Constructing the Perfect Data Base

Brian Rohrback and Scott Ramos,

Infometrix, Inc., 10634 E. Riverside Drive Suite 250, Bothell, WA 98011

In quality assessment, there is often a need to store prior data so it can be used as a reference to improve control. This paper addresses the problem of instrument drift in constructing a data base and marks how an algorithm used in voice recognition can vastly improve the consistency of the data stored.

Abstract

Reference data bases are great. There is tremendous value in having access to spectra of known materials, particularly in research environments where the unknown is being characterized. As data base technology moves into the routine assessment realm, the needs shift so that custom, product-specific data bases are more useful. In this paper,

1. We outline the creation of a Bio-Rad KnowItAll® database containing several years' worth of a refinery's chromatographic data to serve as the basis for a company-wide, refinery-agnostic resource.
2. And, we accommodate environmental effects on Raman spectroscopy so we can directly compare spectra at vastly different temperatures and pressures, in preparation for creating a data base for process Raman.
3. With AnalyzeIt MVP®, we can take advantage of this improvement in data quality, enabling automated assessment of spectra or chromatograms in a global QC arrangement.

The key technology is a multivariate/chemometric technique designed to adjust for retention time or wavelength shifts in a fully automated manner.

Background – Why QC Data Bases are Hard to Build

Analytical instruments are critical to the efficient functioning of nearly any QC laboratory. Unfortunately, an instrument's response will change, albeit usually slightly and slowly, over time. When we look to compare one instrument's result with the analogous result from sometime in the past, the two traces will not superimpose precisely. This can be seen both in the intensity dimension and in the relative position along the abscissa (time in chromatography, wavelength in optical spectroscopy). Intensity changes are most often caused by hardware (*e.g.*, the detector is fatiguing, the flow path is changing) and there are accepted transfer of calibration mechanisms to correct the problem.

This discussion deals with the "x-axis shift" problem. The most generic solution we have found was borrowed from voice recognition technology. Here the signal is shifted or warped in a non-linear fashion to make the time axis (in chromatography) or the wavelength axis (in spectroscopy) be a closer match to a standard run of the same or a similar material. The approach used here is called Correlation Optimized Warping and is contained in the commercial product LineUp™.

Chromatography

Chromatographic data was assembled from a refinery support laboratory that spanned a 6-year period and was collected on 5 different GCs, all running the same chromatographic conditions and using the same column characteristics. The conditions were:

- 100 m dimethylpolysiloxane column, 0.25 mm i.d.
- Helium carrier, FID
- 175 minute run (35°C for 10 min, 1°/min to 200°)
- Data rate 5 Hz

To build this into a reliable database, we needed to align all of the fractions to make the time axis consistent. The problem is exemplified by looking at one collection of fractions, these of the naphtha samples as in Figure 1 below:

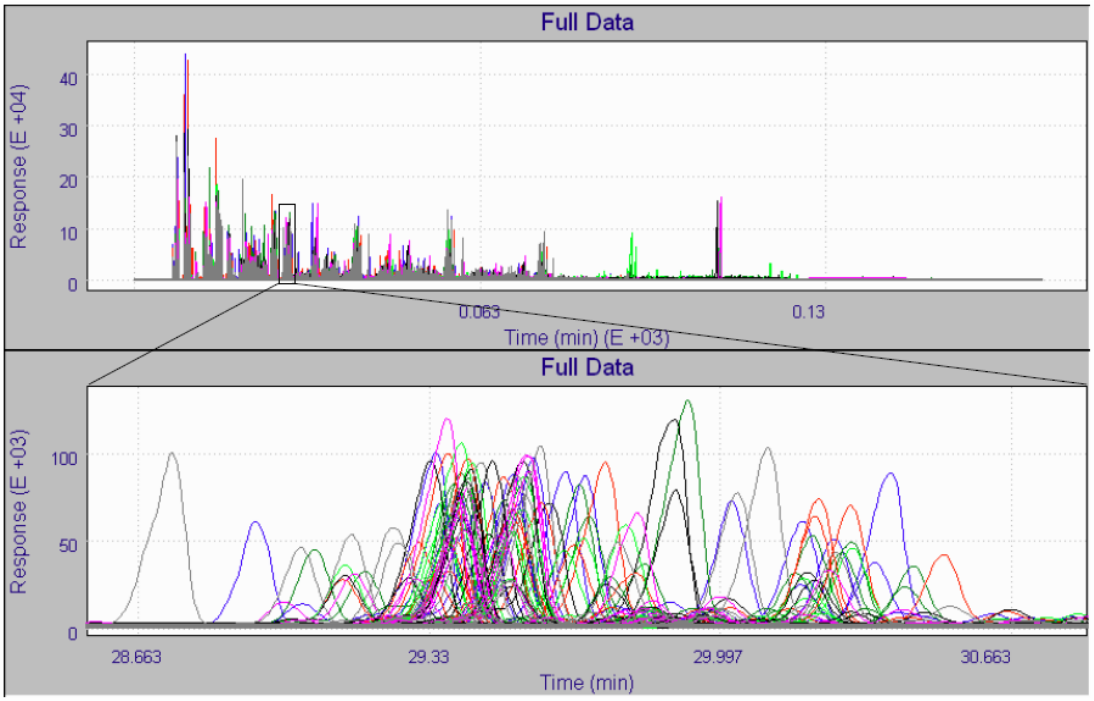

Figure 1: Naphtha profiles showing variation in retention time.

Using alignment technology in LineUp, we can remove most of the misalignment seen in data collected by different instruments over several years' worth of time

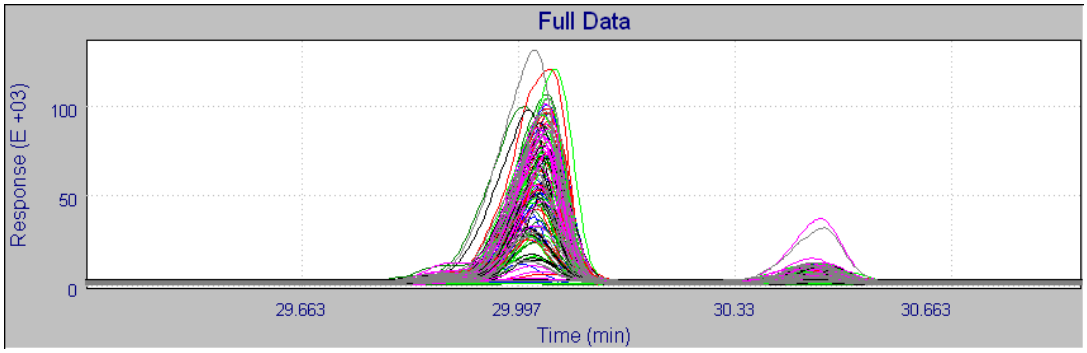

Figure 2: Naphtha profiles after alignment.

Because the data included fractions with very different populations of analytes, there could be no single alignment standard that would serve all fractions. This leads to a different problem. We can align naphthas to a single naphtha standard, but because naphthas and alkylates do not have peaks in common, these different materials do not align to one another.

The solution was to perform a two-step alignment, using a Kovats-index approach first (part of the Pirouette® software) to align the chosen fraction target chromatograms to an n-paraffin standard run. Once this alignment was completed, the once-aligned targets were used to LineUp all of the other chromatograms. The procedure:

1. For each target, we identify the immediately-prior QC sample in the same batch;
2. We then choose one material type as the global target (we used gasoline for the most complete set of peaks);
3. We then align each QC sample to the global QC sample using the Kovats-index approach;
4. We align each target to the global QC sample using the alignment positions calculated in the previous step; and
5. We re-align the sample profiles to these newly-aligned targets using LineUp.

The result is a time-axis-consistent set of chromatograms, such that overlays are possible regardless of the GC of origin, the date of analysis or the fraction used. The original chromatographic fragments are shown compared to the aligned traces in Figure 3.

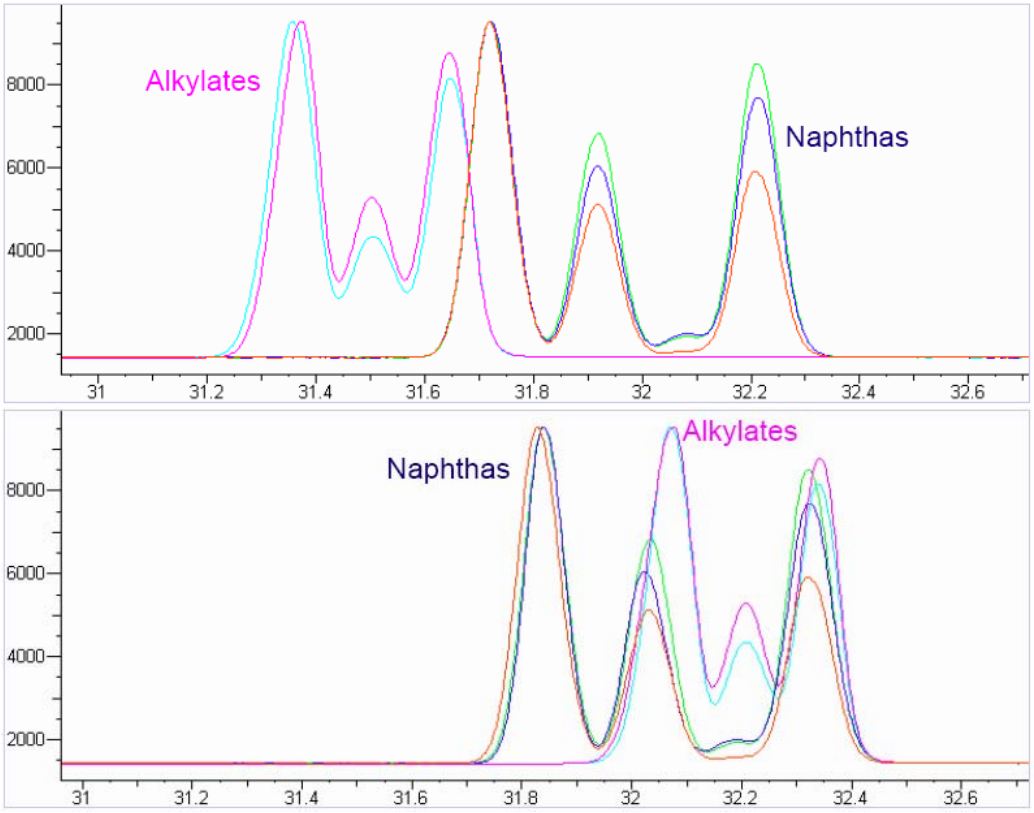

Figure 3: Unaligned and aligned naphthas and alkylates.

This allowed us to import a significantly more consistent set of chromatograms into a search and sort facility within the Bio-Rad KnowItAll database. A PCA plot of the entire set of refinery alkylate, naphtha and gasoline chromatograms are shown within KnowItAll in Figure 4.

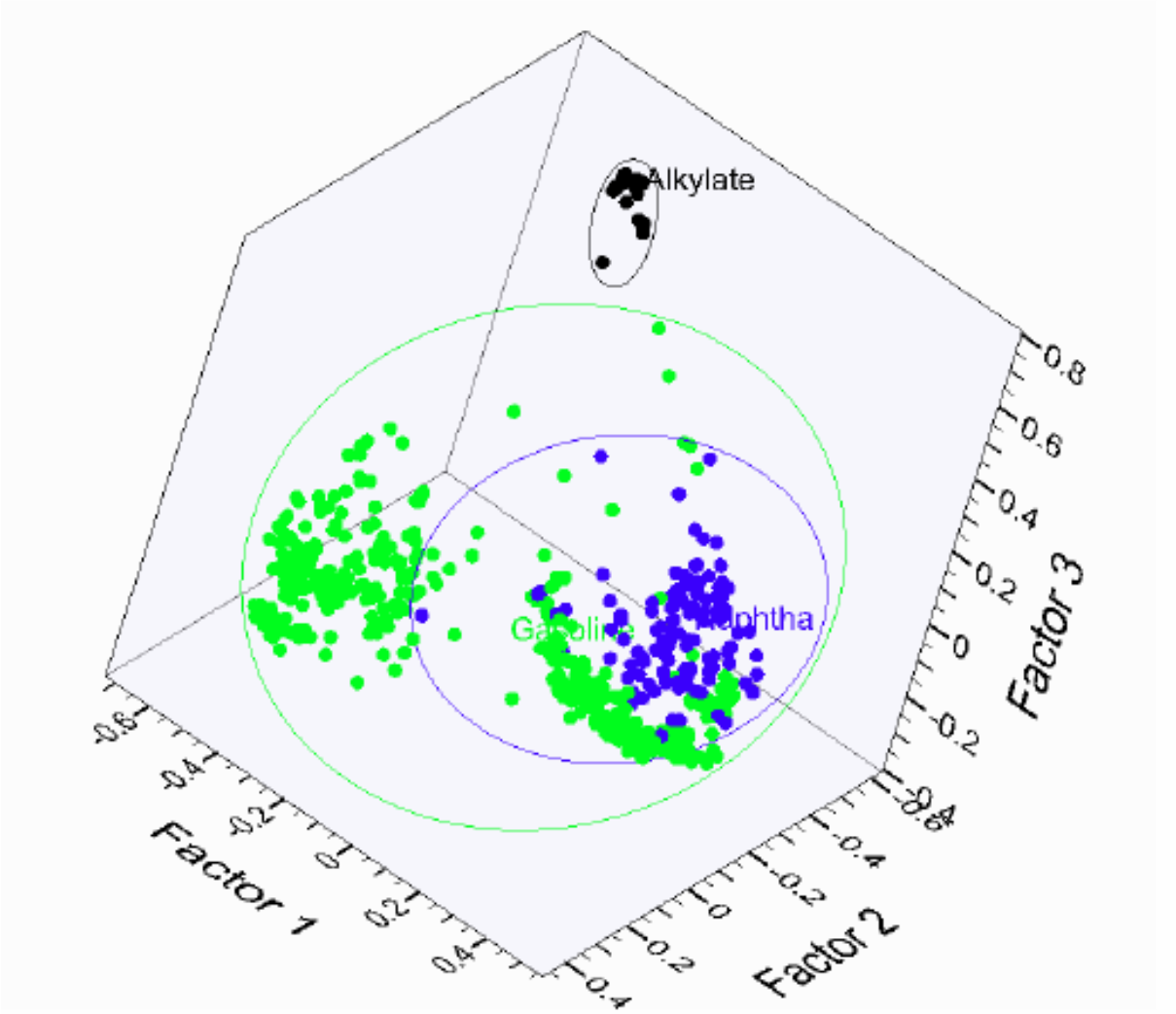

Figure 4: Three fractions within Bio-Rad's AnalyzeIt MVP environment.

The alignment standards selected in this study could be used immediately to reprocess all data collected by this refinery, even those collected by outside contractors or other oil companies, and add them to the existing database.

Spectroscopy

Alignment issues plague spectroscopy as well as chromatography, making spectra difficult to compare when collected at different times, by different instruments, or under different environmental conditions. The collection of spectra reported here is impacted by environmental conditions and requires some correction for temperature and pressure variability under extreme process conditions. In this case, we are modeling the collection of spectra from deep ocean vents, where extreme pressure and 300°C temperature swings are common. The changes in environmental conditions were responsible for significant relocations of the analytical peaks in the Raman spectra.

To simulate the environment in the lab, a NeSSI permeation tower was made of acrylic and utilized a silicone membrane. Because the permeation of vapor is temperature dependent, a heater was added to maintain constant temperature. Hydrocarbon vapors were generated using nitrogen as carrier gas supplied by a compressed gas tank. A ballprobe immersion optic with synthetic sapphire as the focusing element and sampling window was used to collect the data.

Raman spectroscopy is a great optical technique for fingerprinting molecules, however accurate identification of molecules requires that the wavelength axis be properly calibrated. Changes in environmental conditions (*i.e.*, temperature and pressure) can cause shifting of spectral features and must be removed to retain accuracy and reproducibility of multivariate models, similar to the shifting problem seen in chromatography. An example is shown in Figure 5.

Alignment of spectral features is not traditionally applied to Raman spectroscopy as Raman scattering is a fundamental property of a molecule based on its vibrational energy. Thus, because the wavelength axis (Raman shift) is calculated from the excitation wavelength, resulting spectra are comparable regardless of the wavelength used for excitation if the calibration is accurate. However, high temperature and pressure effects can induce a shift in Raman spectral features, as shown in Figure 6. These effects must be reduced or removed in order to generate accurate calibration models.

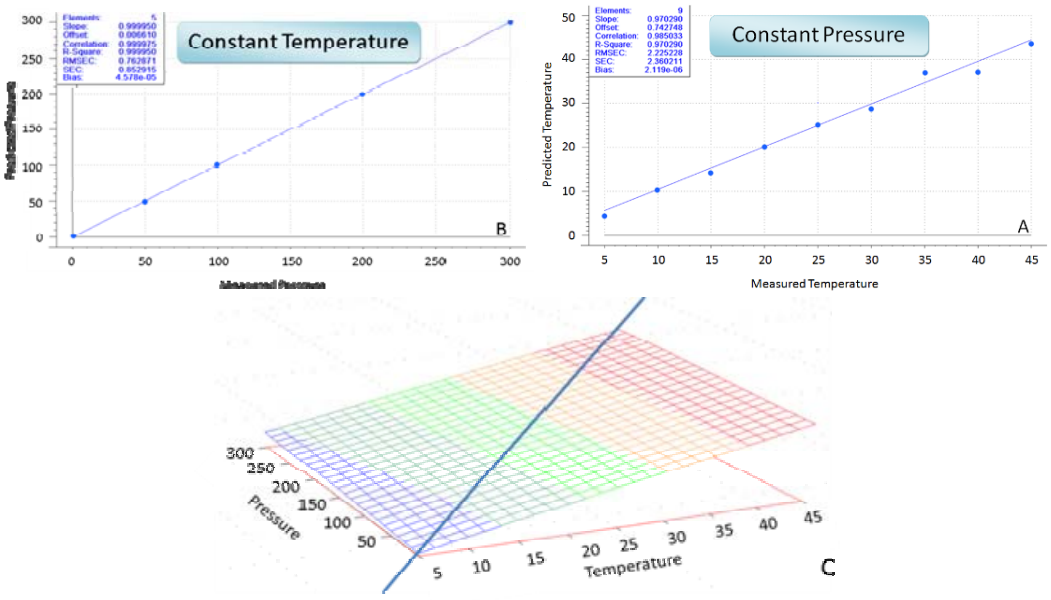

Figure 6: Peak shift is linear with pressure and temperature.

In addition to aligning spectra collected at differing environmental conditions, the feasibility of the LineUp algorithm to maintain calibration was evaluated. A test dataset of neon spectral lines was collected for this experiment. Three spectra were purposely calibrated incorrectly by slightly changing the input wavelengths. These three spectra were then aligned using the algorithm to a properly calibrated neon spectrum (Figure 7). The resulting four spectra are all properly aligned and hence calibrated offline after data collection. This algorithmic approach holds potential for maintaining/correcting instrument calibration during long term experiments by aligning each new spectrum to a standard calibrated Raman spectrum as a post measurement data treatment protocol.

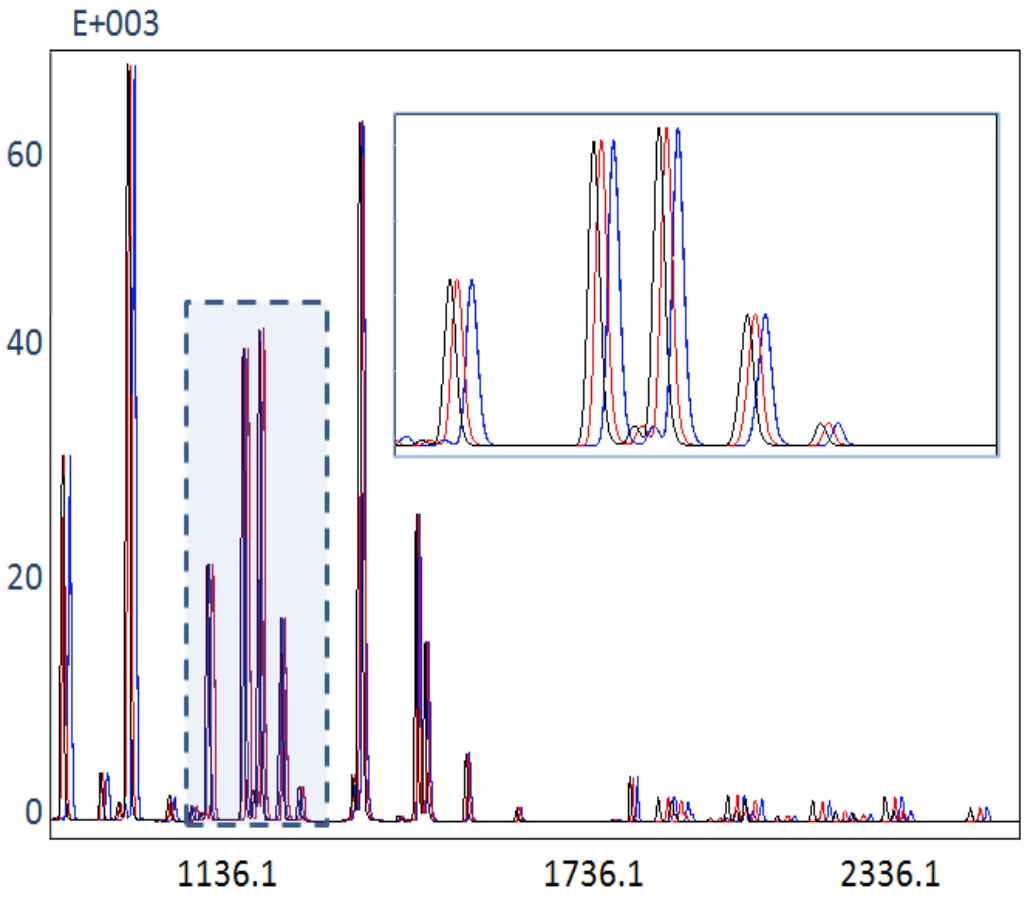

Figure 7a: Changing the input wavelength to cause misalignment.

In this case, the purposely-misaligned data can easily be corrected by aligning to a sample with a properly-calibrated neon spectrum.

This opens the possibility for long term "auto-calibrated" deployments of Raman instrumentation for process analysis or environmental monitoring when applying the alignment algorithm to maintain calibration.

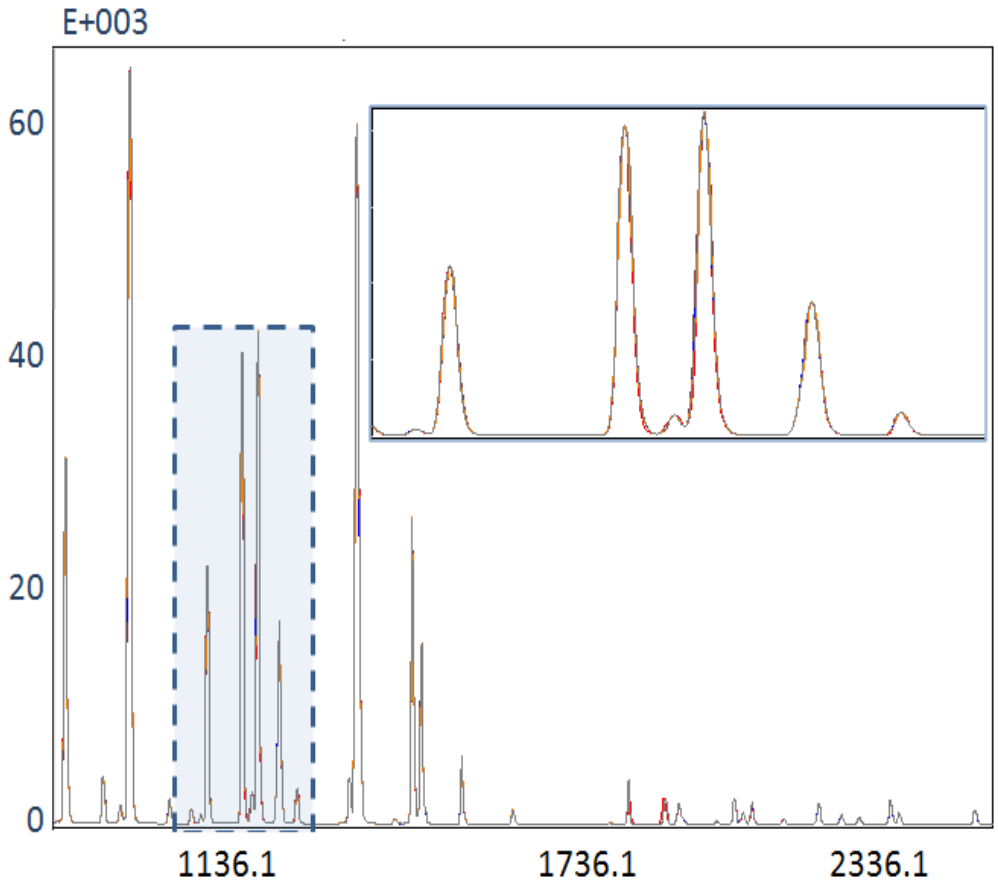

Figure 7b: Aligning the same spectral features.

Conclusions

Data collected by analytical instrument systems can shift with time and environmental conditions. This causes a problem when assembling a large-scale data base. The problem worsens as multiple instruments supply data to a single data base. The use of a multivariate alignment algorithm allows data to be corrected for these changes. The algorithm found to be the most useful originated from warping studies in voice recognition, and can be used to correct both chromatographic and spectroscopic data. The requirements are that the data base contains one sample that can serve as a gold standard or alignment target.

In chromatography, the approach works well on chromatograms collected on different instruments and columns. In cases where there are vast differences among samples (possibly no peaks in common), a two-step alignment, first to an internal or external standard using the Kovats-index and second using the multivariate warp will result in a consistent data base. In this way, even profiles of samples with few peaks in common can be aligned using marker positions from other samples.

In order to apply multivariate prediction algorithms on spectral data accurately, the location of the information peaks must remain consistent. However, environmental conditions can have an effect on peak location and thus cause error in predictions. The same multivariate alignment algorithm used to correct drift in chromatographic data can also be effectively applied to Raman spectra to correct for shifts in peak locations. It is shown that instrument drift can be removed and correct calibration maintained using this approach.

- Chemometric alignment will be essential for data-rich measurements to assure consistent data.
- Alignment can be automated for plant use.
- These techniques can reduce the burden on highly skilled manpower to interpret complex data

As a result, the data can be used to create and maintain consistent corporate-wide databases and applied to historical and future analysis. We have found Bio-Rad's KnowItAll well suited to handling the large, custom data bases required to deploy best-practices worldwide.

Acknowledgements

John Crandall, Falcon Analytical, Ronceverte, WV
Gregory Banik, Bio-Rad Informatics Division, Philadelphia, PA
Carl Rechsteiner, Chevron Energy and Technology Company, Richmond, CA
Brian Marquardt, Applied Physics Laboratory, University of Washington, Seattle, WA