

Multivariate Analysis of Hit Lists from Spectral Searches

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

Abstract

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 characterizing control versus treated samples in a metabolomics experiment.

However, the use of PCA to analyze and visualize 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 analyze a query and the hit list resulting from an IR spectral search.

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 (Euclidean Distance search algorithm; maximum hits retrieved set to 50) in Bio-Rad's KnowItAll® Informatics System.

The resulting hit list was transferred to Analyzelt™ MVP, an application that is a joint development between Bio-Rad Laboratories and Infometrix, Inc., a leader in chemometrics.

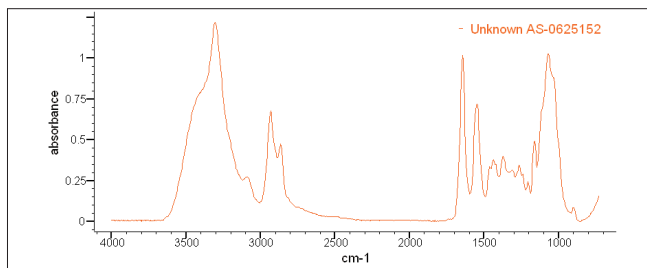


Figure 1. IR Spectrum: Combination of Nylon and Rayon.

In Analyzelt 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.

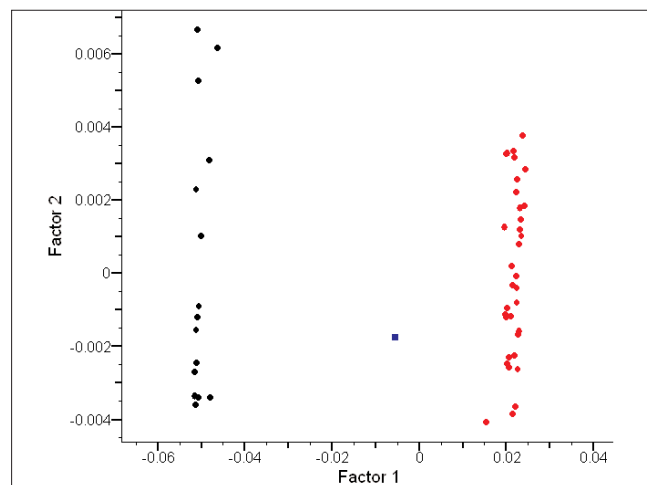


Figure 2. PCA Scores Plot ■ - Query Spectrum; ● - Rayon Hit List Spectra; ● - Nylon Hit List Spectra.

Overlap Density Heatmaps

Comparative visualization 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).

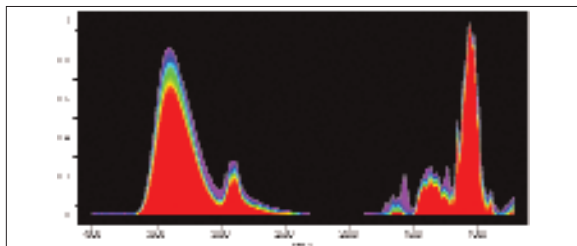


Figure 3. Overlap Density Heatmap of 35 Rayon Hits from Reference Database (OD Level = 0).

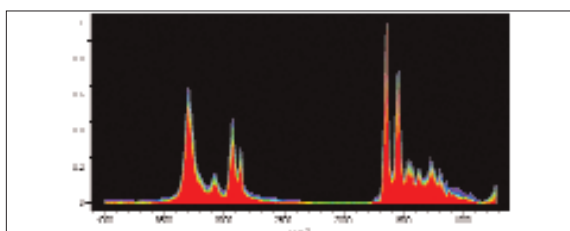


Figure 4. Overlap Density Heatmap of 15 Nylon Hits from Reference Database (OD Level = 0).

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.

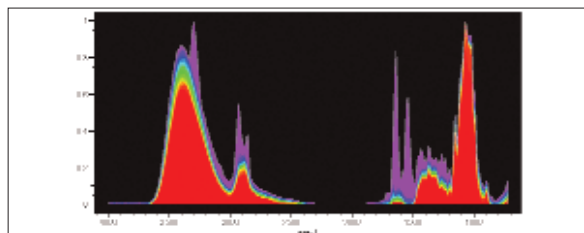


Figure 5. Overlap Density Heatmap of Query Spectrum and 35 Rayon Hits from Reference Database (OD Level = 0).

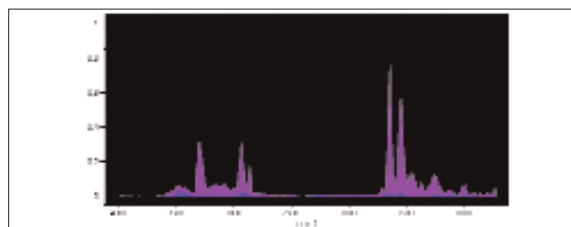


Figure 6. Overlap Density Heatmap of Query Spectrum and 35 Rayon Hits from Reference Database (OD Level = -88).

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.

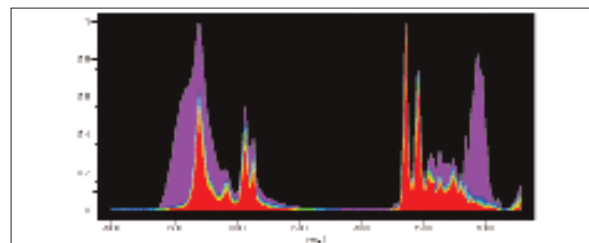


Figure 7. Overlap Density Heatmap of Query Spectrum and 15 Nylon Hits from Reference Database (OD Level = 0).

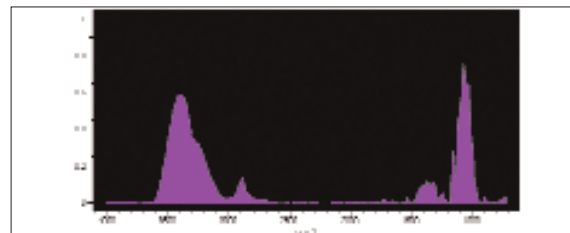


Figure 8. Overlap Density Heatmap of Query Spectrum and 15 Nylon Hits from Reference Database (OD Level = -88).

Conclusions

Principal Component Analysis appears to be a valuable tool to analyze 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.

First published *Int. Labmate* (2006) Vol. XXXI Iss. II., p. 76



**Bio-Rad
Laboratories, Inc.**

Informatics Division
www.knowitall.com

China

Phone: +1 215 382 7800 • E-mail: informatics.china@bio-rad.com

Europe

Phone: +44 20 8328 2555 • E-mail: informatics.europe@bio-rad.com

Japan

Phone: +81 03 (5811) 6287 • E-mail: informatics.nbr@bio-rad.com

Rest of World

Phone: +1 215 382 7800 • E-mail: informatics.row@bio-rad.com

U.S. Sales

Phone: +1 215 382 7800 • 1 888 5 BIO-RAD (888-524-6723) • E-mail: Informatics.usa@bio-rad.com