IT Solutions Focus

MULTIVARIATE ANALYSIS OF HIT LISTS FROM SPECTRAL SEARCHES

Gregory M. Banik, Ph.D. and Marie Scandone, Bio-Rad Laboratories, Inc.

Principal Component Analysis (PCA) is a well-established technique in chemometrics for performing multivariate analyses on spectral and chromatographic data to simplify and clarify the massive amount of data that can result from a typical experiment.

The application of this technique has covered many fields such as the evaluation of quality control spectra or characterising control versus treated samples in a metabolomics experiment.

However, the use of PCA to analyse and visualise spectral hit lists generated from searching one or more reference databases is not well known. This technical note describes an example of the successful use of PCA to analyse a query and the hit list resulting from an IR spectral search.

> ***"PCA ANALYSIS OF THE QUERY AND HIT LIST SPECTRA PRODUCED SCORES THAT SHOW THE SPECTRA VERY NICELY SEPARATED ACCORDING TO THEIR TYPE"***

Author Details:

Gregory M. Banik, Ph.D. and Marie Scandone, Bio-Rad Laboratories, Inc., Informatics Division, 3316 Spring Garden Street, Philadelphia, PA 19104, USA

MATERIALS AND METHODS

An IR query spectrum representing a combination of nylon and rayon *(Figure 1)* was searched against the Sadtler "IR - Fibers by Microscope" database using the SearchIt™ application (Euclidian Distance search algorithm; maximum hits retrieved set to 50) in Bio-Rad's KnowItAll® Informatics System. The resulting hit list was transferred to AnalyzeIt™ MVP, an application that is a joint development between Bio-Rad Laboratories and Infometrix, Inc., a leader in chemometrics.

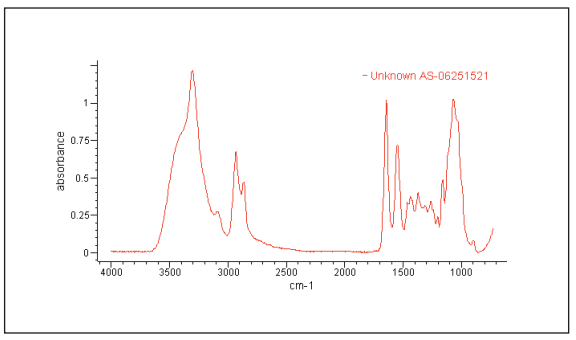


Figure 1. IR Spectrum: Combination of Nylon and Rayon

In AnalyzeIt MVP, the 50 hits and single query spectrum were subjected to Principal Component Analysis using mean-center pre-processing only. No Y-transformations were performed.

RESULTS

Principal Component Analysis

Searching the reference IR database with the query in *Figure 1* resulted in a hit list of reference compounds comprised of 35 rayon spectra and 15 nylon spectra. PCA analysis of the query and hit list spectra produced scores that show the spectra very nicely separated according to their type (*Figure 2*).

Each point in a scores plot represents a spectrum (a point for each hit and a point for the query spectrum), and similar spectra will tend to cluster in similar areas of the plot. From the scores plot, it is clear that all of the rayon hits from the reference database are similar to one another, as are all of the nylon hits from the reference database. The nylon/rayon mixture query spectrum, however, is positioned in the plot between the groups of nylon and rayon reference spectra, suggesting that it is not closely similar to either, but may share some features of each.

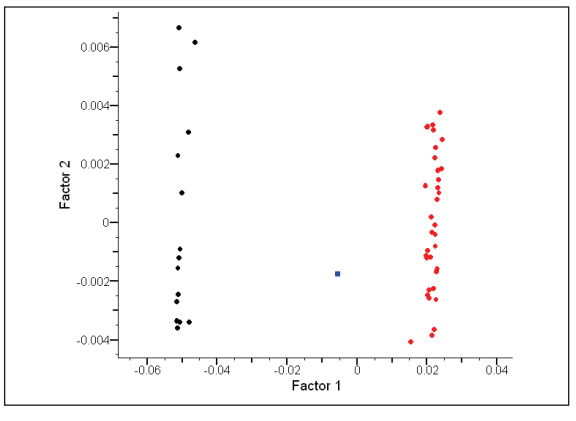


Figure 2. PCA Scores Plot - ■ Query Spectrum; ● Rayon Hit List Spectra; ● Nylon Hit List Spectra.

Overlap Density Heatmaps

Comparative visualisation of all the rayon hit list spectra using Bio-Rad's patent pending Overlap Density Heatmap technology gives a graphical representation of the similarity and dissimilarity of this group of spectra.

In the Overlap Density (OD) Heatmap *(Figure 3)* at OD Level of 0 (showing all areas of overlap density), the areas of highest overlap density in the overlaid spectra are displayed in red, areas of lowest overlap density are shown in purple, and all regions of moderate overlap are displayed in the intermediate colors.

By selecting only the PCA scores corresponding to the rayon spectral hits, the corresponding OD heatmap confirms that there is a high degree of similarity among these spectra: the heat map presents an image that is predominantly red, indicating a high degree of commonality among the spectra. Similarly, the 15 nylon hits display a high degree of overlap density *(Figure 4)*.

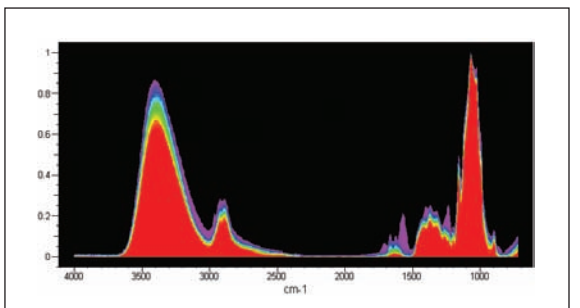


Figure 3. Overlap Density Heatmap of 35 Rayon Hits from Reference Database (OD Level = 0).

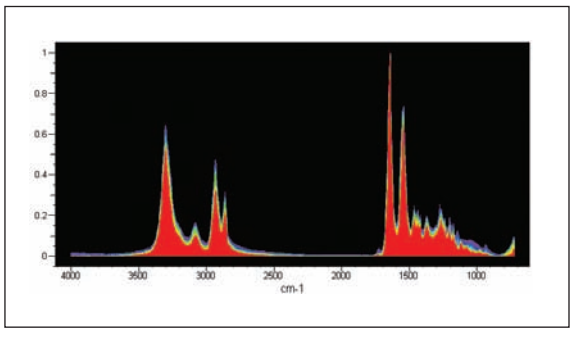


Figure 4. Overlap Density Heatmap of 15 Nylon Hits from Reference Database (OD Level = 0).

When selecting PCA scores that represent the spectra for the query as well as the rayon hits, the unique areas of the corresponding OD heatmap are clearly displayed *(Figure 5)*. The purple area of the heatmap represents those regions of the grouped spectra that are most unique (i.e., least common).

The areas of highest overlap density can be hidden by adjusting the OD level to a negative value which retains in the view only those regions of the spectra not common to all spectra, that is, that are unique to some of the spectra.

The remaining spectral features *(Figure 6)* are very similar to the OD Heatmap of the nylon only reference spectra *(Figure 4)*. In other words, the unique features of the OD Heatmap formed by combining the query spectrum and the rayon reference spectra resemble the nylon spectra.

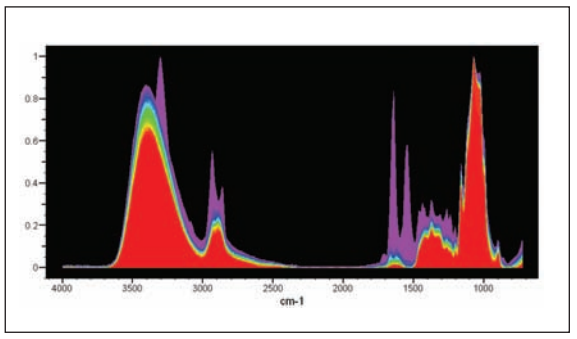


Figure 5. Overlap Density Heatmap of Query Spectrum and 35 Rayon Hits from Reference Database (OD Level = 0).

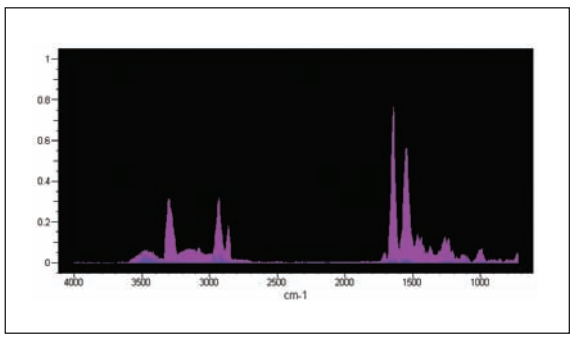


Figure 6. Overlap Density Heatmap of Query Spectrum and 15 Nylon Hits from Reference Database (OD Level = -88)

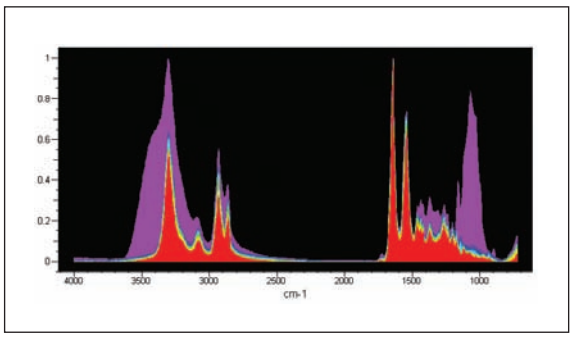


Figure 7. Overlap Density Heatmap of Query Spectrum and 15 Nylon Hits from Reference Database (OD Level = 0).

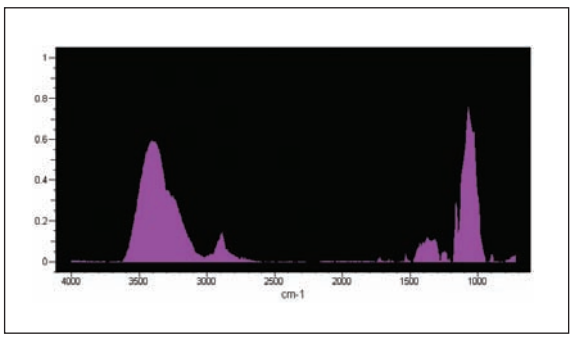


Figure 8. Overlap Density Heatmap of Query Spectrum and 35 Rayon Hits from Reference Database (OD Level = -88)

Conversely, when combining the spectra for the query as well as the nylon hits from the reference database in an Overlap Density Heatmap at OD Level 0, the unique areas of overlap density (the areas in purple) are also clearly displayed *(Figure 7)*. If the areas of highest overlap density are removed by adjusting the OD level, the remaining spectral features *(Figure 8)* are very similar to the OD Heatmap of the rayon only reference spectra *(Figure 3)*. In other words, the unique features of the OD Heatmap formed by combining the query spectrum and the nylon reference spectra resemble the rayon spectra.

CONCLUSIONS

Principal Component Analysis appears to be a valuable tool to analyse the results of standard spectral searches - a spectral query and hit list - providing useful insights into the nature of the compounds in the hit list relative to the query. Overlap Density (OD) Heatmaps not only confirm the value of the technique, but are also a useful complement to the multivariate processing capabilities afforded by PCA.

Acknowledgment

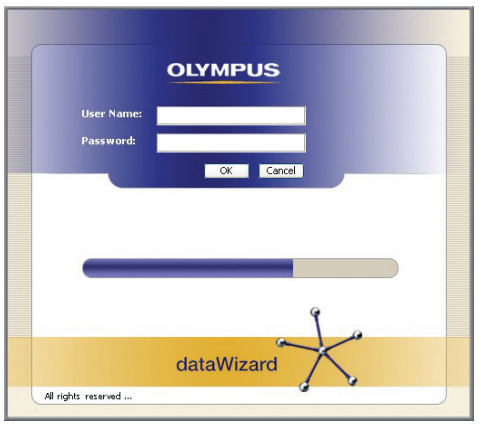
We would like to extend special thanks to Brian Rohrback and Scott Ramos of Infometrix, Inc. for their collaborative efforts.

Lost in Translation?

Even the simplest information can be misunderstood in a multilingual environment, so imagine being able to log in your samples in the US (in English) with the data available immediately and accurately to your laboratory in Japan, in Japanese! The latest LIMS innovation from **Quality Systems International (QSI)** is sophisticated language handling that allows WinLIMS users to run their system in their chosen language. By translating user interface elements like screens, menus and dialogue boxes, etc., WinLIMS gives individual users the convenience and ease of working that comes from using their native language. Global companies using centralised systems can move data around the world without the fear that vital information will be "lost in translation". Operators can have their preferred language assigned to their WinLIMS staff record, and by simply logging in will then immediately see the whole system in that language. Users have the ability to change or update the translations used to reflect changing situations and can even translate actual data - examples might be product names, method or parameter names or details of raw materials. Currently, WinLIMS can be supplied with the following languages already set up: English, German, French, Dutch, simplified Chinese and Portuguese, but the user can set up any other language that might be needed, including languages that require double byte character sets like Chinese, Japanese or Korean.

Circle no. 271

New Data Management System



Olympus Life and Material Science Europa GmbH – Diagnostics, has announced the launch of a new advanced data management system for pathology laboratories in March. Olympus dataWizard offers the latest features, including comprehensive audit tracking, and can be installed for single and multi-site use.

Olympus dataWizard has been specifically designed for use with Olympus AU clinical chemistry analysers, the new AU3000i immunochemistry analyser and the OLA2500 pre-analytical automation system. Already the instruments of choice for "mega" laboratories, the Olympus systems are able to reach new levels of throughput when workflow is optimised by the new software. Another outstanding advantage is not only the connectability of 3rd party analysers but also any other analytical workplace instrumentation. The software allows for retesting depending on initial results. Final results are fed into the existing LIMS system.

Furthermore, Olympus dataWizard meets all the requirements for comprehensive audit trail creation. In addition to sample distribution and processing data, the software also captures reagent information. Every lot and bottle number is stored with every result issued by an Olympus AU system. Another tier of traceability is added by linking and storing the previous and following QC result for the specific parameter.

Circle no. 272

Batch Certificates Available On-line

Barloworld Scientific is offering a new service to customers by placing batch certificates for Pyrex® and E-Mil® flasks on their website.

Batch certificates for volumetric flasks are required by scientists in order to trace their manufacture to known standards. Now, end users can simply access the Barloworld website, enter the batch number into the required field and then download a batch certificate directly, free of charge and without the need to contact a distributor.

The batch certificates can be obtained by logging on the Barloworld Scientific website at and selecting 'downloads' followed by 'batch certificates'.

Thanks to recent advances in laboratory glassware manufacturing technology, Pyrex and E-Mil class A volumetric flasks are also now supplied with individual serial numbers and date of manufacture etched into the base of the product. The date of manufacture allows the exact age of a flask to be known at a glance, without reference to paper certificates.

Circle no. 273