the genera beginning with "Ent..." will be selected as well). If the "***Negative search***" box is checked, all tracks that do **not** match the specifications will be selected. The "***Method***" radio button group determines whether the tracks found will (1) be added to the existing lists (2) replace the existing list or (3) be selected from the already existing list and replace this list.

Pressing **<*OK*>** starts the search. After searching through the current database directory, the list table will show the updated list. As for manual selection, the automatic selection routine will never select tracks which have different resolution or standards. The routine will only add tracks from gels having the same database settings as those previously selected in the current list.

Combining the various possibilities in subsequent topic search calls allows you to search for virtually every possible combination.

#### 6.3.2.3 Band search (Quantification module only)

***Search*|*Bands*** from the main window allows tracks that have bands on specific positions to be searched for. See section 6.11.1 for more information about assigning bands to normalized tracks. This option opens the *band search window*. ***Scale*|*Position*** and ***Scale*|*Metric*** determines whether the band positions will be shown as data point numbers on the normalized tracks or as metrical units (see section 6.3.4 for more information about the use of metrical units).

Use the command ***Add*** to select a band that should be searched for through the database. A dialog box appears with three input fields, asking for (1) the position of the desired band, (2) the position tolerance, i.e. the maximal shift allowed between an actual band found on the track and the predefined band position (a measure expressed in % of the total curve) and (3) the minimal band surface a band should have to be selected, expressed as percentage of the total surface of the bands defined for that track. Press **<*OK*>** to add the band to the list, which shows its position, maximal shift and minimal surface.

It is possible to select up to 20 bands that should occur together in a track in order to be added to the selection list. ***Edit*** is used to re-edit the settings of the currently selected band, whereas ***Delete*** removes the band from the table.

Press **<*Start*>** to start the database search.

### 6.3.3 List management

#### 6.3.3.1 Creating and saving lists

It is possible to assign a name to a list. This name appears in the status bar of the main window; new lists are by default named as "NONAME". ***List*|*Save as*** (Ctrl+F2) from the main window allows a new name to be assigned to a list which will be saved to disk. A dialog box appears, showing the names of all existing lists and an input field for the new name of the list. The list name should be a valid filename without extension, i.e. containing maximally 8 characters and no spaces, periods or other special characters. A warning message is shown if the name is invalid. Press **<*OK*>** to save the list. ***List*|*Load***

(F3) in the main window allows you to load a previously saved list from the disk. Again, a dialog box shows all lists on the hard disk and an input field. A list can be loaded by (1) double-clicking on its name, (2) typing or clicking once on its name and pressing **<*OK*>** or (3) double-clicking on the list name. Note that hot keys F2 and F3 are available to save and load lists, respectively (see also 0). Alternatively, ***List*|*Append*** (Ctrl+F3) works similarly but appends the list to the existing list instead of replacing it. When a list has already a name assigned to it, ***List*|*Save*** (F2) saves it on the hard disk without prompting for a new name, except when the list is new.

> ***NOTE**: Lists are saved in a single directory which can be specified for each user in User Setup (default: \MA-F\LISTS, see also section 8.2.1). In this way, there is no interference between the lists of different users.*

A list is printed out using ***List*|*Print*|*Custom fields, List*|*Print*|*All fields***. When the first option is selected, the list entries are printed with the information fields selected in the Entry description dialog box (see 8.3). The last option prints all available information for every track.

The list which is currently selected can be cleared by the ***List*|*Clear list*** command. Note that, if the list was saved, it will still be present on the hard disk. Hot key F4 has the same function (see also 3.8). The ***List*|*Delete entry*** command removes the currently selected track from the list table. Selecting a track in the list table and pressing the DEL key or the right mouse button has the same result.

The option ***List*|*Maintenance*** from the main window calls the *list maintenance window*, which allows you to remove unnecessary lists from the hard disk. An survey of the saved lists is given, showing the creation date of the selected list, as well as the presence and creation date of its cluster analysis. ***Remove*|*List*** removes the selected list from the hard disk, whereas ***Remove*|*Clustering*** only removes the cluster analysis of that list (if present).

#### 6.3.3.2 Export and import of lists

The function ***List*|*Export*|*Entry codes to clipboard*** will copy the entry codes of the patterns selected in the current list to the Windows clipboard. This feature will allow you to select the same entries in other Windows databasing or spreadsheet applications. ***List*|*Export*|*Custom fields to clipboard*** copies the field combination as defined in the Entry description dialog box (see 8.3) to the Windows clipboard. ***List*|*Export*|*Complete tracks to file*** creates a sequential ASCII text file containing all information fields and numerical densitometric data of the tracks in the current list. The data can be easily imported in other applications.

Similar as the export of entry codes to other applications, lists can be composed in Molecular Analyst Fingerprinting software based upon entry codes imported

via the clipboard. ***List|Import|List from clipboard*** allows you to select entries in Molecular Analyst Fingerprinting software copied to the clipboard by other applications on the basis of their entry codes. ***List|Import|Free strings from clipboard*** is a function which allows you to assign any description to the patterns shown in images, dendrograms, identifications, quantifications etc. The only requirement is that you have created an ASCII text containing for each pattern its entry code field and the name you want to replace with, always separated by line breaks. The file should look as follows:

```
Entry code field of entry X
New label for entry X
Entry code field of entry Y
New label for entry Y
...
```

The contents of this file should be copied to the Clipboard and Molecular Analyst Fingerprinting software will automatically search for the entries having the corresponding entry codes in the current image. The ***Free string*** label must be selected in the Entry description dialog box (see 8.3).

#### 6.3.3.3 Selecting active zones on the densitometric curves

Each list has one or more active or enabled zones from the densitometric curves associated with it. In grouping analysis or identification (see 6.6.1, **Error! Reference source not found.**, 6.8.2 and 6.9.2), only the data points lying within these zone(s) are used for comparison. This option is useful to exclude non-specific or unwanted zones from the gel.

***Edit|Zones*** from the main window displays the *zone window*, showing the densitometric curve and the gel strip (if available) of the currently selected track in the list table. Double-clicking on a track in the list table hasd the same result. The enabled zone is marked in blue, while the rest of the curve is marked in gray (initially, the whole curve is enabled). To desactivate a region on the curve, move the mouse pointer to the beginning of the desired region, click and hold the right mouse button and move the pointer to the end. While moving the mouse, the covered area becomes gray. Selecting a desactivated region works in the same way, but using the left mouse button. At the bottom of the window is the position of the mouse pointer shown, both in data points and in metrical units (if defined). Exit the zone window to make the newly selected zone(s) valid.

***NOTES****:*

*(1) If a list is saved on disk, the active zones are saved along with it.*

*(2) The zone window can be resized to define the regions more accurately.*

### 6.3.3.4 Reconstruction of pattern images

An interesting feature of Molecular Analyst Fingerprinting software consists in the reconstruction of gel images of normalized patterns. There are two methods for displaying the patterns.

***Display|2D-Gel image*** from the main window calls a *gel display window* (info-box supported), showing up to 200 normalized gelstrips of the list (in order to obtain a convenient image, it is necessary to have a graphics adaptor in 256 colours mode). ***Display|Reconstructed gel image*** shows an image of the tracks as reconstructed from the densitometric curves.

The first option is much more truthful but is less fast in operation. Each individual pattern is cut out from the TIFF image as "gel strip" in the Conversion program (see 4.1 and 4.2.4, item 4) which is normalized togehter with the densitometric curves during the normalization process.

> *NOTE: With some non-accelerated graphics cards, displaying the gel strips will be much slower than displaying the reconstructed tracks.*

The colour palette, brightness and contrast of the gel images can be changed with the ***Gelstrips|Palette*** menu (main menu; also available by pressing F6). The ***default*** colour palette is defined by the user in the User Setup menu (see 8.2.3). ***Gray*** and ***gray Inverted*** are predefined normal and inverted gray scales, respectively. Rainbow is a multi-colour palette to reveal more contrast in intensely stained or exposed gels. Use the options *Brightness* and *Contrast* to further optimize the image.

Each track is labeled by a subset of information fields. This subset can be changed by calling the Entry description dialog box (see 8.3) using the F9 function key. If you want to have the complete information of any particular pattern, press CTRL+left mouse button to reveal its info box (see 3.7). ***Show|Bands*** visualizes or hides the bands that were defined for the tracks (if any) on the image, as red lines. ***Show|Geltracks*** allows you to toggle between the 2D-gelstrips (marked with ✓) and the reconstructed tracks. The option ***Show|Metrics scale*** allows a scale of the molecular sizes to be shown next to the fingerprint image (marked with ✓). Obviously, the scale will only be shown when the Metrics are defined for the database in use (see 6.3.4). ***Print*** creates a high-quality graphical print-out of this image respecting the selected colour palette, brightness and contrast. With ***Bitmap*** the program creates a Windows bitmap file of the image in the Molecular Analyst Fingerprinting software home directory, named "PRINT.BMP".

A selection bar can be moved through the image with the mouse or the cursor keys. It is possible to rearrange the displayed patterns using the cut-and-paste option. Tracks on the image can be marked by pressing the right mouse button on the image (or pressing SPACE), and unmarked by pressing the right mouse button or SPACE a second time. Groups of tracks are selected by marking the

first and then marking the last while holding down the SHIFT key. When a track is selected, a red sequential number appears at the top, indicating the order of marking of tracks (the first marked track gets "1", the second "2", etc.) If ***Edit|Cut*** is chosen, the selected tracks are removed from the window. Select then a new position to recover the marked tracks in the selected order at the new position with ***Edit|Paste***. The order of selection will determine the new position after the tracks are cut and pasted. Tracks can be cut and pasted from one gel display window into another. Using the ***Edit|Arrange from clipboard*** command, the gel image will be automatically rearranged according to the order of entry code names (see 8.2.4) if these are contained in the Clipboard.

*NOTE: When the size of the gel display window is changed, the patterns will be automatically resized within the new window dimensions. Use this feature when you have a large screen!*

#### 6.3.3.5 Showing densitometric curves

Select a track from the list and use ***Edit|Densitometric curve*** in the main window to display the densitometric curve togehter with the gel track (if available) of the selected entry (see section 6.3.4 below for more details).

### 6.3.4 Assigning metrical units

The default measure of distance on a track in Molecular Analyst Fingerprinting software is the absolute data point position on the normalized curve. However, it is often preferred to express positions on a gel in units which have some physical importance, such as molecular weights or sizes, isoelectric points, RF values etc. In Molecular Analyst Fingerprinting software, these units are called **metrical units**. The metrical units are calculated using a **calibration sample**, which contains a set of bands with known metrical units. This calibration sample is run on a gel which is normalized in the usual way. It can be any sample of which you know the molecular sizes of the bands.

To assign the metrical units, open the gel containing the calibration track from the file table, select the calibration track and choose ***Display|Densitometric curve*** from the gel window, or in case the track is present in the list table, you can select it there and choose ***Edit|Densitometric curve*** in the main menu. The densitometric curve of the calibration track is now shown in the curve window together with the gelstrip (if available).

*NOTE: Resize or maximize the curve window to get a larger view of the densitogram. You can also display more than one track in the same curve window to define the band positions more accurately.*

If the left mouse button is pressed in this window, a *position ruler* appears at the mouse pointer position. In the upper right field of the window, the data

upper left field, the names of the selected tracks and their respective densitometric values at the position ruler are shown. A part of the curve can be enlarged by dragging the position ruler to that part and calling ***Zoom in*** (or ENTER). The whole curve is recoverd by selecting ***Zoom out*** (or ENTER). Select the first calibration band with the position ruler and - if necessary - by zooming in. Select ***Metrics|Add node*** or press the INSERT key to show the value input window. Enter the metrical unit value of the selected band (real or integer number, without unit of measure) and press **<*OK*>**. The position of the first calibration band is now marked with a dotted line, and the entered value is indicated. This procedure can be repeated for all subsequent calibration bands. Use ***Metrics|Delete last node*** or the DELETE key to remove a wrong calibration point from the list.

If this is completed, ***Metrics|Interpolate*** displays the *interpolation menu*, showing a graph of the metrical units of the calibration bands (Y-axis) against their data point position (X-axis). In order to assign metrical units to every point on the track, Molecular Analyst Fingerprinting software fits a curve through these data points. A radio button box allows choice between four regression methods, listed in increasing order of freedom: (1) L***inear*** curve fitting: $y = ax+b$; (2) ***Exponential***: $\ln(y)=ax+b$, (3) ***Exponential fitting***: $\ln(y) = ax^3+bx^2+cx+d$, (4) ***Spline functions***, which fits the curve through all entered points using "cubic spline" functions, and (5) ***Pole function***, which approaches the curve through a hyperbolic regression (A+x) (B+y) = C.

If the function is expected to be linear (e.g. for isoelectric point estimations), the first equation can be chosen. If the function is expected to be logarithmic, the second is preferable. However, in most gels systems, the migration behaviour of macromolecules is not perfectly logarithmic, and therefore the *exponential fitting* method adds a factor of freedom by incorporating a third power regression, which allows deviations from the logarithmic basis to be followed correctly. The *spline functions* method can follow the most complex migration functions exactly through all points, by the use of combined third power functions. This regression will allow very accurate *interpolations*, but is not intended for *extrapolation* of molecular sizes, i.e. beyond the entered marker bands! Use the latter function only when the size markers cover the entire range of interest of the gels. In addition, you will need at least some 5 or 6 marker bands in order to circumscribe the spline functions accurately. The *pole function* is very reliable for extrapolation, e.g. when few known marker points are available. However, it allows little freedom for curves that deviate from the hyperbolic basis.

With **<*Print*>**, a graphical print job is created for the calibration curve.

The units of the metric (e.g. Da, kDa, pH, bp, kb etc.) can be typed in the corresponding input box (max. 4 characters). **<*Save*>** validates the newly calculated metrical unit. From now on, every position on all tracks from the same database (i.e. which are normalized with the same standard as the

calibration curve and which have the same resolution) can be expressed in metrical units.

## 6.4 Database quality control

### 6.4.1 How database control is measured

Molecular Analyst Fingerprinting software possesses the possibility to estimate the quality of each gel of the database in terms of reproducibility. There are two ways to monitor the reproducibility of a database: by *reference statistics*, and by *band tolerance statistics*. The reference statistics are more direcxted towards fingerprints that can be meaningfully compared by Pearson product-moment correlation (see also 6.6.1.1), whereas the band tolerance statistics are typically designed for patterns that are usually compared using band matching coefficients (see 6.6.1.2).

In the reference statistics, the quality of a gel is determined by the correlation between its reference tracks and the database standard. The quality of each reference track is shown in the gel window (see 6.3.1) as a square in a certain colour. The colour code ranges from red (lowest quality) over orange, yellow and greenish yellow to green (highest quality). The quality indication is also shown in the info-box (see 3.7) of each particular track. The normalization quality of non-reference tracks is calculated from the reference tracks of the gel, assigning relative weight to the neighbouring reference(s) in reverse proportion to their distance. Since the info-boxes can be called from within virtually every analytical application, you can always inspect the quality of any track at a glance.

The band tolerance statistics require that in addition to the reference patterns, a given non-reference pattern is loaded in all or most gels. When a list of these identical patterns is created, the statistics will automatically calculate the average shift on each band for various confidence levels, so that the user can easily estimate the reproducibility in function of parameters such as position on the patterns, dat aof experiment, etc. In addition, the results allow the position tolerance for band matching to be estimated in a meaningful way for each particular electrophoresis system.

### 6.4.2 Creating and updating the reference statistics

With the menu item ***Database|Reference statistics*** you can introduce or update your database quality statistics using the *Reference statistics window*. Press ***Update*** in the menu to create or update a list of the correlations of all reference patterns with the database standard of the current database (shown in the status bar). The reference tracks are sorted independently according to increasing

correlation with the standard reference. A histogram displays the distribution of the correlation values together with a colour scale ranging from green to red. You can drag the colour scale with the left mouse button until the yellow zone corresponds to the top of the histogram. In the right list box, you can easily get an idea of how the adjusted colours correspond to the correlation intervals. Press ENTER or double-click on a reference track to display its gel with the current reference selected.

Press ***Exit*** in the menu to close the reference statistics window. All gel windows will now display the updated quality colours.

## 6.4.3 Band tolerance statistics

When fingerprints are compared using band matching coefficients (6.6.1.2), Molecular Analyst Fingerprinting software uses a so-called band tolerance (see also section 6.6.1.2). If the difference in position of two bands on different fingerprints is less than this value, they are considered as identical. If they are shifted over a larger distance, the program treats them as two separate bands.

The user can freely choose this band tolerance using the *band settings dialog box* (see also section 6.6.1.2). It is expressed as a relative value (in percentage of the total normalized gel length). Obviously, the band tolerance is an important parameter for the comparison of fingerprints, and therefore Molecular Analyst Fingerprinting software offers a tool which helps the user determine the value of this parameter in a meaningful and objective way. This tool is called Band tolerance statistics, and is based on the statistical analysis of the differences in band positions of a list of identical fingerprints.

### 6.4.3.1 The construction of band tolerance statistics

The first step is to make a list of identical fingerprints, collected from a wide range of different runs (experiments). Reference tracks (which were used for the normalization of the gels) should not be used, since using these tracks would would learn you nothing about the reproducibility of other, non-reference tracks. If such a set of identical fingerprints is not available, a set of fingerprints with a very high portion of common bands will also do. Selecting ***Database*** | ***Band tolerance statistics*** from the Molecular Analyst Fingerprinting software main window pops up a dialog box which asks whether optimization should be applied or not. If <***Yes***> is chosen, the program will show the tolerances as in the case that optimization is applied (see also section 6.6.1.2). Otherwise, no optimization is applied. When the calculations are finished, the *band tolerance statistics window* appears. This window is divided in four parts:

- Top: histogram of the band positions
- Middle: overview of all band positions of the different list entries
- Bottom: tolerance statistics graph

- Right: list of the fingerprints

### 6.4.3.2 Tolerance statistics graph

The horizontal axis of this image depicts the normalized run length on the gel, while the vertical axis shows the deviation on the band positions (in points). The vertical lines on the image correspond to the positions of the bands which were used to calculate the tolerance statistics.

The red line shows the RMS (Root-Mean-Square) value of the differences in band positions. One can also interpret this value as a standard deviation. This value is calculated for each band, and the result is drawn as a curve.

The green, cyan, and blue lines show the 50%, 90% and 98% limits on the position differences, respectively. For example, 50% of the fingerprints are shifted over an interval which is not more than the deviation indicated by the green line. This deviation (*d*) is given in absolute data points on the normalized densitometric curves. It can be easily converted to a percentage (*%d*) of the total track length (*n*) as follows:

$$\%d = 100 \times d / n$$

Again, these values are calculated for each individual band. The curves connecting these values give an idea of the average error in relation to the distance on the patterns.

The user can print this image using ***Statistics*** | ***Print***.

### 6.4.3.3 List of the fingerprints

This list shows the gel files and the gel positions of each fingerprint in the list, together with the mean deviation of the band positions, compared with an averaged fingerprint. Use ***Order*** | ***No ordering*** to show this list in original ordering, ***Order*** | ***By deviation*** to show the list in order of increasing deviation from the averaged fingerprint, or ***Order*** | ***By time*** to order the list by the time associated with the normalized .INT file. The fingerprint band positions in the left side window are ordered using the same criterion.

In addition, one can select a particular entry and use ***Statistics*** | ***Show gel*** to pop up the gel window of the corresponding normalised gel file (double-clicking on the entry does the same). Clicking on an entry while holding down the SHIFT key brings up a gelstrip of the fingerprint. These features are particularly useful for detecting aberrant gel files in a database (e.g. due to an error made during normalization).

# 6.5 Combining gels

## 6.5.1 Principle

Molecular typing applications in epidemiology sometimes require that more than one fingerprint is run for each organism, in order to detect clonal variations. The evaluation of different fingerprints is often needed in DNA fingerprinting techniques by using different probes, primers or restriction enzymes, or combinations.

Molecular Analyst Fingerprinting software has the possibility to combine different runs of the same organisms, even from different databases and with own normalization settings, into new synthetic gels, which contain all the information of the constituent patterns. For most functions, these gels behave like other gels: cluster analysis, identification, databasing, imaging etc. The display of metrical values for bands however, is not available for these gels.

Since a combined gel is composed of different gels, each of which having own normalization settings, some restriction must exist which makes it impossible to compare combined gels which have different normalization settings in their constituent gels. Therefore, Molecular Analyst Fingerprinting software will check for each new and unique combination of normalization settings, and will ask you to enter a new *standard* name for the combined gel, if the combination is new. The next time you construct a combined gel with exactly the same set of normalization settings derived from the constituent gels, the same *standard* name will automatically be adopted, and Molecular Analyst Fingerprinting software will not ask to enter a standard name for the new combined gel.

## 6.5.2 Creating a combined gel

A combined gel can be created by selecting ***Database|New combined gel*** from the main menu. This opens the *Combined gel window*. The first gel added to the combined gel will be the *master gel,* which means that the number of tracks as well as all information fields of the tracks will be adapted from that gel. Thus, it is important to select a gel with the right number of tracks and correct and complete names as first component. An empty window leaves the possibility to add the first gel to the new combined gel by ***File|Add new component***. In a file scroll box, you can select any gel from the default database or from any other database.

When a gel is loaded, the complete path and information of the current gel is shown for each track (enlarge the window if necessary). With ***File|Add new component***, another gel can be added. The gel name and normalization settings are indicated for each constituent gel in a separate field above the track list. You can display a particular constituent gel by clicking on that field or by using the TAB key.

*NOTE: Since Molecular Analyst Fingerprinting software adapts the information fields from the first gel entered into the combined gel, it is not necessary in the Conversion program to enter all the information for each of the other gels of which you know they will only serve as part of a combined gel.*

Constructing combined gels will be easiest when the same sequence of patterns is applied for each of the constituent gels, and this is the recommended way of working. However, the option ***Change link*** allows a selected track on one of the constituent gels to be replaced by a track from another gel. This option is in first instance intended to replace bad patterns which have been re-run on another gel. Select the pattern to be replaced on its constituent gel, and choose ***File|Change link*** from the menu. A survey of all database gels is given, from which you can select one. An input field in the left corner lets you specify the track number on the selected gel. Press **<*Ok*>** to confirm the track replacement.

With ***File|Save combined gel***, the combined gel can be saved in the defined directory for combined gels (see 8.2.1) or in any other directory. In a combined gel, the constituent patterns are physically appended to generate a chained composite pattern for each gel lane. The interesting features of this method are that no information of constituent patterns gets lost, and that any combination of electrophoresis patterns can be combined (e.g. an IEF gel with a pulsed-field agarose gel and an SDS-PAGE gel). If the combination of normalization settings did not exist yet, the program first asks to enter a *name for standard* for the gel. Enter a name of maximally 8 characters length, without extension, spaces, or periods.

*NOTES:*

*(1) Once a combined gel is saved to disk, it behaves like a single gel and it is not possible to decompose the gel into its constituents. This implies that it is not possible to re-edit the combined gel and replace tracks after closing the combined gel window. The gel should be recomposed to make such changes.*

*(2) Bands already defined for any constituent gels are saved with the combined gel. It is recommended to edit bands first on the constituent gels and then compose the combined gel, as editing bands on single gels is much easier and more reliable. However, bands can be edited afterwards on the combined gel as well.*

*(3) The 2D-Gel strips are also composed for the combined gel. This is automatically achieved by appending the constituent gelstrips. Note that this may fill quite some disk space.*

With ***File|Save superimposed gel***, a superimposed gel can be saved in the database directory (see 8.2.1) or in any other directory. Superimposed gels can

only be generated from constituent patterns having the same normalization settings. The principle of superimposing gels is that all the constituent patterns are mapped on each other to create an averaged profile. The superimposed gel inherits the same normalization settings as the constituent gels and is by default saved to the database of the active user. Enter a file name for the superimposed gel of maximally 8 characters length, without extension, spaces, or periods.

*NOTES:*

*(1) Once a superimposed gel is saved to disk, it behaves like a single gel and it is not possible to decompose the gel into its constituents. This implies that it is not possible to re-edit the superimposed gel and replace tracks after closing the combined gel window. The gel should be recomposed to make such changes.*

*(2) Bands already defined for one or more constituent gels are saved with the superimposed gel. It is recommended to edit bands first on the constituent gels and then compose the superimposed gel, as editing bands on single gels is much easier and more reliable. However, bands can be re-edited on the superimposed gel at any time.*

*(3) The 2D-Gel strips are also composed for the superimposed gel. This is automatically done by averaging the constituent gelstrips. A superimposed gel with gelstrips uses no more disk space than a single gel.*

When the <***OK***> button is pressed to save the superimposed gel, a dialog box prompts to "Optimize intensities" or not. When <***Yes***> is chosen, the intensity of each pattern separately will be recalculated to use the range of 256 gray levels.

## 6.6 Clustering

The ***Cluster analysis*** module allows comparison and grouping of lists of more than 2500 patterns using the "***Clustering***" tools, and of virtually unlimited numbers of patterns (up to 7000 tested with 32 MB RAM) using the "***Clustering databases***" tools. Some theoretical aspects of clustering are described in this section, which deals with the conventional "Clustering" methods. The next section describes the much more advanced "Clustering databases", and assumes that the fundamentals of clustering are known. It is recommended that beginners start using the "Clustering" tools and read this section (6.6). Those laboratories where extremely large-scale epidemiological and taxonomic studies are conducted will really benefit from the "Clustering databases" principles, which allows unlimited databases to be merged into single dendrograms.

Clustering involves two steps:

(1) calculation of similarity between all possible pairs of tracks from the list;
(2) cluster analysis of the matrix of similarity values.

## 6.6.1 Calculation of the similarity matrix

The calculation of the matrix of similarities can be based either on the Pearson product-moment correlation coefficient or on one of the band-matching coefficients.

### 6.6.1.1 Pearson correlation

The Pearson product-moment correlation coefficient calculates the congruence between arrays of values, typically densitometric arrays. As it compares curves as a whole, it is independent of band definitions and thus is ideally suited for quick comparison of patterns without having to edit bands first. It is largely insensitive to relative concentrations, but is sensitive to differences in background. The Pearson correlation is an objective coefficient in that it does not suffer from typical peak/shoulder mismatches as often found when using band-matching coefficients.

This cluster analysis is started from the main window using the currently selected list by choosing ***Comparison|Clustering (correlation)*** (or Ctrl+C). The *Pearson* correlation coefficient is applied on all points of the densitometric curve which fall within the *active zones* of the list (see 6.3.3.3).

The option ***Optimization*** allows you to perform an ultimate track-to-track correction to compensate for the smallest remaining misalignments.

(1) ***No optimization***. This option is disabled and the correlation is calculated on the unaltered tracks.

(2) ***Global optimization***. Before calculating the correlation matrix, an average pattern is calculated from the list and the tracks are shifted (between certain bounds) until they show the highest correlation with the average pattern. This option works fine for relatively small lists or otherwise for similar patterns. It has the advantage that the corrections applied to each patterns are also used when the tracks are displayed next to the dendrogram.

(3) ***Fine optimization***. While calculating the correlation for each couple of tracks, one of both tracks is shifted with respect to the other (between certain bounds) until they have maximal correlation (this option considerably slows down the calculation). Fine optimization is irrespective of the size of the list and the nature of the patterns, but since the alignment is different from pair to pair, the alignments cannot be displayed.

The ***Clustering method*** box allows you the select between the unweighted pair group method using arithmetic averages (UPGMA), Ward's clustering algorithm and the Neighbour Joining method (see 6.6.2.1).

If **<*OK*>** is pressed, the similarity calculation and cluster analysis is started. If the list contains many entries, this may ask several minutes. The calculations can run in background mode while you are running any application. **Do not change gel information, modify the current selection list or start a second cluster analysis while the clustering is in progress!**

*NOTE: When **Clustering (correlation)** is selected in the short menu mode, no dialog box is shown and the cluster analysis is started immediately using the following defaults: Optimization-Fine and UPGMA.*

### 6.6.1.2 Band-based similarity coefficients

*NOTE: The coefficients described below require that bands have been assigned to the tracks to be compared (see 6.11.1).*

The band-based cluster analysis is started from the main window on the currently selected list by choosing ***Comparison|Clustering (bands)*** (or Ctrl+B). The *band-matching* coefficients are applied on all points of the tracks which fall within the *active zones* of the list (see 6.3.3.3). The similarity between two tracks can be calculated in four ways:

(1) Coefficient of *Jaccard* ($S_J$) using band positions. For each couple of tracks, $S_J$ divides the number of corresponding bands by the total number of bands in both tracks (the corresponding ones plus the track-specific ones for each track):

$$\frac{n_{AB}}{n_A + n_b - n_{AB}}$$

$n_{AB}$ is the number of bands common for A and B, $n_A$ is the total number of bands in A, and $n_B$ is the total number of bands in B

(2) The *Dice* coefficient ($S_D$) which is derived from and very similar to the coefficient of Jaccard but gives more weight to matching bands:

$$\frac{2n_{AB}}{n_A + n_B}$$

(3) Area-sensitive coefficient. This is a more sophisticated similarity coefficient, taking into account the correspondence of bands expressed as $S_J$ as well as the differences of the relative areas under each of the corresponding bands (which is an indication of the concentration of the corresponding band).

$$\frac{\Sigma\Delta S_{AB}}{n_A + n_B - n_{AB}}$$

where $\Sigma\Delta S_{AB}$ is defined as

$$\Sigma \frac{\alpha}{\alpha + |S_{Ai} - S_{Bi}|}$$

$\alpha$ is a constant; $|S_{A,i}-S_{B,i}|$ is the absolute difference between the areas of band i on A and B, and i ranges from 1 to $n_{AB}$

Differences in band areas of corresponding bands in both patterns are penalized accordingly. When the areas of all corresponding bands of two tracks are equal, this coefficient is reduced to a simple Jaccard coefficient.

(4) A "Fuzzy logic" coefficient, which is based on the Jaccard coefficient but assigns scores to corresponding bands proportional to their degree of overlap. If two bands occur at exactly the same position, the score for these bands will be the same as using the Jaccard coefficient; the score will decrease proportionally to the distance between the bands and will be zero when the distance is larger that the allowed tolerance.

(5) Jeffrey's *x* coefficient where the similarity between two patterns is calculated as follows:

$$\frac{n_{AB}}{n_A} + \frac{n_{AB}}{n_B}$$

This coefficient is very similar, but not identical, to the Dice coefficient.

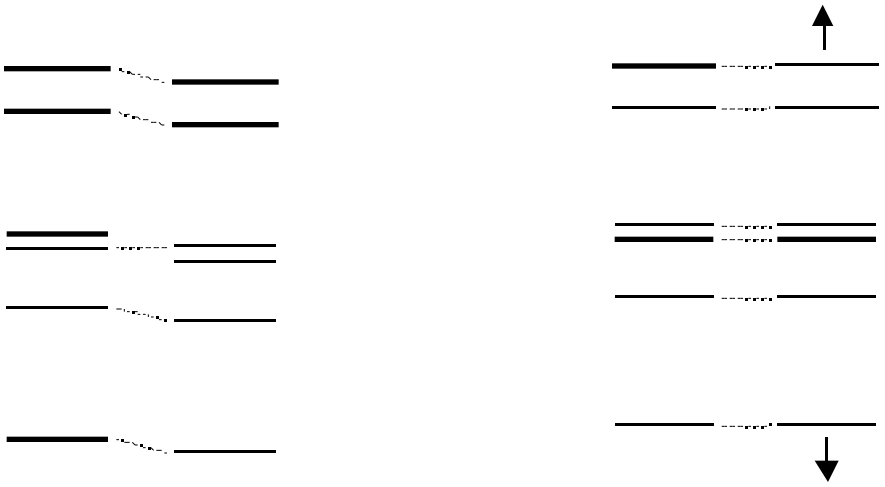
The tolerance on the positional differences between two bands to be considered as matching, is a parameter which can be changed by the user. It can be set in the *Band settings window*, called from the main Molecular Analyst Fingerprinting software menu with ***System|Band settings*** (or press F9) or by pressing the <***Position tolerance...***> button. The dialog box which is displayed contains the *Band comparison settings*, where you can specify position tolerance which is allowed between two bands to be matching. This value is expressed as a percentage of the total length of the pattern (e.g. a tolerance of 1% of a pattern of 500 points means a shift of maximally 5 points which is allowed to occur between two corresponding bands). Further, an linear increase of the tolerance towards the end of the pattern can be defined, also in percentage of the track length. The latter feature may be interesting for gels that have larger error in the low molecular size zone.

In addition to, and independently of the positional tolerance is the ***Optimization*** option available in the ***Clustering (bands)*** dialog box. Similar as for the

optimization explained in 6.6.1.1, an ultimate track-to-track alignment is performed.

To illustrate what the usefulness of this feature is in addition to the position tolerance, two identical patterns are shown in the figures below. As shown in (1), the patterns are not aligned perfectly, and when no optimization is applied, the closest of bands 3 and 4 will be matched. Band 3 in the first and band 4 in the second pattern remain unmatched. Irrespective of the tolerance allowed, these bands will always remain unmatched. Enabling the optimization, the result looks like in (2), again irrespective of the tolerance chosen.

(1) (2)



When the ***Optimization*** option is enabled, the calculation time will be considerably longer.

Select the clustering method and start the calculations as explained in 6.6.1.1.

*NOTE: When **Clustering (bands)** is selected in the short menu mode, no dialog box is shown and the cluster analysis is started immediately using the following defaults: Dice coefficient, Optimization and UPGMA. As position tolerance settings for matching bands, the default value of 1.2% with no increase in tolerance and no area threshold is taken.*

## 6.6.2 Visualization of the groupings

### 6.6.2.1 The dendrogram

When the calculations are completed, a *dendrogram window* (info-box supported) is presented, showing the resulting dendrogram or "tree" (i.e. a

hierarchic representation of linkage levels between pairs of individuals or groups).

In addition to the widely used hierarchic UPGMA clustering algorithm, Molecular Analyst Fingerprinting software offers the less common algorithm of Ward (Ward, 1963, *J. Am. Statist. Assoc*., **58**, 236-244) and the Neighbour Joining method (abbreviated NJ) (Saitou & Nei, 1987, *Mol. Biol. Evol*., **4**, 406-425).

The conceptual difference between UPGMA and Ward clustering is that the Ward algorithm minimizes the overall deviation of the dendrogram from the original matrix of similarities. This implies that, if one recalculates a similarity matrix from the linkage levels indicated on the dendrogram, the overall difference with the basic matrix (i.e. the "cophenetic correlation") will be less than by UPGMA.

The most important difference between NJ on the one hand and UPGMA and Ward's method on the other hand lies in the different interpretation of the tree. In UPGMA, the level of the branch which links two "Operational Taxonomic Units" (OTU's) determines the correlation between the OTU's. In NJ however, the summed lengths (in the *horizontal* direction) of all branches which have to be followed if one goes from one track to another, indicate the distance between two OTU's. This causes the OTU's not to occur at the same vertical position, and in general, the tree offers a more faithful representation of the original matrix of similarities, although less easy to interpret. In combination with similarities based on ribosomal RNA alignment data, a NJ tree is often considered as a biological evolution tree, which means that taxa on a longer branch are evolved further than the ones on short branches. Obviously, one should be very careful with that kind of interpretations.

The NJ method tries to minimize the total branch length of the tree, i.e. the "total evolution distance". For this reason, this class of trees is often referred to as "minimum evolution trees". To emphasize the conceptual difference between the UPGMA and Ward clustering algorithms on the one hand and a NJ on the other hand, the branches of NJ trees in Molecular Analyst Fingerprinting software are drawn in fluent curves instead of perpendicular lines.

The similarity scale of the dendrogram is drawn at the top of the window, while the bottom line of the window shows additional information about the calculation methods. A red cursor can be moved over the branches of the tree by pressing the left mouse button, holding the mouse pointer on a branch link or by using the arrow keys (left arrow key to move towards the root; up or down arrow keys to move towards higher branches in either direction). The similarity level of the cursor position is shown in the upper right corner.

A hardcopy can be made by ***Print|Full size***, which generates a large tree that may contain several pages, or by ***Print|Fit on page***, which rescales the tree to

fit on one page. A possibility to export the entry code fields of any branch of the dendrogram to the Windows clipboard, respecting the order of occurrence as in the dendrogram, is given by ***Print|Export branch to clipboard***. Select the root in order to export the entry codes of the complete dendrogram.

The branches linked at the cursor position can be swapped using ***Arrange|Swap***. In this way, closely related groups (see the dark regions on the similarity matrix) can be brought closer to each other. If the dendrogram is very large and contains many homogeneous subgroups, the structure can be simplified or made more transparent using ***Arrange|Abridge***, which reduces all subbranches linked at the cursor position into one smaller triangle. Choosing ***Arrange|Abridge*** a second time recovers the detailed substructure of the tree, at least when the cursor is positioned on the abridged group. ***Arrange|Correlation minimum*** makes it possible to define the minimal correlation value the scale should start with. By default, Molecular Analyst Fingerprinting software automatically determines the correlation of the root of the tree (the lowest linkage level) as starting point and rescales the tree accordingly. For comparing dendrograms, it can sometimes be convenient to have the same scaling. The size of the tree shown on the screen can be changed using ***Size|Normal*** (i.e. the default size), ***Size|Small***, ***Size|Very small*** or ***Size|Fit in window***, which forces the tree to fit in the current dimensions of the window. The labels assigned to each track are changed in the Entry description dialog box (press F9 or see 8.3).

> *NOTE: Display the full descriptive information of each track by pressing Ctrl+left mouse button on its name (Info-box!), or create its gelstrip using Shift+left mouse button.*

***Show|Error flags*** offers an invaluable feature for interpreting dendrograms, because when this feature is checked, error flags derived from the standard deviations with respect to the similarity matrix are shown for every branch of the dendrogram. This representation gives an easy and reliable visualization of the significance and stability of every cluster from the dendrogram. The exact error values at the the branch at cursor position is indicated in the upper left corner. The error flags are also shown when the graph is printed. This option is not available in the NJ clustering.

***Show|Tracks*** allows the reconstructed pattern images to be displayed next to the dendrogram and the entry names. Note thet you can resize the window if you have a larger screen, or this function can be used in combination with the Size item in the menu, to reveal nice overviews of large dendrograms. ***Show|Gelstrips*** allows you to show the original normalized 2D gel strips next to the dendrogram. The bands can be displayed on the pattern images (if defined) with the ***Show|Bands*** command. The option ***Show|Metrics scale*** allows a scale of the molecular sizes to be shown on top of the fingerprints (marked with ✓). Obviously, the scale will only be shown when the Metrics are defined for the database in use (see 6.3.4).

With ***Show|Image of branch*** you can display an image of all patterns linked below the branch which is currently selected by the cursor. The patterns are displayed in the order they occur on the dendrogram, which saves you the work of rearranging the patterns manually on the list image.

***List|Add branch*** and ***List|Replace with branch*** allows the subbranch at the cursor's position to be added to the selection list in the main window, or to replace the selection list with the branch.

Using the command ***Print|Preview mode***, the *Dendrogram* window changes into a *Preview* window, where the same dendrogram is presented on a white page of paper (sizes as specified in the printer setup) in true WYSIWYG ("what you see is what you get") mode. The yellow borderlines drawn on the sheet of paper can be dragged with the mouse, to adjust the borders of the dendrogram, gelstrips, and text as you wish. Use ***Preview|Zoom in*** and ***Preview|Zoom out***, to view the entire page or to zoom in on specific details. With ***Preview|Page up*** and ***Preview|Page down*** (or using the **PgUp** and **PgDn** keys) you can preview the previous or next pages. ***Preview|Edit mode*** is to return to the normal *Edit* mode. With the ***Size*** menu, you can change the thickness of the gelstrips and the corresponding size of the dendrogram, or fit the whole dendrogram to the page using ***Size|Fit on page***. The ***Show*** menu is the same as for the *Edit* mode. With ***Print|Print all pages***, the dendrogram is printed to the default printing device.

> ***NOTE****: The* Preview *mode does not print to the Molecular Analyst Fingerprinting software printer manager, but sends print jobs directly to the printer instead.*

It is sometimes useful to change the paper orientation when printing dendrograms with or without gelstrips. ***Print|Printer setup*** allows you to change the settings of the default printer, including paper orientation. With ***Print|Copy to clipboard***, a WYSIWYG copy of the current page is generated in the Windows Clipboard. Use ***Size|Fit on page*** to fit the whole dendrogram on one page, if you want a clipboard copy of the whole dendrogram.

#### 6.6.2.2 The similarity matrix

***Show|Matrix*** in the dendrogram window draws a *matrix window*, showing a similarity matrix, rearranged according to the cluster analysis. The similarities between all pairs of tracks are shown in matrix form and visualized as gray-shaded squares (the darker the square, the higher the correlation). If ***Shades*** is chosen, the *index table* giving the similarity intervals for all gray shades is shown. These intervals can be adjusted using the scroll bars.

In case of large matrices only a part is shown in the window. ***Focus*** can be used to display a box containing a scaled image of the whole matrix. A rectangle marks the part of the matrix currently shown in the window and can

be moved through the image using the mouse. Position the rectangle to select a part of the matrix and click the left mouse button. The selected part of the matrix is now shown in the matrix window. In addition, ***Layout|Small*** produces a larger view, but the corresponding names of the track are not shown. ***Layout|Normal*** restores the default state.

***Layout|Show numerical values*** draws an enlarged matrix with the similarity values written in the shaded squares. ***Layout|Original ordering*** will display the matrix respecting the original sequence of patterns as in the selection list. This feature will help you comparing matrices of the same patterns after calculation using different coefficients, band tolerance settings, optimizations etc. Labels for the tracks are changed in the Entry description dialog box (see 8.3).

***Layout|Show dendrogram correlations*** allows you to toggle between the real similarity values of the original similarity matrix (disabled) and the simplified similarity values derived from the linkage levels between all the entries on the dendrogram. While enabling this derived representation, the program calculates and displays the *Cophenetic Correlation* value for the dendrogram. The cophenetic correlation is the product-moment correlation between all original matrix similarities and all corresponding similarity values derived from the dendrogram. If the dendrogram faithfully represents the similarity matrix, the cophenetic correlation will be high (e.g. more than 90%) whereas it will be low (e.g. 70% or less) when the matrix cannot be represented faithfully by an UPGMA dendrogram. As such, the cophenetic correlation is an excellent indication of a quality of a cluster analysis and it is worth mentioning this value with a published dendrogram. The aim of the cophenetic correlation is similar as of the *Error flags* on the dendrogram, which are based on the standard deviations of the similarity values derived from the dendrogram compared to the original values (see section 6.6.2.1).

***Print|Full size*** and ***Print|Fit on page*** can be used to print the matrix. ***Print|Print numerical values*** generates a printout of the similarity matrix with the similarities shown as numerical values, whereas ***Print|Export numerical values*** creates a text file "COR.TXT" in the Molecular Analyst Fingerprinting software home directory containing all data of the similarity matrix in ASCII text format, which can be imported in other applications.

Press the left mouse button on a similarity square to reveal the exact similarity value and the code of the two tracks from which the similarity value was calculated.

#### 6.6.2.3 Storage of cluster analyses on disk

Each time a cluster analysis is calculated on a list, the results, i.e. the matrix of correlations and the dendrogram, are automatically saved on hard disk. ***Comparison|Load cluster analysis*** from the main window creates a dialog box showing all the cluster analyses on the hard disk, which have the same name as

the lists from which they were derived. Double click on an analysis name to open the corresponding clustering window. In this way, it is not necessary to recalculate grouping analyses. See section 6.3.3.1 for information about file management of cluster analyses saved to hard disk.

#### 6.6.2.4 Calculating dendrograms from similarity matrices stored on disk

The menu item ***Comparison|Clustering (load matrix)*** allows you to start calculating a dendrogram from a previously calculated similarity matrix which is stored on disk. Since calculating the similarities is often the most time-consuming part of the analysis, you can execute and display several clustering algorithms without having to recalculate the matrix. When this option is called, a survey of all available similarity matrices on disk is shown (the names are the same as the corresponding list names). Select a matrix and press **<*OK*>**. Then select the clustering algorithm in the dialog box that appears (the optimization does not apply here) to calculate the dendrogram.

## 6.7 Clustering databases

The “Clustering databases” tool is the result of a combination of the fast and efficient database management achieved in Molecular Analyst Fingerprinting software and years of research in PC-programming, which has culminated in the most powerful cluster analysis software presently available. The “Clustering databases” are restricted to the Dice and Jaccard coefficients for band matching, whereas UPGMA is the sole clustering algorithm. This limitation is due to the special computing techniques applied for optimizing both speed and memory management.

With the Clustering databases, you will be able to cluster any database into a single dendrogram, irrespective of its size (up to 7000 patterns tested). All functions are especially designed to interact dynamically between the dendrogram, the similarity matrix and the database. It is in fact your database, which is not represented as a list of patterns or a set of gels, but as a dendrogram, complete with matrix and gelstrips or bar graphs of the patterns. Selected patterns (lists) are highlighted on the dendrogram, whereas groups selected in the dendrogram become automatically selected into lists.

In addition, the Clustering Databases tool offers "*incremental clustering*". This unique feature allows one to add a (relatively small) set of new fingerprints to an existing cluster analysis, without having to recalculate all correlations and all branches of the dendrogram. In this way, it only takes a few seconds to add a new fingerprint to a dendrogram containing several thousands of entries.

### 6.7.1 Creating a new ClusterBase

Molecular Analyst Fingerprinting software can maintain many Clustering Databases at a time for the same user. The list of existing clusterbases is shown in the *ClusterBases List* subwindow in the Molecular Analyst Fingerprinting software main window. Initially, this list is empty, and a new clusterbase can be created using the "***New***" button in the ClusterBases List subwindow, or using the menu command ***Comparison|Clustering Databases|Create new*** from the main window. This pops up the *ClusterBases Creation dialog box*, which allows one to define the various settings for the new clusterbase:

- The name of the new ClusterBase. This name should be a valid DOS filename, without spaces and punctuation marks, limited to 8 characters and without extension.
- The type of correlation which is applied for the creation of the UPGMA dendrogram. One can choose between Pearson product-moment correlation (based on the densitometric curves) and the Dice or Jaccard coefficient (based on the bands). See section 6.6 for further details on these correlations.
- An optional optimization. Check this option for a detailed track-to-track correction of mismatches (see also section 6.6 for more information about optimization).

Pressing the **<*OK*>** button of this dialog box creates a new clusterbase, which is initially empty. This ClusterBase is added to the list of existing ones. For each clusterbase in the list, the number of entries is shown together with details about the type of correlation used and about the space used on the hard disk.

In the *ClusterBases List window*, one can open an existing clusterbase by selecting it in the list and using the menu option *Clusterbase*|*Edit*, or by double-clicking with the left mouse button on the ClusterBase entry in the list.

### 6.7.2 Editing a ClusterBase

In the *ClusterBases List window*, one can open an existing ClusterBase by selecting it in the list and using the menu option ***Comparison|Clustering databases|Edit*** (or by double-clicking with the left mouse button on the ClusterBase entry in the list). This action creates a *ClusterBase Window*. This window is divided in four parts (from left to right):

- A small *dendrogram overview* (outermost left)
- The *dendrogram* itself, together with the names of the database entries (can be customized using F9 to call the Entry description dialog box, see 8.3).
- The *fingerprint image field*. If a band-based correlation is used (Dice or Jaccard), this field initially shows the *Bandstrips* of the fingerprints (*Bandstrips* are small maps showing the positions of all bands on the

fingerprint). This part of the window is also used to show the *Gelstrips* of the fingerprints

- The *matrix of correlations* of the entries in the ClusterBase

Of course, all of these fields are empty if the ClusterBase contains no entries.

#### 6.7.2.1 Adding new fingerprints to a ClusterBase

A selection list, created in the Molecular Analyst Fingerprinting software database manager (see section 6.3.2), can be added to an existing Clusterbase using the ***Clustering|Add current list*** menu option in the *ClusterBase Window*. This command adds those entries of the list to the dendrogram which are not yet a member of the ClusterBase. The first time a selection list is added to an empty ClusterBase, the ClusterBase's correlation zones (see 6.3.3.3) and band tolerance (for Dice and Jaccard coefficients) are set to the values which are valid at that moment (F7; see Band settings dialog box in 6.6.1.2). When new entries are added to a non-empty clusterbase, the parameters valid for that clusterbase are adopted.

In case the number of new entries is small compared to the number of entries which are already present in the ClusterBase, the program prompts a dialog box asking the user whether the dendrogram should be completely rebuilt or not. Pressing ***<Yes>*** causes the program to completely recalculate the dendrogram. This option is slower, but exact. Answering ***<No>*** uses an *incremental cluster analysis* algorithm, which *merges* the new fingerprint to the existing clustering. Instead of recalculating the whole dendrogram, the algorithm places the new entry in the branch where it fits best, and recalculates that branch. This option is much faster, but does not always produce the exact dendrogram topology, especially when the number of new entries is large compared to the size of the clusterbase. However, when one or a few entries are added to a clusterbase of say 1000 patterns, the error will be negligible. The top part of the window shows the number of entries in the ClusterBase, and the number of entries that are merged using incremental clustering. If this number grows too large, the dendrogram becomes less reliable. At any time, the user can decide to recalculate the whole cluster analysis by using the ***Clustering|Rebuild clustering*** menu option. When this is done, the number of merged entries is reset to zero and the dendrogram again has the correct structure.

#### 6.7.2.2 Changing the appearance of the dendrogram

The ***Layout|Big***, ***Layout|Normal***, ***Layout|Small*** and ***Layout|Very small*** menu options determine the vertical distance between the fingerprint entries on the screen. In addition, the user can adjust the horizontal position of the vertical separation line between the fingerprint field and the matrix of correlations by clicking on it with the left mouse button and dragging it to its new position. The style of the information labels next to the fingerprint can be changed by

calling the *Entry description Dialog box* by pressing the F9 key (see also section 8.3).

A particular branch of the dendrogram can be *highlighted* by clicking with the left mouse button on the branch point. The whole highlighted branch, together with the Bandstrips of the fingerprints belonging to that branch, are marked red. One can highlight the whole dendrogram by pressing the left mouse button on the root branch, or by calling ***Clustering|Highlight root***. The two sub-branches which originate from the highlighted branch can be swapped using the ***Clustering|Swap highlighted branches*** menu option.

When a ClusterBase is opened on the screen, no Gelstrips are initially shown. The user can display the Gelstrips on a highlighted branch of the dendrogram by using the ***Layout|Load highlighted gelstrips*** command, and remove them using ***Layout|Remove highlighted gelstrips***. Note that all Gelstrips of the whole ClusterBase can be shown by using ***Clustering|Highlight root***, and then ***Layout|Load highlighted gelstrips***.

The definition of the shadings in the correlation matrix can be changed using the ***Layout|Shades*** command. This pops up a *Correlation shading window*, which displays the correlation limits of all grey shadings. The user can change the limits by dragging them to a new position. ***Preview*** previews the new correlation matrix, while ***OK*** closes the window.

#### 6.7.2.3 Selection lists in ClusterBases

The real power of ClusterBases consists in its relation to selection lists (see also section 6.3.2 for more information on selection lists). At any time, the selection lists which is currently present in the Molecular Analyst Fingerprinting software database is also visualized in the dendrogram of a Clusterbase: the names of the selected fingerprints are shown on a yellow background. Of course, a selected fingerprint which is not yet a member of the ClusterBase, is not shown in this dendrogram.

In addition, the user can manually select or unselect a whole branch of fingerprints directly on a ClusterBase dendrogram by clicking the right mouse button on the corresponding branch origin.

#### 6.7.2.4 Division of fingerprints into groups

A Clusterbase also offers the possibility to divide the various fingerprints into groups. Each group is represented by three attributes: a name, a colour and a small graphical symbol. On the dendrogram image, each member of a certain

group is marked by either the colour code or the graphical symbol (if printed). These groups are very useful to mark members of (a) particular subgroup(s) of the database.

The attributes of the various groups can be changed using the command ***Layout|Groups***, which creates a new window showing a list of all groups. In this window, double-clicking on one of the group names pops up a dialog box which allows you to change the group name or the graphics symbol associated with that group. The entries of the current selection list can be assigned to a particular group by using the menu item ***Layout|Assign list to group***, and selecting one of the available groups therefrom.

#### 6.7.2.5 Printing the dendrogram.

The ***Print|Clustering*** menu option is used to create a print-out of the dendrogram of the ClusterBase. This command pops up a dialog box which prompts for the following print settings:

***Vertical spacing***. This parameter determines the vertical distance (in millimetres) between adjacent entries of the dendrogram. The estimated number of pages of the print-out is also given.

***Margins***. These parameters determine left and right margin of the paper (in cm). If maximum size is required, these values can be set to zero.

***Print bands***. If this option is checked, the Bandstrips are printed next to the dendrogram image (Bandstrips are small maps showing the positions of the band on the fingerprints). This option works only for Dice or Jaccard coefficients.

***Print Gelstrips***. This option determines whether or not Gelstrips are printed next to the dendrogram.

***Use colours***. Check this option if the print-out should appear in colours.

***Page guides.*** If this option is checked, small crop marks are printed on the pages. These lines can be used to align the pages to each other.

***Print names***. Check this option to print the fingerprint information labels next to the dendrogram.

***Print group codes***. This check box determines whether or not the group codes are printed next to the entries.

Pressing the **<*Setup*>** button offers the opportunity to change the printer's settings, while the **<*OK*>** button is used to start printing.

#### 6.7.2.6 Printing the correlation matrix.

Use ***Print|Matrix*** menu option to create a print-out of the correlation matrix of the ClusterBase. This creates a dialog box which prompts for the following settings and options:

***Spacing***. This parameter determines the size (in millimetres) of the correlation blocks.

***Margins***. This parameters determines left margin of the paper (in cm). If maximum size is required, this value can be set to zero.

***Print names***. Check this option to print the entry information labels next to the matrix.

***Use colours***. Check this option if the print-out should appear in colours.

***Page guides***. If this option is checked, small crop marks are printed on the pages. These lines can be used to align the pages to each other.

Pressing the <***Setup***> button offers the opportunity to change the printer's settings, while the <***OK***> button is used to start printing the matrix.

## 6.8 Identifying with database patterns

*NOTE: The Identification tools are part of the **Identification module**.*

There are three methods for identification available in Molecular Analyst Fingerprinting software. The simplest one, as described in this section, is to compare and identify a list of unknown patterns against a list of database patterns stored on disk. A more sophisticated method is to identify the list of unknown patterns against an identification library (see 6.9). The third method is intended for identification based on band matching only, which assigns importances to bands within known groups, based on their frequencies, in order to arrive at a statistical identification of unknown patterns (see section 6.10).

### 6.8.1 Creating lists of database patterns to identify with

You should first have a list of database patterns on hard disk which you will use to identify with. To create such a list, use the ***Search*** option to add all database patterns or only tracks of certain species or groups (see 6.3.2 and automatic topic search in 6.3.2.2). Note that the more patterns you add to the list, the longer the identification will take. Save the list to disk as described in 6.3.3.1. Now clear the list from the list table with ***List|Clear list*** or the F4 key.

### 6.8.2 Identifying a list of patterns

Create a list of new or unknown tracks to be identified. Up to 50 tracks can be identified in batch; if the list is larger, it will be truncated so that you should identify in two batches. Choose ***Comparison|Identification (with list)*** from the main menu. Select the list intended for identification from the table of available lists and the similarity coefficient for identification (see also 6.6.1.1 and 6.6.1.2). The scroll bar ***Minimum correlation*** allows you to display only database patterns that show correlation higher than the value specified. If the scroll bar is moved completely to the left, no limitation by correlation will exist ("No" appears). The scroll bar ***Entry limitation*** allows you to specify a maximal number of most likely database patterns to be displayed for each identified pattern.

### 6.8.3 Global identification report

Press **<*OK*>** to start the identification. The calculations can run in background mode as long as you do not try to change the current list or start a second identification. When completed, the best matching database pattern (within the used list) is shown for each unknown pattern. This report can be printed using the ***Print*** menu.

### 6.8.4 Detailed identification report

With ***Detailed*** or by double-clicking on a track, a detailed identification report for the selected pattern is shown, listing the best matching database patterns (within the used list), in decreasing order of similarity. The number of entries displayed will be according to the limitations defined in the Identification dialog box. Either ***Entry limitation*** or the ***Minimum correlation*** specified will be the limiting factor. In case the Entry limitation is the limiting factor, you can choose ***All*** in the menu to cause all the patterns to be listed, scoring above the minimal correlation, in decreasing order of similarity. The menu option ***List|Select all entries*** allows you to select all the patterns in a list that match according to your specified criteria. In addition, you can limit the number to be selected by placing the cursor bar on any pattern and choosing ***List|Select down to cursor***. Print this report using the ***Print*** option.

## 6.9 Identification using libraries

### 6.9.1 Construction of libraries for identification

The Molecular Analyst Fingerprinting software library manager (***Identification module***) can generate and manage up to 100 libraries for each user. A Molecular Analyst Fingerprinting software library consists of user-defined

units. A virtually unlimited number of units can be defined within one library. Each unit represents a homogeneous electrophoretic type and may contain one or more (max. 15) representative patterns. Since Molecular Analyst Fingerprinting software offers extensive grouping analysis functions, it is easy to define homogeneous groups and representative tracks for these.

A new library is created by selecting ***Library|Create*** from the main window. An input box appears, asking for the library name. This should be a valid directory name of max. 8 characters, without extension, periods or spaces.

> ***NOTE:*** *Libraries are saved as subdirectories of the library directory, which can be specified for each user in the* ***User Setup*** *menu to be found in the Startup program (8.2.1).*

#### 6.9.1.1 Editing a library

***Lirary|Edit*** from the main window shows a list of all existing libraries. By double-clicking on one name a *library window* is shown, containing a list of all units which currently exist in within the library (initially the list is empty). A selection bar can be moved through this list. The bottom line of the window shows the standard and resolution of all tracks of this library (it is impossible to mix tracks which have different standards or resolutions into one library). ***Library|Print*** creates a detailed print-out of all units from the library. ***Library|Delete*** removes the library from the hard disk (for safety reasons, this option works only for an empty library). A new unit is added to the library by selecting ***Unit|Add new***. This calls a dialog box prompting for the name of the new unit (maximally 40 characters). An existing unit is removed from disk by selecting it and calling ***Unit|Delete***. Double-click on a unit's name or select ***Unit|Open*** to display a *unit window*, which allows the selected unit to be edited.

#### 6.9.1.2 Editing a unit

The unit window shows a list of all entries which are defned within a library unit (initially the list is empty). A selection bar can be moved through this list. ***Rename*** assigns a new name to the unit, while ***Print*** creates a detailed print-out.

**a) Adding and removing entries.** If one or more tracks are to be added to the unit, first create a list containing the desired entries (section 6.3.2). ***Entries|Add list*** appends the current list to the list of entries of the unit. Note that the number of entries is restricted to 15. If this number is exceeded, the list is truncated. For compatibility reasons, only tracks that have the same standard and resolution as the existing entries can be added to a unit. ***Entries|Delete*** deletes the currently selected entry from the unit.

**b) Selecting active zones for a unit.** Units also have active zones associated, which are used for identification. ***Edit|Zones*** creates a zone window showing the currently selected entry, together with the current active zones. See section 6.3.3.3 about enabling and disabling zones for comparison. Note that the zones selected in the library are specific for each unit, which allows the user to create very specific identification cases.

**c) Marking bands on the densitometric curve.** Identification of unknown tracks also can be based on a set of predefined bands, specifically assigned to each unit of a library. ***Edit|Bands*** creates a *band edit window*, allowing you to edit the set of bands (see also 6.11.1). This window shows the densitometric curve of the currently selected track, with the defined bands marked as Gaussian curves (initially, there are no bands marked). A position ruler can be moved through the window by clicking the left mouse button and moving the mouse. The status bar of the window shows the number of selected bands and the position of the selected band in data points and metrical units (if present). A band is selected with the position rule. The Gaussian shape of the band is then bordered by three red squares. The area around the position ruler can be enlarged using ***View|Zoom in***. ***View|Zoom out*** restores the previous view. You can also press ENTER to toggle between the normal and enlarged mode. In the enlarged mode it is possible to scroll through the densitogram.

A band is added to the band list by moving the position ruler to its maximum and selecting ***Add|***. To remove an existing band, select the band with position ruler and choose ***Remove|***, or simply press DEL. In addition, an automatic band search is performed by ***Search|***. This automatic search is executed according to the band search settings specified in the band settings window (see 6.6.1.2).

In the enlarged mode, you can adjust the Gaussian shape of a band by dragging one of the three bordering squares with the left mouse button. ***Print|Text report*** creates a table of all selected bands, whereas ***Print|Graphics image*** prints the band positions on the densitometric curve. Choose ***OK*** to update the band list of the unit and exit the band edit window.

***Show*** gives an image of the patterns of the library unit, with the active zone(s) marked as gray background and the defined bands shown as lines. This image can be printed.

*NOTE: Marking zones and/or bands can be done on the "mean pattern" of the library unit. This patterns is usually the most representative of the unit!*

## 6.9.2 Identification against a library

*NOTE: The Library Manager and Identification tools are part of the* ***Identification module****.*

First select a list of new or unknown tracks to be identified. Up to 50 tracks can be identified in batch; if the list is larger, it will be truncated and you should identify in two batches. Lists of new or unknown tracks are identified by the option ***Comparison|Identification (with library)*** from the main window. A list box shows all existing libraries, allowing you to select which library to use for identification. A radio button box offers the choise between seven possible similarity measures. The first three use the Pearson correlation. To understand the difference between the choices, you should realize that the **library units** will be considered as identification cases, not the individual patterns within the units. When an unknown pattern resembles best a given unit, the name which you have assigned to the unit will appear as identification. However, since a unit may consist of more than one pattern, there are several ways to calculate the overall similarity of a unit with the unknown pattern.

(1) the average value of the correlation with all entries of the unit;

(2) the highest among the correlations found with all entries of the unit;

(3) Pearson correlation with the mean entry of the unit;

The other choices are all band-based similarity coefficients, and should only be applied when bands were defined for the library units and the unknowns.

(4) and (5) are the $S_J$ and $S_D$ value of band positions defined in the Library manager and for the unknowns;

(6) Similarity based on both band position or and relative area;

(7) a "fuzzy logic" coefficient, penalizing distance between close bands.

For the Pearson correlation, i.e. options (1), (2) and (3), the fine optimization (see section 6.6.1.1) is applied and a second radio button box allows you to determine whether the correlation will be based on the zones specified for each unit separately (see 6.9.1.2) or on the zones provided with the list of unknown tracks (see 6.3.3.3).

For coefficients 4 to 7, (see also 6.6.1.2), when the radio button ***Library bands only*** is checked, the identification will only take account of the bands that occur on the patterns of the library units (see 6.9.1.2) to assign scores to the unknowns, and will not penalize additional bands that occur on the unknown patterns but not on the library patterns.

Press **<*OK*>** to start the identification.

> ***NOTE:*** *(1) Only tracks normalized with the same standard and having the same resolution as the specified library can be identified.*

*(2) When the option "Library bands only" is checked, one can mathematically prove that the Dice coefficient is reduced to the Jaccard coefficient.*

#### 6.9.2.1 Global identification report

When the identification is finished, an *identification window* is shown, listing the best matching unit for each track, as well as the corresponding similarity (info-box supported). A selection bar can be moved through the entries. ***Print*** creates a print-out of the data. When the third option for identification is applied (correlation with mean), a coloured square at the right side of the similarity value for each track indicates the quality of the identification (only when more than two entries were defined for the library unit). Similar to the database quality control indication (see 6.4.1), the colour ranges from red, over orange, yellow to pale and dark green. Red means an unreliable identification whereas green indicates the most confident identification. The origin of this second indication of quality is explained below.

#### 6.9.2.2 Detailed identification report

Choose ***Detailed*** or double-click on a track name to create a detailed *report window* of the selected track, showing the 10 best matching units with the corresponding correlation in decreasing order, and the identification quality indication (see below). ***Print*** creates a print-out.

#### 6.9.2.3 Statistics of an identifcation

In the detailed report window, you can select one of the 10 best matching library units and choose ***Statistics*** from the menu to reveal the *identification statistics window*. The statistics apply only to library units for which more than two entries were defined. This report is divided in three parts.

The left part shows a blue curve, which represents the variance of the entries of the library unit from the unit's mean track along the densitometric curves. The red curve is the variance of the unknown track from the unit's mean track. The colours of the bar at the right side of the variance curves indicate the deviation of the red curve with respect to the blue curve. The scientific interpretation of these statistics is as follows.

If more than one track is defined in a library unit, there may occur some regions with low variability on the patterns, as well as regions with high variability. The regions with high variability are bands on the blue curve. The regions of the unknown patterns that deviate from the mean unit pattern are bands on the red curve. If the bands on the blue and red curves are the same, one can conclude that the variable zones within the library unit and on the unknown are same, and even though the correlation may be low, the

identification will probably be confident, since the regions in which the unknown differs from the library unit are not representative of the unit. This is reflected as green regions on the bar at the right side of the curves.

The active zones for identification are shown as gray background behind the patterns in the middle part of the window. The correlation of each pattern with the unknown is shown. AV is the unit's mean pattern (average), the sequential number of the other patterns corresponds to the sequence number in the Library Manager.

**A second-place identification with a green square is likely to be a more probable identification than a first-place identification with a yellow or reddish square!**

In the right part of the report, the graph marked (1) is a plot of the correlation of the unknown pattern with all unit patterns on the one hand (red) and the correlations between the unit patterns internally (blue). Ideally, the red lines should occur in the middle of the blue lines. The graph marked with (2) is similar, but showing only the correlations with the unit's mean pattern. Ideally, the red line should occur on top of the graph. The *probability ratio* is an important value expressing the overall position of the unknown pattern within the library unit. If this value is significantly larger than 1, the unknown is more deviant than the unit patterns on the average. If the value is less than 1, the unknown fits well in the library unit. The coloured squares shown on the identification reports reflect this ratio. A ratio less than 1 is shown as a dark green square next to the identification score, whereas a ratio of more than 2 will be shown as a red square.

With ***Print***|, you can create a hard copy of this report.

## 6.10 Identification using Identification Groups

It often occurs that a taxonomic group, which cannot be split into smaller subgroups, still has some remaining variability because it does not represent a real genetic clone. Electrophoretic study of a set of entries from such a group will result in fingerprints which have a number of “stable” bands in common. However, due to the remaining variability, there will be a few “unstable” bands which are only present a part of the samples. For identification purposes, it is interesting to rely more on the “stable” bands than on “unstable” or variable ones.

For this reason, Molecular Analyst Fingerprinting software offers a third alternative to the identification of unknown fingerprints, based on statistical occurrence of bands. For each taxonomic unit, an *Identification Group* is created, containing a collection of representative fingerprints, representing the

each Identification Group, the software counts the relative frequency of occurrence for each individual band on the fingerprints. During an identification, more weight is assigned to bands with a high occurrence frequency (typical bands). On the other hand, bands with a low occurrence frequency (atypical bands) have a small influence on the identification.

## 6.10.1 The construction of Identification Groups

The menu option ***Library|Edit Groups*** creates a *Groups list window*, which gives a list of all existing Identification groups. This window offers the possibility to

- Add a new Identification Group, using ***Group|Add new***. This menu option pops up a dialog box which prompts for the name of the new Identification Group.
- Delete an existing Identification Group, by selecting a Group name and calling ***Group|Delete***.
- Edit an existing Identification Group, using ***Group|Edit***. This option creates a *Group Edit Window*.

### 6.10.1.1 Assigning a list to an Identification Group

The *Group Edit Window* allows you to assign a list of representative fingerprints to the edited Group, and to calculate the statistical occurrence of the individual bands on the fingerprints. When a new Group is created, it will initially be empty. First create a selection list (see also section 6.3.2) containing the representative fingerprints, and assign it to the Identification Group using ***Group | Import from list***.

### 6.10.1.2 Calculating the band occurrence frequencies

After a list has been imported in the group, the occurrence frequencies of the bands can be calculated using the menu option ***Group|Calculate***. This command brings up a dialog box, which asks for four parameters:

- Peak smearing (in points). This parameter gives the maximal spread (tolerance) on the positions of the bands, used for the determination of the common band classes
- Optimization (in points). This parameters (further called $S_A$) determines the maximum shift which is applied during the optimization procedure (see also section 6.6.1.2; under *Optimization*)

- Stretching (in points). Apart from simply shifting the fingerprints to the left or to the right, the optimization procedure can linearly stretch or compress the fingerprints in order to reach a maximum match of the common bands. The maximum amount of stretching or compression allowed is given by the value of this parameter (further called $S_B$ )

- Minimal peak occurrence. This parameter determines the minimal relative offset of a band (in %) in order to be used for identification.

Pressing the **<*OK*>** button in this dialog box starts the calculation of the band occurrences, using the following steps:

1. The program divides all the bands on the fingerprints into classes of bands with the same run length (molecular size or whatever physical parameter). The peak smearing parameter determines the maximum deviation which may occur on the run length of the different bands in one class. In order to correct for small mismatches, the program also optimizes the band positions of the fingerprints by using a global shift or rescaling. Thereto, the run length *r* of all bands on each fingerprints is converted using the formula

$$r \blacklozenge \ (r + a)\frac{r_{tot} + b}{r_{tot}}$$

where $r_{tot}$ is the total run length of the gel, $-S_A \le a \le +S_A$ , and $-S_B \le b \le +S_b$ .

2. When the bands are divided into classes of constant run length, the relative occurrence of each band is calculated as the fraction of the fingerprints in the group sharing that band.

When these calculations are completed, the program displays a blue frequency curve in the *Group Edit window*. For each point on the gel, this curve gives the relative frequency of the bands having this particular run length (taking into account the value of the peak smearing). On top of this distribution curve, the various band classes are vertical green lines.

The user can manually select or unselect a band class. Only selected band classes will be used during the identification. The band classes which are currently selected are marked with a small green rectangle at the bottom of the frequency curve (initially, all the classes are selected). To alter the selection of a class, click with the right mouse button on the position line of the band class.

The menu option ***Group|Show Gelstrips*** loads the Gelstrips of the entries of the Identification Group, and shows them on the screen, together with the positions of the bands. In addition, the user can create a separate window containing a report of the run lengths, the metrics and the relative occurrences of all band classes in an Identification Group by using the menu option ***Group|Band***

***information***. The menu option ***Bands|Arrange by position*** shows the classes in order of increasing run length, while ***Bands|Arrange by occurrence*** displays the classes ordered to decreasing relative occurrence. ***Bands | Print*** creates a print-out of this report. The Group editor window and the Band statistics information dynamically interact with each other as a band selected in one window will automatically be selected in the other too. A band selected in the Group editor is marked with a yellow pointer; in the Band statistics window, it is marked with a selection bar.

#### 6.10.1.3 Searching an Identification Group in a selection list.

Molecular Analyst Fingerprinting software offers the possibility to search in a selection list for fingerprints that match a certain Identification Group. First create a selection list of fingerprints in the Database Manager (6.3.2), and then use the command ***Group|Search in selection list*** in the *Group Edit Window* of a particular Identification Group. A dialog box is shown which prompts for the tolerance on the band positions (again, in points on the gel). In addition, it shows a checkbox called “penalize extra bands”. If this option is not checked, bands that are present on the fingerprints and that are absent in the Identification Group, are ignored. When this option is checked, such bands will decrease the correlation. Press **<*OK*>** to start the search. Upon completion, a correlation value is assigned to each entry of the selection list, and a *Group Entry matching window* is created. This window lists all entries of the selection list, in order of decreasing correlation with the Group. Next to this list, a table containing all bands of the Group is shown. For each entry, the bands of the groups which are present on the fingerprint are checked in this table. The last column at the right hand side of the window shows the number of bands present on the fingerprint that not matched to any band class of the group.

The program also allows you to copy the best matching list entries into a new selection list. First place the cursor to the last entry which should be copied to the selection list, and then use the menu option ***Copy|Down to cursor***. This option creates a new selection list, containing all entries with a correlation higher than or equal to the entry at the cursor position. Alternatively, one can use ***Copy|Entries having selected peaks***. This option creates a new selection list, containing all the fingerprints which possess all the selected bands of the Identification Group. Selected bands are marked with a red line (initially all bands are selected). You can manually select or unselect bands of the Identification Group by clicking with the right mouse button on the top line of the band table (this line shows the relative occurrence and the metric of the bands).

### 6.10.2 Identification using Groups.

The main purpose of the Groups database is meaningful statistical identification based on relative band frequencies within biological entities

First, create a selection list containing the unknown fingerprints (6.3.2). Selecting the option ***Comparison|Identification with Groups*** in the main window creates a dialog box which prompts for the tolerance on the band positions (again, in points on the tracks). It also shows a checkbox which is called “penalize extra bands”. If this option is not checked, bands which are present on the fingerprints but which are absent in an Identification Group, are ignored. When this option is checked, the presence of such bands will decrease the correlation. Press **<*OK*>** to start the identification. Upon completion, a *Group Identification window* is created. This window lists all the list entries, together with the best matching Identification Group and the corresponding correlation.

Double-clicking on a particular list entry displays a *Group Identification Detailed Report window*, which gives more detailed information about this particular identification. This window gives a list of the 10 Groups which are most similar to the unknown pattern, ordered by decreasing correlation. Next to each Group, a small graph is displayed, showing the details of the band matching. The blue lines show the positions of the bands on the unknown fingerprint, while the red and green lines show the bands on the Groups. A band is marked green if it is found on the unknown pattern, and it is marked red when it does not correspond to a band on the unknown pattern. The height of the red and green lines is an indication of the relative occurrence of each band in the Group. Note that a green (matching) band and the corresponding band on the unknown (blue) do not always fall at exactly the same position. This is due to the tolerance allowed at the start of the identification.

## 6.11 Comparative Quantification

*NOTE: The features described in sections 6.11.2 and 6.11.3 are part of the **Quantification module**.*

### 6.11.1 Assigning bands to tracks

Molecular Analyst Fingerprinting software offers the possibility to define a set of bands (bands) to each normalized densitometric curve. Each band is approximated by a *Gaussian* curve, with variable position, height and width. Open the gel window of the gel and choose ***Gel|Assign bands*** from the menu to call the *band image window*, showing the image of all the patterns of the gel. Initially, no bands are defined for the patterns. With ***Show|Geltracks*** the 2D-gelstrips are shown instead of the reconstructed images.

With ***Bands|Auto search***, the program automatically searches bands through all the patterns and decomposes them into Gaussian curves. You can change the band search settings from the main Molecular Analyst Fingerprinting software menu using ***System|Band settings*** or by pressing F7. The *band search*

*filters* involve a minimal area as percentage of the total area of the pattern and a minimal profiling which is the elevation of the band with respect to the surrounding background, also as percentage. A more advanced tool based on deconvolution algorithms, ***Shoulder sensitivity,*** allows shoulder without maximum as well as doublets of bands with one maximum to be found. We recommend to start with a sensitivity of 5, but optimal parameters may depend on the type of gels analyzed. In addition, a ***Non-linear fit*** based on curve fitting allows the Gaussian shapes of densitometric curves to be decomposed in a much more refined way. The mathematics behind are quite exacting, which implies that this option is rather slow!

The defined bands are marked with a red line. You can select a band with a rectangular (green) cursor using the mouse. The band can be deleted by using the DELETE key or the ***Bands|Remove*** menu item. Similarly, ***Bands|Remove all*** will remove all defined bands from the gel. Pressing the right mouse button on any region of a pattern causes a detailed view of that area of the pattern to be popped up with a densitometric curve next to the pattern image. The defined bands are marked with a red line. By selecting a band using the left mouse button, the approximated Gaussian shape as calculated by the program is shown. The shapes can be modified by dragging the nodes while pressing the left mouse button. Press the **<*Del*>** button to remove a selected band. At any position, press ther **<*Add*>** button to insert a band. The **<*Up*>** and **<*Down*>** buttons allow you tyo scroll up and down the zoomed image. Press **<*Exit*>** to quit the zoom window and update the band image window.

With the ***Bands|Edit*** option from the menu, a detailed *band densitogram window* of the selected pattern is shown, the area around the selected band being magnified. The defined bands are drawn on the densitometric curve as a summed graph of the constituent Gaussian curves, approximating the contour of the bands. A *position ruler* can be moved along the curve by pressing the left mouse button and moving the mouse cursor. ***Add*** forces a new band to be added at the position of the ruler. A new Gaussian curve is added, and the height and width are calculated by iterative curve fitting. ***View|Zoom out*** restores the full preview of the curve. ***View|Zoom in*** gives a detailed view on a small part of the track around the pointer line, allowing you to tune the position of a band accurately. It is even possible to scroll through the detailed view of the densitogram. By pressing ENTER you can toggle between the detailed and normal modes. The selected Gaussian curve is marked by three red *dragging nodes* (squares) which are designed to adjust the Gaussian shape manually, one at the top and two at the left and right edges. When zoomed, you can click on the top node to drag it to a new position (while holding down the left mouse button) or to change the height of the band. Drag one of the two outer nodes to a new position to adjust the width of the curve. ***Remove*** or pressing DEL deletes the selected band.

This procedure can be repeated to mark other bands on the track. An existing bands can be edited by putting the position ruler on its top position. The three

dragging nodes appear, allowing you to change the Gaussian shape as described, after zooming in.

In addition, ***Search*** automatically searches bands on the whole pattern, respecting the band search filters defined in the band settings window. ***Print|Text report*** creates a report of the positions, heights and surfaces of all bands. ***Print|Graphics image (bands)*** and ***Print|Graphics image (envelope)*** prints the calculated band shapes on the densitometric curve, as separate bands or summed envelope, respectively. Select ***OK*** to validate changes made and to return to the band image window of the gel or ***Cancel*** to exit without performing any changes made.

In the bands image window, the menu item ***Bands|Print*** creates a text-oriented print report of all defined bands for each pattern with the metrics (if defined) and the relative area indicated. Choose ***OK*** to save the bands for the gel to disk and exit the band image window. With ***Cancel***, you exit the window without saving the bands to disk.

## 6.11.2 Quantification of bands

The concentration of any band on a gel can be determined based on a set of calibration bands with known concentration, occurring on the same gel. Because gels can be differently stained, or scanning parameters such as brightness, contrast, resolution etc. can be different, each calibration is restricted to the gel from which it is derived. This means that, for every new gel you want to determine band concentrations on, you will have to add a set of calibration bands.

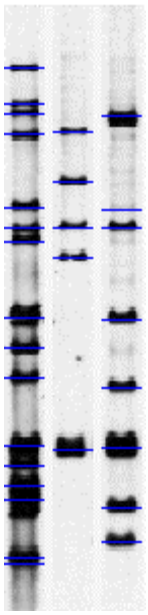
The quantification of bands is based on the delineation of surface contours on two-dimensional images. In other words, the gels must be scanned as TIFF image and the gelstrips must be present in the database (see **Gelstrip** settings in 4.2.4 and 5.2.2). The band quantification cannot be used when patterns are only available as line tracks (densitometric curves) in the database.

### 6.11.2.1 Definition of band contours

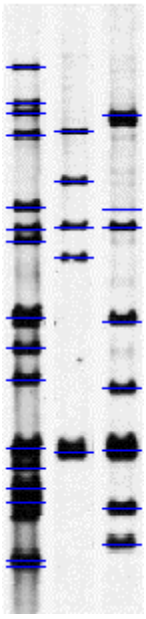
Double-click on a gel in the database to open its gel window. When bands are not yet defined for the gel, first define and save them using the menu option ***Gel|bands***, as explained before (6.11.1). The *Band quantification* window is called using menu command ***Gel|bands quantification***. All bands defined for the patterns are displayed as blue lines on the gelstrips.

***NOTE****: The* band quantification *window is not a band search window. This implies that bands cannot be added or searched for in this window. It is, however, possible to delete a band using* **Bands|Delete band**. *Bands deleted in the Band Quantification (.QNT file) are not automatically*

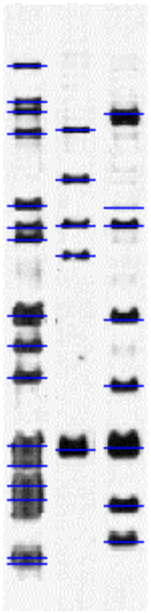
Before finding the surface contours of the bands on the gelstrips, you can eliminate the background on the pattern images using ***Bands|Subtract background (1D)*** and ***Bands|Subtract background (2D)***. The difference between the 1D and 2D algorithms is made clear in the example below. The 1D method uses a densitometric curve (line track) derived from the gelstrip and subtracts background from that curve. The calculated background contour is then subtracted from the gelstrip. Typically, differences in background from the top towards the bottom of the patterns are eliminated. The longitudinal background smear on the first pattern is not eliminated. The 2D method treats the 2D gelstrip directly, subtracting background in both directions (lateral and longitudinal). As a result, the longitudinal stripes are removed from the first pattern.



Before background subtraction



1D background subtraction



2D background subtraction

The 2D background subtraction takes much more time than the 1D method. It is therefore recommended to use the 1D algorithm unless you observe background appearances as shown.

The surface contours of the bands are calculated automatically by ***Bands|Calculate surface***. The search process involves two passes, and when searching is finished, all bands that were defined are contoured by blue lines. A band can be selected by pressing the **left** mouse button on the green spot in the centre of the band. The contour line of the selected band becomes red. At the bottom of the window, the status line displays the absolute position of the band (and its molecular weight, if present) and its absolute volume calculated from the gelstrip.

The surface contour of each individual band can be edited and modified by pressing the **right** mouse button at the green spot. This pops up a detailed view of the selected band, allowing you to redraw the contour lines by dragging the mouse pointer while pressing the **left** mouse button. You can exit the zoom window by clicking the **right** mouse button again. While editing band contours, you may want to ***Save*** the work now and then.

#### 6.11.2.2 Calibrating the gel using known band concentrations

Once the band contours are defined (automatically or manually), Molecular Analyst Fingerprinting software has automatically calculated the absolute volume of each band. As the absolute volume is a measure without physical significance, tools are available to translate these volumes into real concentrations.

First locate the bands with known concentrations and double-click on the first one. An input dialog box appears, prompting you to enter a concentration value (decimals are allowed). Press **ENTER** or click **<*OK*>** to enter the value. The band is now marked as calibration band by a pink marker line. You can correct the value at any time by double-clicking again on the band and entering a new value. A band to which a calibration value was assigned can be removed from the calibration table by double-clicking and deleting the value from the editor line in the input dialog box.

Continue to assign values to known bands until all calibration bands are marked. Then, with ***Calibration*|*Calculate interpolation***, a calibration curve is shown, fitting through all entered values. The regression is based on the *cubic spline* mathematical algorithm. This curve can be printed using the ***Calibration*|*Print*** menu. When you open the calibration window, the regression is automatically applied for all unknown bands on the whole gel. You can perhaps ***Save*** the calibration again after this step.

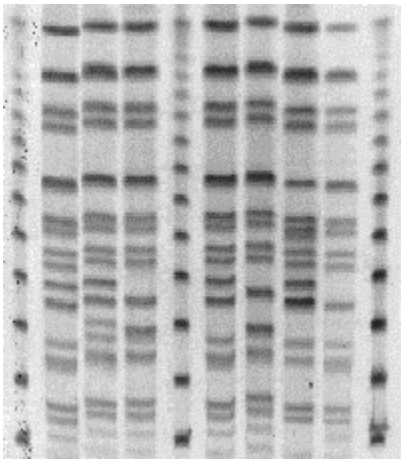
The menu ***Calibration*|*Concentration unit*** allows you to enter a metrical unit for the concentration (e.g. pg, ng or µg; the default unit is ???). Type a unit name of up to 7 characters. ***Save*** the process again after this step. When any band is now selected, its concentration is shown on the status line. The status line also shows KNOWN if the band was used for calibration. If the concentration of a band is higher than the most concentrated calibration band, Molecular Analyst Fingerprinting software is not able to reliably calculate its concentration and will notice this with a *greater than* sign (>) followed by the highest concentration entered.

#### 6.11.2.3 Calibrating the gel for band morphology differences

In most electrophoresis systems, the morphology of bands differs with their molecular weight, i.e. their position on the gel. High molecular weight bands

are usually better stacked, resulting in sharp, thin and dark bands whereas low molecular weight bands have migrated a longer distance and can diffuse easier, causing broad, fuzzy and weak bands. Other factors such as staining also may be dependent on the molecular weight of proteins or nucleotide sequences.

The size marker patterns on the PFGE gel below are an example of how bands may differ gradually with their size and position.



Thus using a single marker band with a given molecular weight to calibrate the whole gel may be subject to considerable error when bands with different molecular sizes are quantified (particularly because the scanner response curve is nonlinear and has a saturation level). Molecular Analyst Fingerprinting software allows a secondary calibration to be performed based on *groups* of bands, each group consisting of bands with the same molecular size but different concentrations. In other words, if we take the example above, we could have taken a 1µl sample for the left, a 2µl sample for the middle, and a 3µl for the right marker. Each set of corresponding bands of the three markers, i.e. with the same molecular weight, constitutes one group, and Molecular Analyst Fingerprinting software will calculate a *horizontal* calibration regression for each of the 13 groups in total, based on the differences in *concentration*. In addition, the program will calculate a secondary *vertical* regression based on the differences in *morphology and/or staining* of the successive band groups.

In practice, enter first all concentrations for a band with a given size, as described above. When this is done, call ***Calibrate|Add group***. You enter now in the *group adding* mode, which allows you to simply click on all bands which you want to unify into one group (bands with the same size!). They become connected with a yellow line. When all corresponding bands are grouped, press ***OK***. Otherwise, if you made a mistake, press ***Cancel*** and restart.

Repeat the same steps for the next group and so on, until all groups are defined. Then choose ***Calibration|Calculate interpolation*** to calculate the regressions. The first group is shown, and the window shows for example "**Curve 1/4 (position at 94 pts)**". Press the ***PgDn*** key to view the successive group curves until 4/4 is reached, then ***Exit*** the calibration window. ***Save*** the calibration as necessary. When any band is now selected, its concentration is shown on the status line. The status line also shows KNOWN if the band was used for calibration. If the concentration of a band is higher than the most concentrated calibration band, Molecular Analyst Fingerprinting software is not able to reliably calculate its concentration and will notice this with a *greater than* sign (>) followed by the highest concentration entered.

The thus obtained quantification of the gel is saved in the database and can be used in the comparative quantification as discussed below (6.11.3). A full text report of the band quantification statistics can be printed using the ***Bands|Print*** command. The report lists for each track the positions of all its bands, their molecular sizes or other metric if defined (see 6.3.4), their absolute band volumes, relative band volumes as percentage of total volumes, and concentrations in physical units. Bands indicated with > are out of range provided by the calibration marker bands and their concentration cannot be reliably extrapolated. Bands marked with (*) are known calibration bands. Note that it is at any time still possible to recall the calibration window of a particular gel and to re-edit the calibration.

### 6.11.3 Comparative quantification and polymorphism analysis

The so-called *Comparative Quantification* of Molecular Analyst Fingerprinting software is a very valuable combination of two types of analysis: comparison and grouping of patterns using conventional clustering algorithms and on-screen evaluation of common and different bands between groups of patterns. Comparative quantification can be executed on any selection of up to 1000 patterns from the database. In a first step, Molecular Analyst Fingerprinting software divides all the bands found among the selected patterns into *classes of common bands.* 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. Clearly, the number of band classes distinguished will depend on the *position tolerance* that is allowed between bands considered as matching. When a larger position tolerance is specified, more bands will be grouped in the same class than when a small position tolerance is chosen.

For each pattern, a particular band class can have two states: present or absent. This is the basis for the *Polymorphism Analysis*, a tool which allows comparative binary (+/-) tables to be generated, displaying all or only polymorphic bands between the selected patterns. These tables, created as text or tab-delineated files, are ready for export to other specialized software for statistics, genetic mapping or other further analysis. Instead of using binary (+/-

) data, the same tables can be generated with absolute band volumes, percentage band areas or real band concentrations.

To make the polymorphism analysis even more performant, different patterns from the same organism, obtained by using different primers, restriction enzymes or whatever, can be combined in one single analysis. The limit is 1000 bands per single entry, on a total of maximally 1000 entries! This means that a total of one million bands can be compared in a single analysis.

#### 6.11.3.1 Band grouping

Before starting the comparative quantification, select a list of patterns you want to analyze (see 6.3.2 and 6.3.3 to select lists). Then select ***Comparison|Polymorphism Analysis*** from the Molecular Analyst Fingerprinting software main menu. A dialog box *Comparative quantification* prompts you to select various options.

The quantification part of the analysis can be based either on the one-dimensional data (***1D***) derived from the peak areas on the densitometric curves (provided in the .PKS files, see 6.11.1) or on the two-dimensional (***2D***) band volumes calculated from the two-dimensional gelstrips (.QNT files, see 6.11.2).

Within the 1D (.PKS) option the quantification can be based on the peak's relative ***Surface (%)*** or on the ***Max height*** of the peaks.

Within the ***2D (.QNT)*** option, the band intensity can be expressed as relative ***Surface (%)*** where the total pattern surface is set to 100%, as absolute ***Surface*** directly calculated from the band contours, or as ***Concentration*** based upon the calibration with known reference bands.

The band matching part of the analysis involves two parameters for searching corresponding and not corresponding bands: the ***Tolerance***, in points, and the ***Optimization***, also in points. Setting a tolerance of 5 points allows bands shifted over an interval of maximally 5 points at either side to be added to the same class. An optimization of 5 points means that each track is shifted over an interval of 5 points at either side of the original position, and the position with the largest number of bands grouped in existing classes will be used for the final output. This optimization is similar to the optimization discussed in the grouping analysis, section 6.6.1.2. Both the tolerance and optimization depend on the track length and the quality of the patterns. Indicatively, a tolerance of 0.8% and an optimization of 0.8%, e.g. 4 points each for a 500 points resolution, will be convenient.

When the **<*OK*>** button is pressed, the *Polymorphism analysis window* appears after some calculations. In this window, the images of all tracks of the current list are displayed. On top of the tracks, a red dotted line marks the position of

the bands. All bands found on all tracks are grouped and divided into classes, dependent on the tolerance interval specified. The central positions of these classes are marked with vertical magenta-coloured lines drawn behind the gel tracks. Every band is connected to its class with a full red line. Due to the allowed tolerance, there might be some distance between a band and its class. Note that every band on all tracks allways belongs to exactly one class.

#### 6.11.3.2 Viewing options

By default, ***View|Gelstrips*** is enabled (marked by ✓). With ***View|Band positions***, the patterns are simplified as bar graphs on which the band positions are represented as rectangles. The rectangles have a colour ranging from green, over yellow and orange to red, which reflects the intensity of the bands. The menu command ***View|Intensity*** is to enable (✓) or disable the intensities. When disabled, all bands are yellow rectangles.

While the bar graphs still respect the relative positions of the bands on the patterns, the option ***View|Band table*** does this not. All band classes are listed as columns of a matrix where the patterns are rows. If a particular band class is present on a particular pattern, the intersection of the corresponding row and column is shown as a green to red square, reflecting the intensity of the corresponding band (when intensity is disabled a green square appears). A band class which is absent on a track is shown as a dark blue square. This visualization of a +/- table is easy to interpret, but does not reflect the distance between the bands neither the shift that was allowed for each band to match with a class. The position of the band classes in points relative to the track length is shown in the header of the table.

In both ***View|Band positions*** and ***View|Band table*** modes, you can zoom in on the patterns by using the ***View|Zoom in (X dir)*** command. This menu item is marked with ✓ when zoomed. Zooming in the X direction only makes sense for tracks with a high resolution and/or with many bands. In addition, the patterns can be shown as normal or narrow graphs using ***View|Small (Y dir)***. In small mode, this menu item is marked by ✓.

A real table of intensities can be shown by the command ***View|Numerical data***. The patterns are rows and the band classes are columns. The position of each band class is shown in the table header. Only when a band is present, its intensity is shown in the table cell. Depending on the choice made in the polymorphism analysis dialog box, the intensities are shown as 1D surface of peaks, 1D heights of peaks, 2D relative surfaces of band contours, 2D absolute surfaces of band contours or 2D concentrations. The type of quantification chosen is shown on the status line.

Bands which are common for all patterns are sometimes not interesting for the analysis. The menu command ***View|Discard trivial bands*** allows all common

(= non-polymorphic) band classes to be discarded. This feature can be enabled (✓) or disabled.

To leave more space for the name field or for the image, the horizontal separator line between the entry labels and the band table or image can be moved by positioning the mouse pointer on the line and dragging it left or right. Similarly, the vertical separator line between the table header and the table or image can be moved up or down.

#### 6.11.3.3 Changing band assignments

The assignment of each band to a class is done by Molecular Analyst Fingerprinting software, using an algorithm that takes account of several factors such as the most prominent bands, the position tolerance, the optimization (if enabled). However, the user may wish to assign some bands differently upon visual inspection of the result. Molecular Analyst Fingerprinting software allows you to override every assignment made, provided that you are in the ***View|Gelstrips*** or ***View|Band positions*** mode.

> ***NOTE****: Before we start explaining how to change the assignments, remember that you can resize the window in order to enlarge the image if necessary, or that you choose* ***View|Zoom in (X dir)*** *when the bar graphs are enabled.*

A band can be selected by clicking the left mouse button on its position. In the gelstrips mode, it becomes marked with a red rectangle, whereas in the bar graph mode it is encircled by a light blue rectangle. When you want to assign the band to a neighbouring band class, click and hold down the left mouse button on the band while dragging it to the neighbouring class. While dragging, a cross (x) indicates which class it will be assigned to. Then release the mouse button. Since each class can only be represented on a track by one band, this will not work if another band on the same pattern is already assigned to that neighbouring class!

When a band is incorrectly assigned to a class, the user may wish to create a new class for that band if no other class exists where it belongs. This can be achieved by selecting the band and using the menu command ***Bands|Split apart***. Bands on other patterns are assigned to the new class if this is the closest. Conversely, if bands are assigned to two different classes whereas they actually are considered to belong to a single class, select a band on one of both classes, and then use the menu command ***Bands|Merge to the left*** or ***Bands|Merge to the right*** if the other class is left or right from the selected class. This will only work if the merged class does not infringe the rule that each class can only be represented on a track by one band!

Note that, in the ***View|Gelstrips*** or ***View|Band positions*** mode, clicking with the left mouse button on a band pops up small flags on all the bands that belong

to the same class as the band which is selected. In addition, these flags show the intensity of the bands in the currently defined way.

#### 6.11.3.4 Comparative Quantification of multiple gel types

It is often necessary that different patterns be obtained for the same organism in order to increase the amount of information. Depending on the technique used, they are obtained by using different primers, restriction enzymes, etc. The comparative quantification tool allows you to combine this extra information in one single merged analysis.

First, start by selecting all the database entries which are the subject of the comparison on one of the gel types and perform ***Comparison|Polymorphism analysis*** on these data. Then proceed by creating a new list of the same database entries, now coming from the second set of gels, generated by e.g. other restriction enzymes. These new tracks can be appended to the structure by choosing the menu option ***File|Add current list*** in the polymorphism window. The program then pops up a dialog box which prompts for the information fields which should be used to link the different entries of both lists to each other. If no fields are checked, both lists are coupled by linking the entries without changing the ordering. If one or more of the information fields are checked, then the program will link all entries from both lists for which these fields are the same and will change the ordering of the newly appended list as necessary, to match with the entries present in the created polymorphism analysis. When a track matches multiply or remains unmatched, the program will report this error and the new list won't be appended.

When the new list is joined succesfully, the same process can be repeated to add subsequent gel types. In this way, up to 10 different types of gel information can be merged in one analysis.

In the ***View|Gelstrips*** or ***View|Band positions*** mode, the tracks from only one gel type are shown at a time. At the bottom of the ***View|*** submenu, a list of items show the list names of all the lists which are currently joined in the analysis. You can use these items to show the tracks of an other gel type. On the other hand, in the ***View|Band table*** or ***View|Numerical data*** mode, the information of all the bands coming from the different gel types is merged into one comprehensive table.

#### 6.11.3.5 Rearranging the ordering of the tracks

By default, the tracks are displayed in the polymorphism window in the original ordering of the selection list. It might often be interesting though to rearrange the entries in order to show closely related tracks next to each other. You can select a subset of tracks by clicking the right mouse button on the track images (or data rows when the numerical table mode is active). The

selected tracks show a label displaying a sequential entry index, which reflect the new ordering as defined by the selection process. ***Selection|Rearrange*** then rearranges the tracks into the new ordering. ***Selection|Clear*** removes the current selection.

Using ***Selection|Discriminative*** bands, the band classes in the *band table* or *numerical table* are rearranged in such a way that bands being typical of the selected fingerprints are displayed at the left hand side of the table. This means that the most left bands are occurring on as many of the selected patterns as possible, whereas they occur in as few of the other patterns as possible. Conversely, the most right bands occur on as many of the non-selected patterns as possible whereas they occur on as few of the selected patterns as possible. Discriminative bands for the selected patterns are those bands that occur either on all of the selected patterns and none of the others, or only on the non-selected patterns and none of the selected ones.

#### 6.11.3.6 Polymorphism analysis combined with Cluster Analysis of tracks

If the comparative quantification and the cluster analysis are available in the same Molecular Analyst Fingerprinting software package, both modules can be combined in a very powerful way. The program allows a hierarchical cluster analysis (using UPGMA) to be executed based on the band information provided by the polymorphism analysis. Selecting ***Clustering|Cluster tracks*** from the polymorphism window pops up a dialog box which prompts for the type of correlation coefficient, offering the choise between the Dice and Jaccard's coefficients, and a correlation coefficient which takes also into account the relative differences band intensities (depending on what kind of intensity is currently applied). Pressing ***<OK>*** starts the clustering. When the calculations are finished, the tracks are automatically rearranged following the clustering, and the corresponding dendrogram appears at the left of the image (see section 6.6 and particularly 6.6.1.2 for more details about clustering and band matching coefficients).

Selecting ***Clustering|Cluster tracks*** a second time removes the cluster analysis from the image (ordering of the tracks remains unchanged). Note that some features (such as manual rearrangement of the tracks) do not work when a clustering is present. The reason for this is that entries cannot be moved independently in a dendrogram.

#### 6.11.3.7 Polymorphism analysis combined with Cluster Analysis of band classes

The present/absent table which arises from a polymorphism analysis can be viewed in two ways. In the "classical approach", one is interested in tracks and wants to determine which bands are present in a particular track. This approach results in a so-called *R* matrix and corresponding dendrogram. In the "dual

approach", one is mainly interested in bands and wants to see which tracks possess a particular band and which lack it.

In the second approach, one can define a correlation coefficient that measures resemblance between band classes. This correlation is high if both bands are in general present and absent on the same tracks. The correlation is low if there are a lot of tracks which have one of the bands and lack the other. This approach results in a *Q* matrix, in which the characters (bands) are compared for the number of entries they share. Although such an analysis is, up to now, not very common, it often reveals useful information about the link between bands among groups of individuals and the underlying mutations in the organisms. In addition, it offers information about the discriminative power of groups of bands.

Molecular Analyst Fingerprinting software allows correlations between band classes to be calculated and the results to be presented in a hierarchical clustering based on UPGMA. Select ***Clustering|Cluster bands*** from the polymorphism window to show a dialog box which prompts for the type of correlation coefficient. Again, there is a choise between the Dice and Jaccard's coefficients, and a correlation coefficient which takes also account of the relative differences band intensities. Press **<*OK*>** to perform the clustering. The band classes are now rearranged into a dendrogram, and hence, the results of this calculation are only visible in the ***View|Band table*** mode or in the ***View|Numerical data*** mode. The program automatically switches to one of these display modes when the calculations are finished. All the bands (which may come from different gel types) are rearranged upon the results of the clustering, and the highly correlated bands are brought in each others neighbourhood. On top of the image, the dendrogram of the clustering is shown. Selecting ***Clustering|Cluster bands*** a second time removes the clustering of the bands. Note that certain features (such as manual reassignment of bands to an other class) do not work when a band clustering is present. It is obvious that the band clustering has to be recalculated when bands are brought into different classes.

A simultaneous clustering of both the tracks and the band classes is an extremely powerful tool to uncover hidden relations in the data. In addition, this combination is convenient to determine sets of bands that might be responsible for the discrimination of the distinct groups in the tracks.

#### 6.11.3.8 Detailed comparison of two tracks

When two tracks are selected in the polymorphism window, selecting ***Selection|Detailed report*** creates a comparison window. This window shows all the bands occurring on both tracks, divided in three groups: (1) bands which are common for both tracks, (2) bands which only occur on the first one and (3) bands which only occur on the second one. For all bands, the position and

### 6.11.3.9 Exporting results from the polymorphism analysis

Select ***File|Print (as graphics)*** to create a graphical print-out of the results. The appearance of the print is in accordance with the current viewing mode.

A text file containing the matching table can be created using the ***Bands|export band table*** option. This pops up a dialog box prompting for the presentation (eighter binary +/- data or the exact band intensities) and allows you to choose whether the file should contain spaces or tabs. The result is stored in a file "TABLE.TXT".

# 7. Database Sharing Tools, a platform for the exchange of fingerprint information

## 7.1 Purposes

Today, the exchange of information among different laboratories is of the utmost importance for the advancement of epidemiological research. However, there is currently no framework for the exchange of electrophoresis fingerprints, partially due to the lack of a general and common file format, and partially due to differences in the standardisation of electrophoresis and processing techniques. The ***Database Sharing Tools***e of Molecular Analyst Fingerprinting software defines an open standard for the exchange of fingerprint information among different research sites. This standard addresses both the issues of a common and flexible file format and the comparison of differently standardised gels.

An exchange file (called ***bundle***) contains a set of fingerprints, ranging from one single entry to several thousands of entries. Each fingerprint is represented by a set of information fields and by the positions of the bands on the track. However, much more advanced information can be added, such as two-dimensional images called *gelstrips*, band concentrations, and an unlimited number of database information fields. These *bundles* have an open, tagged file format, which can easily be extended in the future.

Because different research laboratories often use different electrophoresis parameters and/or reference markers, molecular weights of bands are currently the only kind of information which can be shared with other institutes. However, the Molecular Analyst Fingerprinting software software also allows one to deal with more advanced data, such as normalized *gelstrips*. These gelstrips have become an invaluable tool for the visual interpretation and control of numerical comparisons, identifications, etc. (Good Laboratory Practice) and therefore have a particular value for the exchange of fingerprints. The Database Sharing Tools therefore include a *remapping* technique which makes it possible to convert the full fingerprint information from one reference system to another, including *densitometric* and *2D-image* information. This remapping is based either on the *molecular weight calibration curves* in both systems, or on a dedicated *remapping gel*, which combines both reference patterns in one run.

The *bundle* file format, combined with the *remapping* technology, can serve as a basis for the development of an international network for the exchange of fingerprint databases. The practical implementation of such a network remains

open, but a *client-server approach* based on the Internet is certainly a good option.

## 7.2 The bundle file format

In order to provide a convenient way to exchange fingerprint information and to construct unified databases, the concept of fingerprint **bundles** is introduced here. A bundle is one single file, containing a collection of fingerprints, and has the following features:

- The number of fingerprint entries in a bundle is completely free, and can vary from one single entry to many thousands of them.
- All information (fingerprint images, densitometric curves, band positions and intensities, database information) is collected in one single file, and compressed.
- The entries of a bundle can come from different gels and even from different institutes.
- A bundle always contains information that allows the user to track back the origin and history of its entries.
- A bundle contains maximum flexibility with respect to the type and amount of information for each fingerprint.
- Bundles have an open file format, easily allowing for future extensions and refinements.
- Built-in remapping information enables easy, transparent conversion between different reference systems (see section “conversion between standards”).
- Bundle files are NOT ENCRYPTED; highly confidential or secret information should not be transferred over the Internet without prior encryption.

A bundle contains a header, followed by the list of fingerprint entries. The header contains the following information (fields marked with (*) can be absent)

- Name of standard reference pattern
- Calibration curve of molecular sizes (*)
- Names of all database information fields

- Remapping history
- Comments (*)

Each individual fingerprint entry has the following information:

- Name of the original gel and lane number on that gel
- Name of research site where the fingerprint was created
- Database information fields (no limit on number) (*)
- Band positions (molecular size) (*)
- Band positions (run length) and intensities (*)
- Band concentrations (*)
- Densitometric curve (*)
- Normalized gelstrip image (*)

The Molecular Analyst Fingerprinting software software contains two databases: the **Local Database**, containing ordinary gel files, and the **Shared Database**, which consists of bundle files. Both databases can be mixed for analysis in a completely transparent way, and every feature of the software can be applied on ordinary gel files as well as on bundle entries.

## 7.3 Practical implementation: a Client-Server set-up.

Obviously, bundles are the preferred way for database exchange. Although the implementation of this exchange is completely free, a client-server setup is certainly an interesting option. Such an approach is based on three components.

1. The **shared database**, containing all fingerprints which are shared amongst the different institutes.
2. The **server site**, where the shared database is maintained
3. One or more **client sites**, which are involved in the exchange project.

The procedure of maintaining, updating and distributing the database happens in the following way:

When one of the clients has one or more new fingerprints which it wants to add to the shared database, it prepares a bundle containing the fingerprints and

database information as desired, and submits it to the server institute (e.g., by FTP, or even as E-mail attachment).

The server institute collects all incoming new bundles, checks for duplicates and performs a acceptability control (standardization, quality, reproducibility). It adds all interesting new fingerprints to the shared database.

On a short time scale (e.g. once in a week), an incremental update of the shared database is distributed (e.g., using WWW or FTP server), containing only new fingerprints which are not present in the last major upgrade.

Periodically (e.g., once in a year), the server distributes an updated version of the complete shared database to all clients (e.g., using CD-ROM or DAT tapes).

Of course, clients can also directly exchange fingerprint bundles over the net, but these ones are not “official members” of the shared database.

In this approach, all clients have a local copy of the most recent shared database. This approach offers some important advantages over the approach where clients send queries to the server (e.g., for identification of new tracks):

1. It offers maximal flexibility with respect to the application of this database: clients are completely free in the way they perform their analyses, and present their results.

2. It avoids long delay times due to net activity or server computer shutdowns

3. It avoids heavy stress on the computers of the server.

4. The shared database can be used “off-line” from the net, using the local copy.

It is obvious that other implementations, for example where client sites have no copy of the complete database, can be designed if necessary in certain instances.

## 7.4 Conversion between standards

The Molecular Analyst Fingerprinting software software uses the concept of “standards settings” of a database, which holds the information about the reference pattern used to normalize (to make compatible) all the gels, and the resolution (number of densitometric values per track) of the normalized fingerprints. In the past, one could only compare gel files if they had the same standard settings. However, this is an inconvenient situation if one wants to set up a cooperation between different institutes, since they will often have

bundle files and sharing databases is combined with the introduction of a option to convert standards into one another, called *remapping*, as part of the Database Sharing Tools. The remapping function allows the user to easily convert bundle files from one standard system to an other. The only information required for this is the **calibration curves of the molecular weights in both systems**. A very powerful feature of the remapping function is that all possible information can be converted from one system int another, including 2D-images (gelstrips), densitometric curves, band positions (run lengths), etc. Even research institutions using different electrophoresis settings, for example different pulse and ramp settings and/or size markes in pulsed field gel electrophoresis, can convert one anothers complete databases with no loss of information, provided that the molecular weight curves for both systems are available. The reliability of the remaping will depend on the accuracy and reproducibility of the size markers used in both systems.

Using the same concept, it is even possible to import fingerprint information exported from other software systems, where only molecular weight information is known, into the Molecular Analyst Fingerprinting software software. In this way, the bundle concept offers an attractive way to make the shared databases compatible with other fingerprint software systems. Of course, more advanced information (normalized gelstrips, band matching tolerance, densitometric curves etc.) will only be available when exchanges between Molecular Analyst Fingerprinting software software systems are made.

## 7.5 Using the Database Sharing Tools.

### 7.5.1 Creation of a new bundle.

First, create a selection list in the Analyze program, containing the fingerprints which are to be included in a bundle. Then select from the menu ***Bundles/New bunde*** to open the *Bundle creation dialog box*. This dialog box allows you to specify:

- The filename of the new bundle
- The name of the institute where the fingerprints were created
- Database information fields which should be included in the bundle
- Which information should be included: densitometric curves, bands or normalized Gelstrips
- Whether the gelstrips should be compressed or not

· A comment line about this bundle

When everything is filled in properly, press **<*OK*>** to create the bundle. The bundle appears as a ".BDL" file under the MA-F database directory (where the ".INT" gel files are stored). A bundle which is imported from an other system (e.g. copied from disk) should always be copied to this directory.

***IMPORTANT NOTE.*** *It is also possible to select bundle entries into a selection list in MA-F (see below). In this way, one can incorporate existing bundle entries into a new bundle. This option can be used to merge several small bundles into one bigger file, or to add new fingerprints to an existing bundle. In case two identical fingerprints are present (e.g. coming from two different bundles), the program checks for this and warns the user, asking which one to use.*

## 7.5.2 Opening a bundle file.

When the MA-F analysis software is opened, no bundle is loaded and the Shared Dabatase window is empty. To open one or more bundles, select the menu option ***Bundles/Open bundle***, which creates a dialog box that lists the existing bundles in the current database directory. The user can select different bundle files, and use the **<*Open*>** and **<*Close*>** buttons to load bundles into the shared database, or to remove them. Alternatively, one can use the ***Bundles/Open all bundles*** menu option to load all the existing bundles at once.

Double-clicking with the left mouse button on one of the entries in the shared database opens a new window showing additional information about that bundle entry.

## 7.5.3 The construction of lists of fingerprints containing bundles.

The Molecular Analyst Fingerprinting software handles bundle fingerprints almost in exactly the same way as fingerprints coming from ordinary gel files. The user can manually select bundle entries into the selection list by pressing the right mouse button on the corresponding entry in the shared database. In addition, one can also use the automatic search function, called with ***Search/Topic***. The automatic search dialog box has two additional checkboxes, "Local database" and "shared database", which allow the user to specify whether only the local database is scanned, or only the shared database, or both are searched simultaneously.

In addition, additional search functions exist, which are specific for the shared database:

· The menu option ***Bundles/Search bundle*** is used to select all fingerprints which come from one single bundle.

- The menu option ***Bundles/Search institute*** allows one to select all fingerprints which come from a specific institute or research centre.

Lists containing bundle entries are saved and loaded in the same way as ordinary lists.

### 7.5.4 Analysis of shared database entries.

The most powerful aspect of the shared database module is that almost every analysis (cluster analysis, band matching, identification,...) can be done on a mixed list, containing ordinary fingerprints as well as bundle entries.

In order to indicate that a particular fingerprint originates from a bundle rather than a local gel file, its name is written in blue on the screen (e.g. when showing a dendrogram).

## 7.6 Using the standard conversion utility.

MA-Fingerprint offers the opportunity to automatically convert fingerprints of the shared database from one *normalization standard* to an other. A normalization standard consists of two components:

1. The resolution of the normalized fingerprints. This part never causes any problem because the Molecular Analyst Fingerprinting software automatically rescales bundle entries to a new resolution if necessary.

2. The name of the standard reference fingerprint which was used to correct the references (i.e. a gel name and a gel index). In order to be able to convert between different standard references, the software needs to know the metrics calibration curve (e.g. fragment lengts of molecular weights) of both both reference systems. This information is stored in a *remapping file*.

Remapping always causes a slight decrease in quality of the fingerprints, because standardization using Metrics calibration curves is intrinically inferiour to standardisation based on run lengths. Therefore, the software warns the user that a particular fingerprint has been converted by showing its name in red.

It is important to realize that the Molecular Analyst Fingerprinting software never changes the information if the bundle entries in the files on the hard disk, but performs the conversion during an analysis (e.g. a cluster analysis) *on the flow* in the computer's memory. In this way, bundles are allways preserved in their original state.

### 7.6.1 Creation of a new remapping file

Select the menu item ***Bundles/Create standard conversion*** file pops up the *Standard conversion dialog box*. On the left, the existing standard conversions are shown. On the right, two lists show the existing normalization standards for which the Metrics calibration curve is known. Select the source standard in the left column, and the destination standard in the right column. An edit field allows the user to enter an additional comment, which will be saved in the remapping file. Pressing **<*OK*>** creates a window where both calibration curves are shown, as well as the remapping curves. Press **<*Save*>** to store the remapping file on disk.

# 8. User Setup

The Startup program allows various parameters to be customized, such as screen colours, field labels, data directories, etc. To make Molecular Analyst Fingerprinting software even more flexible, it is possible to define multiple users, each defining their own preferential settings and separate directories.

## 8.1 Creating and naming users

When the <***User***> button in the Molecular Analyst Fingerprinting software Startup screen is pressed, a floating menu lists a number of commands with regard to users. Choose ***Add new user*** to add a new user to the list of the currently established users. An input box appears, prompting for the name of the new user. Type a name of maximally 12 characters and press <***OK***>. The program now asks "*Automatically create new user directories?*". If you choose <***Yes***>, a subdirectory will be created automatically in the Molecular Analyst Fingerprinting software directory with the name of the user (abbreviated to 8 characters), and all directories necessary for databases, lists etc. for that user (see section 2) will be created as subdirectories of that user directory. If <***No***> is chosen, you will have to create (sub)directories manually and specify them for the new user in the User Setup menu (section 8.2). The name of the new user appears in the list. (note that each user should have a unique name). A user can be removed from the list by selecting the user name and calling ***Remove*** from the User button's menu. The program asks to confirm that the user will be deleted. If confirmed, it asks "*Automatically remove empty user directories?*" By asking <***Yes***>, empty directories specified for that user will be removed. ***Rename*** allows a selected user to be renamed. The *User Setup window* for the selected user is loaded by calling ***Setup "Username"*** from the User button's menu.

## 8.2 The User Setup menu

### 8.2.1 Databases and directories

To avoid interference between the application directories and databases of the different users, the directories can be specified for each user independently. The directories which are entered are not created by the program; they should first be created in the File Manager or from a DOS prompt. The following directories can be defined:

(1) Database (default: \MA-F\GELS.INT). This directory specifies the default directory for the storage of normalized gels (".INT" files). However, other

directories from the same drive can be loaded at run-time. The directory will be displayed when the command ***Database|Go to database*** is selected from the Molecular Analyst Fingerprinting software main window.

(2) Gel images, where you can specify the directory used by the scanning software to write gelscans.

(3) Raw gels (default: \MA-F\GELS.RAW). Unnormalized gel files are stored and loaded in this directory. Other directories can be specified at run-time.

(4) Lists (default: \MA-F\LISTS). All selection lists and accompanying grouping analyses are stored in this directory.

(5) Libraries (default: \MA-F\LIBS). Identification libraries are stored as subdirectories of this directory.

(6) Combined gels (default: \MA-F\GELS.CMB). Combined gels will be stored in this directory. The directory will be displayed when the command ***Database|Go to combined gels*** is selected from the Molecular Analyst Fingerprinting software main window.

To change the name of one of these directories, be sure that the new path exists on the hard disk; otherwise, the program produces an error message and selects the Molecular Analyst Fingerprinting software home directory.

### 8.2.2 Window colours

This option allows you to choose between a gray window background using standard windows system colours and a user-defined background from an RGB palette of $10^6$ colours, which is independent of the Windows system colours but requires 256 colour graphics. A sample background and selection bar is shown. Choose the option "***Use system colours***" when the custom mode conflict with other software that is run simultaneously.

### 8.2.3 Gel staining

The Gel colours allows you to adjust the colour palette of the gel display window in order to reproduce the real colours of the gels. The "*Background*" colour represents the background colour of the gel matrix, while the "*Stain*" colour represents the colour of the used stain. Both colours have three scroll bars, allowing you to adjust the red, green and blue components independently. A 256 colour graphics adaptor is necessary to achieve acceptable representations.

### 8.2.4 Information field labels

Gel tracks are characterized by 7 information fields, which are used to label a gel track and for the automatic search option. The names of these fields can be changed for the current user. Each field can have a name of maximum of 12 characters The fields 1, 2 and 3 are supposed to contain the name(s) of the track (each track name can be 15 characters long). Fields 4 and 5 are designed for a code name and an additional code of the track, respectively (each track code can be 12 characters long). Fields 6 and 7 are comment lines, with a restriction of 40 characters for each comment.

When all settings are customized, press **<*OK*>** to return to the Startup screen and update the user settings on disk. Using **<*Cancel*>**, the changes are not saved.

### 8.2.5 "Doublegel" option

Checking this option will enable the Doublegel conversion and normalization (see 4.5) for the specified user.

## 8.3 Customizing entry description in Molecular Analyst Fingerprinting software

After customizing the information field labels in the User Setup, each user can design the combination of information fields to be displayed in images such as gels, dendrograms etc.

In the Molecular Analyst Fingerprinting software Main program, the ***System*|*Names*** menu item or pressing the F9 function key allows you to display the *Entry description dialog box*. Six predefined schemes are available of which one can be selected. With **Gel name and index**, the name of the gel and lane number will be shown for each entry. **Abbreviated name** will show the three name fields abbreviated to four characters whereas **Full name** will not abbreviate the name fields. The next two options are for the two entry code fields (the name that appears depends on the field names you have defined). The last option will display the full name fields plus the first entry code field.

Each of these schemes can be used as basis for customizing the display, using the scroll bars at the left side of the dialog box. In order to have an overview of how the customized layout will look, it is recommended to display the requested image (pattern selections, dendrograms) before calling the Entry description dialog box. Make the screen image large enough to have the full entry descriptions displayed e.g. by maximizing the window. Then call ***System*|*Names*** or press F9 to customize the descriptions. The first item which can be displayed is the *gel name and lane number.* If you want to show these,

click on the appropriate check boxes. Allow the necessary space between the gel index fields and the next fields to be displayed, by increasing the ***Space*** scroll bar. Try perhaps 80. Then define a cut-off length for each of the *name fields*. If you want to hide a given name field, decrease the cut-off length to zero, or if you want to display it untruncated, move the scroll bar to the right. Press the <***Preview***> button to update the current screen image(s) in the background and, if necessary, change the ***Space*** allowed. Now, if you want to add *entry code fields*, determine the necessary space between the name fields and the first entry code field (the name displayed for that field is the one you choosed in the user setup) and define the maximal length of the entry code field(s) and the space between them. An additional field is available to mark each pattern with an imported string, called a *free string*. How you can tell Molecular Analyst Fingerprinting software which string you want to assign to each entry of the list, is described in 6.3.3.2. After each change you can press the <***Preview***> to update and check the image(s).

You can save a customized scheme using the <***Save***> button. Enter a name of maximally 25 characters; any characters allowed. Delete an existing scheme with the <***Delete***> button. Predefined schemes (marked with >) cannot be deleted.

# 9. The Molecular Analyst Fingerprinting software Printer Manager

## 9.1 Principles

Print jobs from Molecular Analyst Fingerprinting software are not immediately sent to the printer, but are written to disk as print job files. The advantage of this system is that printing in the programs demands almost no time, so that one can work further and executing the printing of all print jobs at another time, without having to interrupt the session. Printing is done by the Molecular Analyst Fingerprinting software Printer Manager. Each time a new print job is launched, the job list is automatically updated and brought to front.

The Molecular Analyst Fingerprinting software Printer Manager uses the print routines provided by Windows 3.1. This means that the installation and setup of a particular printer should happen using the "**Printers**|**Add**>>" item from the Windows Control Panel.

## 9.2 Previewing and printing

The *Molecular Analyst Fingerprinting software Printer Manager main window* displays a list of all available print jobs and is brought in front each time a job is launched or by the ***System***|***Molecular Analyst Fingerprinting software Printer Manager*** command in the Molecular Analyst Fingerprinting software main menu, or the shortcut F8 (function key). One job can be selected by moving the bar using the mouse or ⇧⇩ keys. To print a job, select the job name and choose ***Print***|***Print selected job***. With ***Print***|***Print all jobs***, the printer driver will print out all print jobs of the current list, one by one. The Molecular Analyst Fingerprinting software header on each printed page can be enabled or disabled by checking the menu item ***Print***|***Print with header*** (marked with ✓ when enabled). ***Print***|***Print using colours*** makes it possible to print various images and reports in colours on a colour printer supported by Windows (marked with ✓ when enabled). ***Print***|***Printer setup*** calls the setup dialog box of the default printer, allowing you to change the various printer settings. ***Preview*** or double clicking on the print job name shows a scaled image of the printed papers of a selected job. In case the job will be printed on more than one page, press ***PgUp*** or ***PgDown*** to view all pages successively. To cancel a job that has been printed and/or is not needed anymore, select it and use ***Delete***|.

## 9.3 Exporting print jobs to other applications

***Export|To Clipboard*** allows you to copy the selected job to the clipboard. The figure or text can then be loaded in most other Windows applications by simply using an ***Edit|Paste*** command. For text-oriented files (such as list information, identifiction reports, etc.), the format is the normal OEM-text. For graphical images (such as densitometric curves or dendrograms) the Windows Metafile format is used. For text-oriented print jobs, ***Export|Text file*** creates an ASCII text file in the Molecular Analyst Fingerprinting software home directory, named "PRINT.TXT".

# 10. Molecular Analyst Fingerprinting software import of band size tables

## 10.1 Installation

The program will use the settings of the *current active user* and copy the created gels to the database directory specified for that user. Since the generated gels will be of a new type, it is recommended to first create a new user in the Molecular Analyst Fingerprinting software User Setup menu, and have Molecular Analyst Fingerprinting software create appropriate directories for that user.

## 10.2 File format

### 10.2.1 Genescan™ band size tables

This program allows the import in Molecular Analyst Fingerprinting software of band tables, containing the molecular sizes, heights, and areas or volumes of the bands. Tab-delineated ASCII-text files as generated by the Genescan software (ABI sequencers, Applied Biosystems Division, Perkin Elmer Corporation, Foster City, CA) are directly imported when they are converted to MS-DOS format and have the following layout (the header should not be included):

| Lane & band no. | Rf | Size | Height | | Area |
|---|---|---|---|---|---|
| 1B,1 | 24.85 | 60.61 | 195 | 933 | 932 |
| 1B,2 | 26.24 | 70.95 | 269 | 1401 | 984 |
| 1B,3 | 26.88 | 75.41 | 98 | 767 | 1008 |
| 2B,1 | 26.19 | 71.12 | 2670 | 17545 | 982 |
| 2B,3 | 26.59 | 73.91 | 360 | 2740 | 997 |
| 2B,4 | 26.83 | 75.62 | 974 | 6467 | 1006 |
| 2B,5 | 29.41 | 95.01 | 104 | 744 | 1103 |
| 2B,6 | 30.83 | 104.67 | 96 | 739 | 1156 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2B,7 | 36.93 | 143.31 | 128 | 993 | 1385 | |
| 2B,8 | 41.47 | 170.85 | 317 | 2799 | 1555 | |
| 2B,10 | | 47.47 | 204.57 | 1594 | 15557 | 1780 |
| 2B,11 | | 47.97 | 207.36 | 263 | 3479 | 1799 |
| 2B,12 | | 48.48 | 210.15 | 522 | 7452 | 1818 |
| 3B,1 | 23.07 | 70.01 | 57 | 479 | 865 | |
| 3B,2 | 26.11 | 86.59 | 91 | 562 | 979 | |
| 3B,3 | 47.23 | 204.74 | 207 | 1836 | 1771 | |
| 3B,4 | 48.24 | 210.35 | 68 | 837 | 1809 | |
| 3B,5 | 49.87 | 219.34 | 57 | 585 | 1870 | |
| 3B,6 | 89.01 | 432.80 | 198 | 3467 | 3338 | |
| 3B,7 | 101.20 | | 112 | 2187 | 3795 | |

...

It is important that the numbers are separated from each other by ONE TAB; the entire file should contain NO SPACES. The lines should be separated by one hard return.

### 10.2.2 Tab or space delineated band size tables from other sources

The program further allows the import of some simple text tables, delineated by tabs or spaces. If you want to import band size profiles from other software or from spreadsheets, we recommend to use the following format:

>**SAMPLE_NAME1**<RETURN>

**SIZE1**<TAB>**HEIGHT**<TAB>**AREA**<RETURN>

**SIZE2**<TAB>**HEIGHT**<TAB>**AREA**<RETURN>

**SIZE3**<TAB>**HEIGHT**<TAB>**AREA**<RETURN>

<RETURN>

>**SAMPLE_NAME2**<RETURN>

**SIZE1**<TAB>**HEIGHT**<TAB>**AREA**<RETURN>

**SIZE2**<TAB>**HEIGHT**<TAB>**AREA**<RETURN>

**SIZE3**<TAB>**HEIGHT**<TAB>**AREA**<RETURN>

...

Example:

```
>Lane 1B
60.61    195  933
70.95    269  1401
75.41    98   767
>Lane 2B
60.73    1703 10208
71.12    2670 17545
73.91    360  2740
75.62    974  6467
95.01    104  744
104.67   96   739
143.31   128  993
170.85   317  2799
204.57   1594 15557
207.36   263  3479
210.15   522  7452
>Lane 3B
70.01    57   479
86.59    91   562
204.74   207  1836
210.35   68   837
219.34   57   585
```

```
432.80   198  3467
```

Separation between the lanes may contain multiple returns (blank lines).

The molecular sizes may be preceded by a TAB.

It is not necessary that the heights or the areas are present; if not, the program will assign fixed values to them.

The “>” character that characterizes the lane names, should be the first character in that line.

If the lane name is absent, the lanes should be separated from the previous one by at least one return (blank line). In that case, the program automatically assigns the lane numbers as lane names.

Molecular sizes, heights and areas may be separated by one or more spaces instead of tabs.

Thus the program would also accept the following type of input:

```
60.61

70.95

75.41


60.73

71.12

73.91

75.62

95.01

104.67

143.31

170.85

204.57

207.36

210.15
```

```
70.01

86.59

204.74

210.35

219.34

432.80
```

It is obvious that in this example, only the band positions are present, and consequently, that only the Dice, Jaccard and Jeffrey's X coefficients will provide meaningful comparisons.

## 10.3 Features

### 10.3.1 Creation of files

Program MWTOGEL saves the gels as IMP_XX files in the Database directory of the active user. XX represents a sequential number. The first gel will be saved as IMP_1, the second as IMP_2 and so on. The program automatically counts up by looking at existing files. The maximum number of lanes a gel in Molecular Analyst Fingerprinting software can contain is 50. If the file contains more than 50 lane records, the program will automatically save the lanes in two or more gels, respecting the sequential numbers as found in the first column of the input file.

### 10.3.2 Molecular weight regression

To revert the size tables into gels, the program uses a reverse MW-regression. A slightly exponential *pole* function of the type (A+x) (B+y) = C is used. The function applied is automatically saved in the Molecular Analyst Fingerprinting software METRICS subdirectory, so that it is possible to assign the exact molecular weights to the bands of the reconstructed gels. The error due to reverse regression is less than 0.2%, which is far below the error of the most accurate MW estimations possible.

A linear regression is also available.

### 10.3.3 Use of band size parameters

The program makes use of both the height of the peaks (bands) and their area (or volume) to recalculate a *Gaussian* profile for each pattern. Each band is presented as a Gaussian curve with a specific height and width, so that both the height and the area of the Gaussian curve are proportional to the height and area, respectively, of that band as given in the input file. In the case of very close bands, the profile will be the sum of the overlapping Gaussian shapes, as is the case in real gel scanning profiles.

### 10.3.4 Applicability

Program MWTOGEL not only creates the gel files (.INT files), but also the associated peak description files (.PKS files). The peak files also contain the peak heights and areas. By having both the densitometric curves and peak profiles, it is possible, using Molecular Analyst Fingerprinting software, to accomplish the following types of analysis:

- Calculate dendrograms and identify based on *Pearson* product-moment correlation
- Calculate dendrograms and identify based on *Dice*, *Jaccard*, and *Jeffrey's* X binary coefficients
- Calculate dendrograms and identify based on area matching coefficients and fuzzy logic
- Show reconstructed gel images as independent figures and next to dendrograms.
- Show Metric scales next to or on top of gel images and dendrogram images

The only feature which is in fact not available, is the possibility to show 2D-gelstrips and perform 2D-quantifications.

## 10.4 Using the program

### 10.4.1 Loading files and creating gels

Before running MWTOGEL, run Molecular Analyst Fingerprinting software to verify if the right user is set active. If you run MWTOGEL for the first time, it is recommended to define a new user. In MWTOGEL, select ***File|Open*** in the menu to locate and open an input text file as specified in this chapter. If the file format is incorrect, the program will not warn, but will show up only one

(sometimes two) lanes found, rather than the expected number. After a file is loaded and read correctly, the program shows:

**File loaded: FILENAME (XX lanes)**

FILENAME is the name of the input file, XX is the number of lanes found.

Then, select ***File|Create*** to save the gel file(s) in the Database directory of the current active user. When this has happened, the program reports:

**Created: IMP_1**

**Database: C:\MA-F\USER1\GELS.INT of user "USER1"**

For next gels the filename will be IMP_2, IMP_3 and so on. USER1 is an example here, as well as the full path listed. You can now import more files, or quit MWTOGEL by ***File|Exit*** and analyze the gels in Molecular Analyst Fingerprinting software.

### 10.4.2 Settings

With ***Edit|Settings***, you can modify settings like the ***Regression method*** applied and the MW range to be used. Choose between a weakly exponential *pole* function (recommended in most cases) and a linear regression, which may be interesting to focus on high molecular weight bands.

With the ***Resolution***, you can specify the number of points of the densitometric tracks that will be reconstructed from the band size table. Typical setting is 1000 points. If the files are characterized by complex patterns with well-defined sharp bands, the bands may become too overlapping, so that it may become useful to increase the resolution.

The ***Range in base pairs*** defines the maximal band size that will fall within the reconstructed pattern. It is useful to check the text files for the largest molecular weights that occur, before you start the conversion.

The settings are saved in a file MWTOGEL.GCD, in the Molecular Analyst Fingerprinting software home directory.

# 11. Configuration, Diagnosis & Information

## 11.1 Objectives

Most of the problems that cause errors or problems when running Molecular Analyst Fingerprinting software can be detected and explained by the GCINFO program. This program determines the hardware configuration and can display a complete report of characteristics, including errors and incompatible features. The program discriminates between critical errors and non-critical errors or warnings. In addition, GCINFO can generate a comprehensive report of all Molecular Analyst Fingerprinting software settings of the selected user, including all problems found. This file, of which hardcopies can be made as printouts, is a very valuable backup of the current normalization and database settings, and enables the user to reconstruct his database when files and/or directories have been damaged by any reason.

## 11.2 Directories

From the Molecular Analyst Fingerprinting software Startup screen, the <***Diagnose***> button loads the GCINFO program with the selected user. The *Main Diagnostics screen* is shown, on which the user name is displayed and the directories are listed, together with the available space for each directory. A green lamp next to a directory means that the directory is valid and no problems will occur. A yellow lamp means that the selected directory may cause problems in certain circumstances, for instance in case a networked drive, or a removable drive such as optical disks. A red lamp means that the drive/directory specification or the media is invalid.

## 11.3 Problem list

In the main diagnostics screen, a list box shows all problems detected. These include description of disk problems as indicated above, and non-critical errors (warnings) such as lost or unnecessary files (such as ".PKS" files without associated gel), gels normalized with another standard than the one currently defined in the Normalization settings, marked with a yellow lamp. A detailed list of gels that are incompatible with the current Normalization settings is shown when the <***Gel files***> button is pressed.

The current Normalization settings are shown. Errors provided with a red lamp are critical and will always cause problems when running Molecular Analyst Fingerprinting software.

## 11.4 Diagnostics and User information report

A report of all hardware characteristics, user settings and problems detected is created in the Windows Notepad when the **<*Create report*>** button is pressed. GCINFO will load the Notepad program, where you will be able to view the report and print it out.

# 12. Troubleshooting

## 12.1 General

Errors during installation:

The INSTALL program displays an error message such as "**Unable to find installation file: XXX**" where XXX is a file name. You may have inserted the wrong disk into the disk drive. If this is not the case, the file is lacking or corrupted by physical damage to the disk and the installation will be aborted. Contact your dealer to ask new installation disks.

During installation, the computer displays a system message "**System error: cannot read from drive A:**" or similar. This means that some physical damage has occurred to the installation disk. The installation will be halted. In some cases, disk repair tools may help solve such problems. Contact your dealer to ask new installation disks.

The INSTALL program displays messages such **as** "**Could not create directory**", "**Could not create file**" or "**Not enough space on hard disk**". It may be that there is not enough free space (4 MB required for installation programs only). If this is not the case, check your hard disk for bad blocks using CHKDSK/F or SCANDISK from the DOS prompt.

Errors at startup:

The Startup program is started, or any of the application buttons in Startup is pressed and the error "**Not enough memory to run application**" is displayed. Instead of loading the program, Windows returns to the Program Manager. First, check if there is enough free memory available (2 MB free memory required only to LOAD the programs). You may run the Diagnostics program if this is possible. If the required memory is available it means that one or more program files are corrupted on the installation disks during shipment or on the hard disk. Reinstall Molecular Analyst Fingerprinting software and if the same happens, contact your dealer to ask new installation disks.

When the program ***Normalize*** or ***Analyze*** is started, a fatal error: "**Essential files have been modified. Reinstall software from disk**" is generated. This means that the code in one of the essential files to run Molecular Analyst Fingerprinting software has been changed for any reason. The only solution is to reinstall the software from disk. If the same happens after reinstalling the software, it may be that one of the essential files is corrupted on the installation disk. Contact in that case your dealer to ask new installation disks.

When the program ***Normalize*** or ***Analyze*** is started, a fatal error: "**Security key not recognized. Key inserted...? Printer on...?**" is shown. This means that the protection key necessary to run Molecular Analyst Fingerprinting software, is either not present in the parallel port (LPT1) or does not function properly. In the latter case, check whether the printer is switched ON, since some printers obstruct the key when they are off. If the same occurs when the printer is on, try disconnecting the printer from the key and running Molecular Analyst Fingerprinting software again. When the key is still not recognized, return it to your dealer to be replaced.

When the program ***Normalize*** or ***Analyze*** is started, or when a gel is saved, opened, a clustering is started, or another action is done, a fatal error: "**Security key not recognized. Check network status**" is shown. If this message occurs when starting the *Normalize* or *Analyze* program, this means that the Network security key necessary to run Molecular Analyst Fingerprinting software, is not found by the computer on which you are running Molecular Analyst Fingerprinting software. In this case, check whether the computer running the Security Key Server program (see 2.2.2.2) is still running properly, and whether the network is still functioning. If the message occurs within the program while doing certain operations, the reason can be that your license is stopped by the server program by the preset time-out of approx. 30 minutes. If this is not the case, the reason can be that the network is down, the server computer is down, or the network is overloaded and too slow.

In all of these cases, if you encounter security key recognition problems regularly, check whether the network is configured properly. In general, you should only install the network protocols that are used for each adaptor. For example, if NetBEUI is the default protocol used with your network adaptor, remove IPX/SPX protocols if this is installed but not used. If TCP/IP is used with a different adaptor, e.g. a dial-up adaptor, then install TCP/IP only with the dial-up adaptor, and so on.

The settings of the ***Conversion*** program are not user-specific, notwithstanding the fact that several users are defined. When they are changed for one user, they will change for others as well. The reason is that the IMAGE directory is the same for different users. Molecular Analyst Fingerprinting software stores the conversion settings in that directory, and if it is the same for different users, it will be overwritten by the last working user. Create separate directories in the User Setup, and specify the IMAGE directories accordingly.

The Normalization and Database settings are not user-specific, notwithstanding the fact that several users were defined. When they are changed for one user, they will change for the other(s) as well. The reason is that the DATABASE directory is the same for different users. Molecular Analyst Fingerprinting software stores the normalization and database settings in that directory, and if it is the same for different users, it will be overwritten by the last working user.

Create separate directories in the User Setup, and specify the DATABASE directories accordingly.

When the ***Conversion***, ***Normalization*** or ***Analyze*** program is run, no files or directories are shown in the file-load menu, or in the database. A non-existing path has been specified in the User Setup, or the directory has been removed or renamed. Run the User Setup and define a correct path for the IMAGES, RAW GELS and DATABASE.

The window background of the Molecular Analyst Fingerprinting software ***Conversion***, ***Normalization*** or ***Analyze*** programs are dithered instead of having a pure colour, and text becomes almost illegible. The graphics mode has only 16 colours. Select a 256 colour mode in the Windows Control Panel or the graphics driver, or check "***Use system colors***" in the Molecular Analyst Fingerprinting software User Setup.

When a new user name is entered to add to the user list in the Startup program, the user is not actually created. The maximum number of users is reached. Delete one or more "obsolete" users.

Starting ***Normalize*** or ***Analyze***, both fatal errors "**Essential files have been modified. Reinstall software from disk" and "Protection key not recognized. Key inserted...? Printer on...?**" are shown and the program does not run. One of the reasons can be that Molecular Analyst Fingerprinting software is installed on a network drive. Never install Molecular Analyst Fingerprinting software on a network drive since it searches for the protection key and several essential files on the computer from which it is run. You may use network drives as database and image directories, provided that these can be read-and-write accessed all the time. A second possibility is a combination of problems that have been explained earlier.

## 12.2 Conversion program

The conversion settings are not user-specific, notwithstanding the fact that several users are defined. When they are changed for one user, they will change for others as well. The reason is that the IMAGE directory is the same for different users. Molecular Analyst Fingerprinting software stores the conversion settings in that directory, and if it is the same for different users, it will be overwritten by the last working user. Create separate directories in the User Setup, and specify the IMAGE directories accordingly.

No files or directories are shown in the file-load menu, or in the database. A non-existing path has been specified for IMAGES in the User Setup, or the directory has been removed or renamed. Run the User Setup and define a correct path for the IMAGES.

The window background is dithered instead of having a pure colour, and text becomes almost illegible. The graphics mode has only 16 colours. Select a 256 colour mode in the Windows Control Panel or the graphics driver, or check "***Use system colors***" in the Molecular Analyst Fingerprinting software User Setup.

When a TIFF file is loaded, an error message "**Not enough memory to load image**" is shown. There is not enough free memory available to load the image into the computer's memory. While large memory blocks are asked by an application, Windows reallocates and rearranges the free memory, and often, the problem will be solved by trying to load the file a second time. If the problem happens often, one should extend the RAM so that at least 4 MB memory becomes available. Another solution is by increasing the Swapfile size (see Windows Control Panel ---> Enhanced ---> Virtual Memory --->Change). This is, however, not the fastest solution!

While loading a TIFF file, the error "**Invalid or unknown file type**" is shown. (1) The file is not in TIFF format; (2) the file is TIFF for Macintosh or is compressed; (3) the file is a 24 bit true RGB TIFF file. Load only uncompressed IBM-PC TIFF files, which may be 8 bits or 16 bit type.

After loading a TIFF file, the image appears inverted on the screen. The image can be inverted by checking the "***Negative***" item in the file load dialog box. Make sure that the gel is always seen as dark bands on a bright background, NOT inverted. If on the original the bands are lightening on a black background and the image is scanned or photographed as such, the ***Negative*** option should be used.

After loading the gel, the image appears but is very bright and bands are hardly seen. This will often be the case with 16 bit TIFF files. Press F6 to call the Palette edit window, decrease the brightness and increase the contrast.

When a converted gel is saved, an error message "**Gel window not present - could not save additional information**" is shown. The gel image window was closed after the tracks were defined. After defining tracks, minimize the gel image window or move it away from before the track list; do not close it.

## 12.3 Normalization program

When the Normalization program is loaded, a warning "**Normalization Settings not found - using defaults**" is produced. This happens when Molecular Analyst Fingerprinting software is freshly installed and no gels have been normalized yet. It also may occur when the file SETTINGS.NOR in the database directory has been removed, or is corrupted.

The Normalization settings are not user-specific, notwithstanding the fact that several users were defined. When they are changed for one user, they will change for the other(s) as well. The reason is that the DATABASE directory is the same for different users. Molecular Analyst Fingerprinting software stores the normalization settings in that directory, and if it is the same for different users, it will be overwritten by the last working user. Create separate directories in the User Setup, and specify the DATABASE directories accordingly.

No files or directories are shown in the file-load menu. A non-existing path has been specified for RAW GELS in the User Setup, or the directory has been removed or renamed. Run the User Setup and define a correct path for the RAW GELS.

The window background is dithered instead of having a pure colour, and text becomes almost illegible. The graphics mode has only 16 colours. Select a 256 colour mode in the Windows Control Panel or the graphics driver, or check "***Use system colors***" in the Molecular Analyst Fingerprinting software User Setup.

When a converted gel (raw gel) is loaded, the error message "**TIFF file not present**" is shown and the program displays reconstructed patterns instead of real two-dimensional strips. The most probable reason is that the converted file was not saved in the RAW GELS directory defined for the user. Correct this problem by moving the raw gel file and the corresponding "**.I1**" file to the specified path, or by reconverting the image and saving it to the right directory. In the future, always save converted gel files to the default directory for the user. A second reason could be that the TIFF file has been removed from the hard disk. A second reason could be that the *image window* in the Conversion program was closed after defining tracks and before the gel was saved. 2D-information can only be saved if the gel image window in the Conversion program is minimized or moved away from the track list; NOT closed.

When a converted gel (raw gel) is loaded, the error message "**Inconsistent track resolution of current gel and standard**" is shown, and the database standard pattern will not show at the left hand side of the window. The track resolution of the gel containing the database standard and the current gel were not the same in the Conversion program (see section 4.2.4, Track resolution). Correct the problem by noting the track resolution of the database standard in the status bar of the Normalization screen (indicated "Resolution"), and convert new gels using the same "Track resolution" value in the Track Scanning Settings in the Conversion program (section 4.2.4).

While loading a raw gel file, an Error (usually 204 at XXXX:XXXX) occurs, and the program halts. There is insufficient free memory to load the gel with 2D-strips. Extend the RAM so that at least more memory becomes available. A second solution is to increase the Swapfile size (see Windows Control Panel ---

> Enhanced ---> Virtual Memory --->Change). This is, however, not the fastest solution! A third inferior solution is to disable "***Show of 2D-gelstrips***" in the normalization settings (press F7).

After loading the gel, the gelstrips appear but are very bright and bands are hardly seen. This will often be the case with gels converted from 16 bits TIFF files. Press F6 to call the Palette edit window, decrease the brightness and increase the contrast.

Loading and saving 2D-gelstrips is extremely slow. This happens when the hard disk is not sufficiently cached to load the complete TIFF file into the caching buffer. When you are using Windows 3.1, install SMARTDRIVE with a Windows caching size of at least the size of the largest TIFF file used in Molecular Analyst Fingerprinting software. When you are using Windows 3.11 install 32 bit file access, with a cache size of at least the size of the largest TIFF file used in Molecular Analyst Fingerprinting software. This feature increases the speed of handling of large TIF files and gelstrips in Molecular Analyst Fingerprinting software enormously. Windows 95 automatically assigns a file cache depending on the available memory.

While associating a band on a pattern with a selected reference position (blue selector), the band becomes automatically associated with another, closer reference position, which is not the correct one. The selector jumps automatically to the closest reference position. If you want to keep reference position fixed, select it first, then hold the CTRL key while associating any bands with it.

## 12.4 Main program

The Database settings are not user-specific, notwithstanding the fact that several users were defined. When they are changed for one user, they will change for the other(s) as well. The reason is that the DATABASE directory is the same for different users. Molecular Analyst Fingerprinting software stores the database settings in that directory, and if it is the same for different users, it will be overwritten by the last working user. Create separate directories in the User Setup, and specify the DATABASE directories accordingly.

No files or directories are shown in the database window. A non-existing path has been specified for DATABASE in the User Setup, or the directory has been removed or renamed. Run the User Setup and define a correct path for the DATABASE.

Lists cannot be saved and loaded from the hard disk. The LISTS directory is invalid. Specify a valid directory for lists in the User Setup.

Libraries cannot be created or edited. The LIBRARIES directory is invalid. Specify a valid directory for libraries in the User Setup.

The window background is dithered instead of having a pure colour, and text becomes almost illegible. The graphics mode has only 16 colours. Select a 256 colour mode in the Windows Control Panel or the graphics driver, or check "***Use system colors***" in the Molecular Analyst Fingerprinting software User Setup.

When a gel image is shown, the image has false colours that do not match in the selected palette. You are probably running other applications that require colour palettes simultaneously on a display in 256 colours mode. Try minimizing and restoring Molecular Analyst Fingerprinting software. If the problem persists, restart Molecular Analyst Fingerprinting software and do not bring the other conflicting application in the foreground as long as you are working in Molecular Analyst Fingerprinting software. If you are often working with colour-requesting applications in combination with Molecular Analyst Fingerprinting software, choose a 65000 colours mode.

When changing the brightness and contrast of an image using the palette edit window, the colours are not adjusted dynamically ("animation"), but only after pressing OK or Preview. Your computer operates in a graphics mode other (higher) than 256 colours. This is not a real obstacle.

When ***2-D Gel image*** is shown for a gel or list, crossed lines are seen instead of gelstrips. The gelstrips were not normalized with the patterns. This can have several reasons: (1) the tracks were scanned as line tracks (densitometric curves) using a laser scanner or similar; (2) the TIFF file for the gel was not found by the Normalization program (see 12.3) (3) the "***Save 2D-gelstrips***" option in the Normalization settings was not checked when the patterns were normalized. In the first case, select "***Reconstructed gel image***" instead of ***2-D Gel image***. In the second case, reconvert the gel if necessary and renormalize it. In the third case, renormalize the gel with the ***Save 2D-gelstrips*** enabled.

The gel in the database looks not normalized. While normalizing the gel, probably all the associations were made, but they were not carried out using the ***Align associated peaks*** command before saving. Renormalize the gel.

A number of menu options such as ***Clustering (correlation)***, ***Clustering (bands)***, etc. are grayed and are not available. This is because there is no list present in the List window (the right subwindow). These tools are based on selections of patterns from the database, called lists. First select a list of patterns you want to study.

When patterns from different gels are selected into a list, an error message "Different standard or track resolution. Run diagnostics" is shown. You are trying to select tracks that are normalized using a different database standard or

using the same standard, but with a different track resolution specified in the Normalization program (see Normalization settings, section 5.2.2). Find out where the problem lies by calling ***File|Information*** in the gel window of incompatible gels, and comparing all settings, in particular the items "**Standard**" and "**Resolution**". Renormalize incompatible gels with the same resolution and database standard.

When Clustering (bands) is selected for a list of patterns, the error message "**No bands defined**" is shown. Before a cluster analysis or identification based on band positions and/or area can be performed, the bands must be defined for the patterns. This is achieved for a gel by calling ***Gel|Bands*** in the gel window.

When ***Gelstrips*** are shown in the dendrogram window, crossed lines are seen instead of gelstrips. The gelstrips were not normalized with the patterns. This can have several reasons: (1) the tracks were scanned as line tracks (densitometric curves) using a laser scanner or similar; (2) the TIFF file for the gel was not found by the Normalization program (see 12.3) (3) the "***Save 2D-gelstrips***" option in the Normalization settings was not checked when the patterns were normalized. In the first case, select "***Reconstructed gel image***" instead of ***2-D Gel image***. In the second case, reconvert the gel if necessary and renormalize it. In the third case, renormalize the gel with the ***Save 2D-gelstrips*** enabled.

While calculating a large dendrogram, an Error (usually 204 at XXXX:XXXX) occurs, and the program halts. There is insufficient free memory to calculate the dendrogram. Extend the RAM or close other programs so that more memory becomes available. Another solution is to increase the Swapfile size (in Windows 3.1: Control Panel > Enhanced > Virtual Memory > Change. In Windows 95: Settings > Control Panel > System > Performance > Virtual Memory > set maximum). Releasing or extending real memory is highly preferable in view of computing speed!