

New Approach for Spectroscopic Analysis Applied to Infrared Spectroscopy

Marie Scandone*, Gregory M. Banik, Ph.D., Ty Abshear, and Don Tucker

Bio-Rad Laboratories, Inc., Informatics Division, 3316 Spring Garden Street, Philadelphia, PA 19104 USA

Abstract

Methods such as Principal Component Analysis (PCA) to perform multivariate analyses on spectral and chromatographic data have been a mainstay of chemometrics for years. This poster describes a new method that combines cheminformatics tools with chemometrics tools for PCA in an intuitive environment for performing such analyses. A new patent-pending technology—Overlap Density Heatmaps (ODH)—now allows the comparative visualization of heretofore unheard of numbers of spectra or chromatograms. Overlap Density Heatmaps are used for visual data mining and analysis to assess the similarities and dissimilarities in large amounts of spectral, chromatographic, and other graphical data. This new approach for spectroscopic analysis will be examined in specific case studies as applied to IR and Raman data. We will demonstrate the successful use of PCA and ODH to analyze a query, and the hit list resulting from an IR spectral search, as well as an overall analysis of a database.

Materials and Methods

A polymer mixture was used as a query spectrum and searched against the Sadtler "IR-Monomers & Polymers (Comprehensive) Database which contains over 11,000 spectra. Using the SearchIt™ application of Bio-Rad's KnowItAll® Informatics System, the parameters were set for the Euclidean Distance search algorithm and the maximum hits retrieved were set to 50. 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.

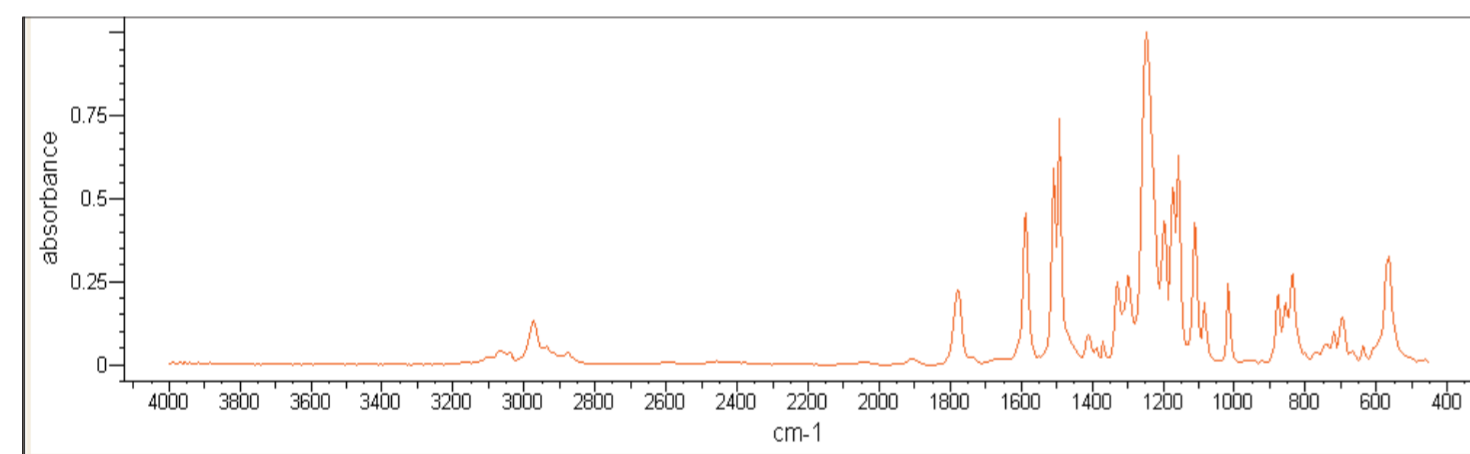


Figure 1. Unknown polymer mixture.

In Analyzelt MVP, the 50 hits and single query spectrum were subjected to Principal Component Analysis using mean-center pre-processing only. The "Maximum Factors" were set to three. There was no "Binning/Bucketing" and "Ranges" used. The Y-Transforms selected were 2nd Derivative, with the number of points set to 15, Smooth, with the number of points set to 15, and SNV (Standard Normal Variate).

The 2nd derivative and smoothing transforms are based on a Savitzky-Golay polynomial filter. This method applies a convolution to independent variables in a window containing a center data point and n points on either side. A weighted second-order polynomial is fit to these $2_n + 1$ points and the center point is replaced by the fitted value. The transforms differ in the weighting coefficients.

When using the "Number of Points" to specify the number of (window) points, the number of points must be less than the number of independent variables; otherwise the run aborts.

SNV is another approach to compensating for scattering. It can be described as row-autoscaling. The mean and standard deviation of a sample are first computed based on included variables; the value for each included variable is corrected by first subtracting the mean, then dividing by the standard deviation.

Results

Searching the reference IR database with the query spectrum in Figure 1 resulted in a hit list of reference compounds comprised of 34 polysulfone spectra and 16 polycarbonate 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). When viewing the scores plot, each point 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. Therefore, from the scores plot, it is clear that all of the polysulfone hits from the reference database are similar to one another, as are all of the polycarbonate hits from the reference database. The mixture query spectrum, however, is positioned in the plot between the groups of polysulfone and polycarbonate reference spectra. The positioning suggests that it is closely similar to the polysulfone group.

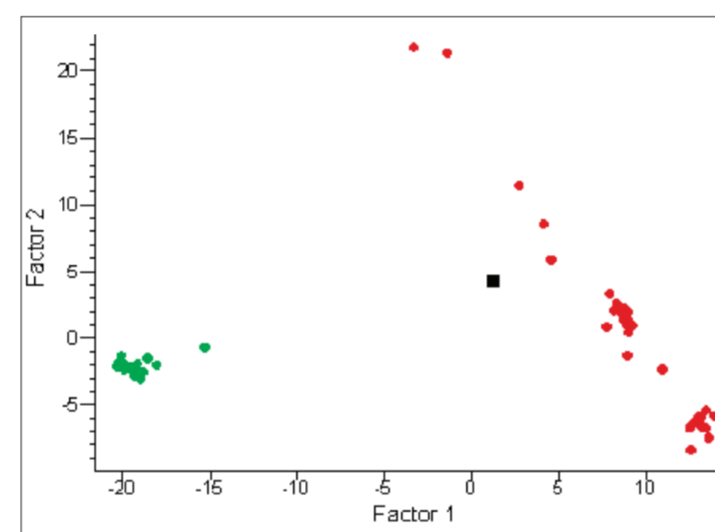


Figure 2. 2D PCA scores plot displaying Polycarbonate (●) Hits, Polysulfone (●) Hits, and Query (■).

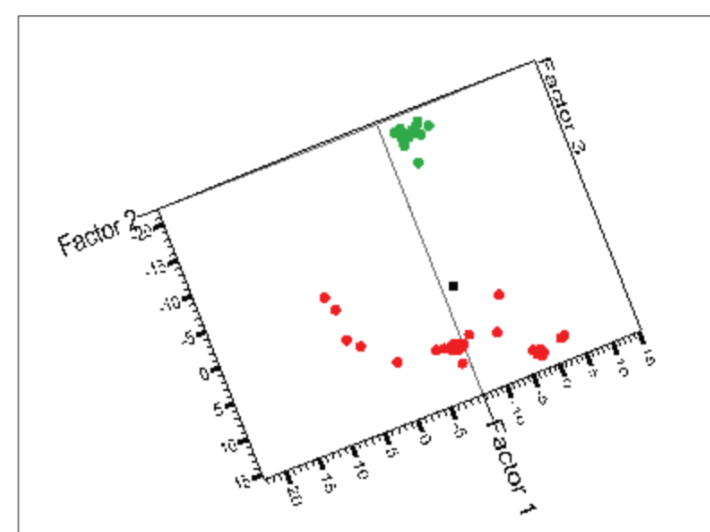


Figure 3. 3D PCA scores plot displaying Polycarbonate (●) Hits, Polysulfone (●) Hits, and Query (■).

Using Bio-Rad's patent-pending Overlap Density Heatmap technology, a comparative visualization of all the polycarbonate hit list spectra gives a graphical representation of the similarity and dissimilarity of this group of spectra. In the Overlap Density (OD) Heatmap (Figure 4) 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 polycarbonate 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 polysulfone hits display a high degree of overlap density (Figure 5).

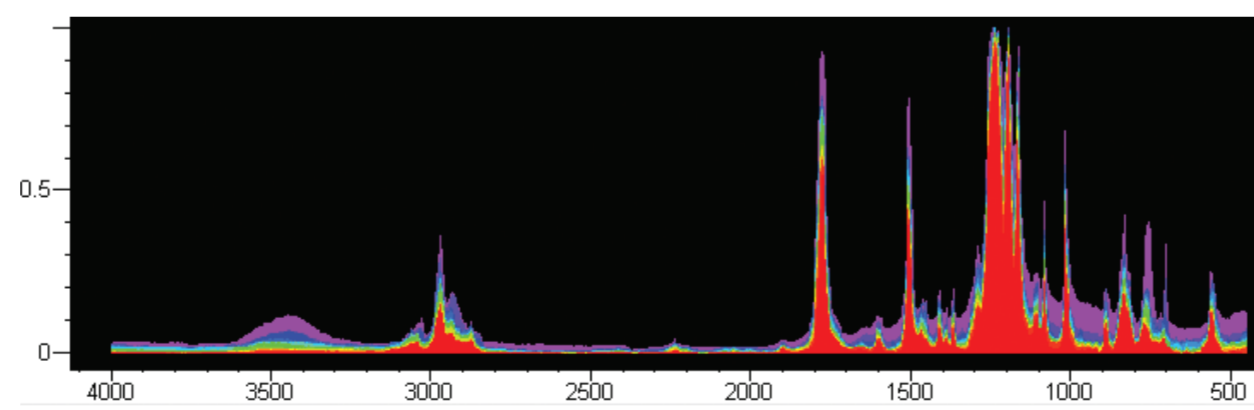


Figure 4. Overlap Density Heatmap of Polycarbonate Hits from the Monomers & Polymers (Comprehensive) Database (OD Level= 0).

Using the Overlap Density Heatmap, a consensus spectrum can be created. By tracing the outline of the highest level of overlap at a given OD level, it is possible to mathematically reconstruct a spectrum by using the maximum spectral Y-values at each spectral X-value in the OD Heatmap. This consensus spectrum is the visual representation of the spectral areas under the curve of the OD heatmap. The top part of the OD heatmap is "traced" and becomes the OD Consensus Spectrum. This Overlap Density Consensus Spectrum can then in turn be used as the spectrum in a spectral search query to find similar spectra in user or reference databases, as an entry to be stored in a database for future reference, or for reporting. The Consensus Spectra constructed from the scores plot confirms the presence of polycarbonate and polysulfone in the mixture after searching each against the Monomers and Polymers Database and confirming the components of the query (Figures 6 and 7).

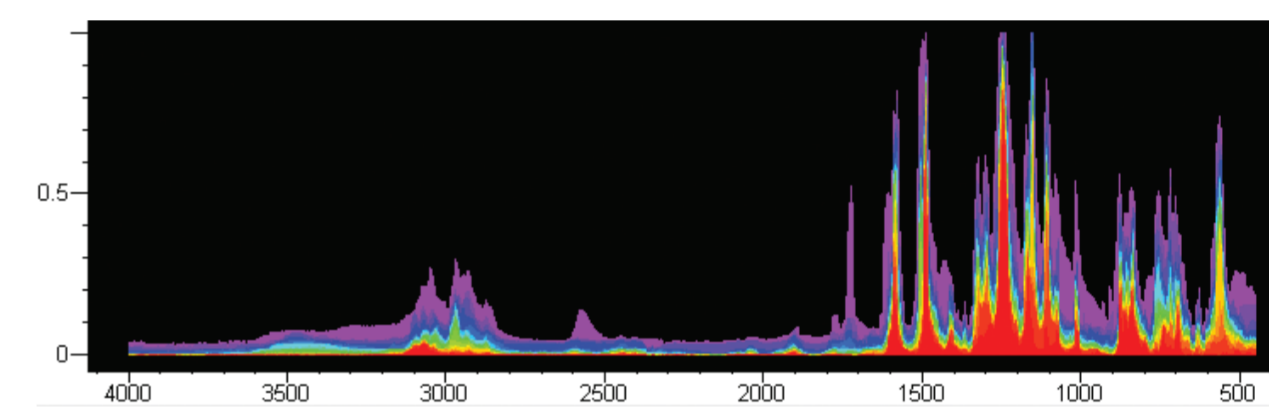


Figure 5. Overlap Density Heatmap of Polysulfone Hits from the Monomers & Polymers (Comprehensive) Database (OD Level = 0).

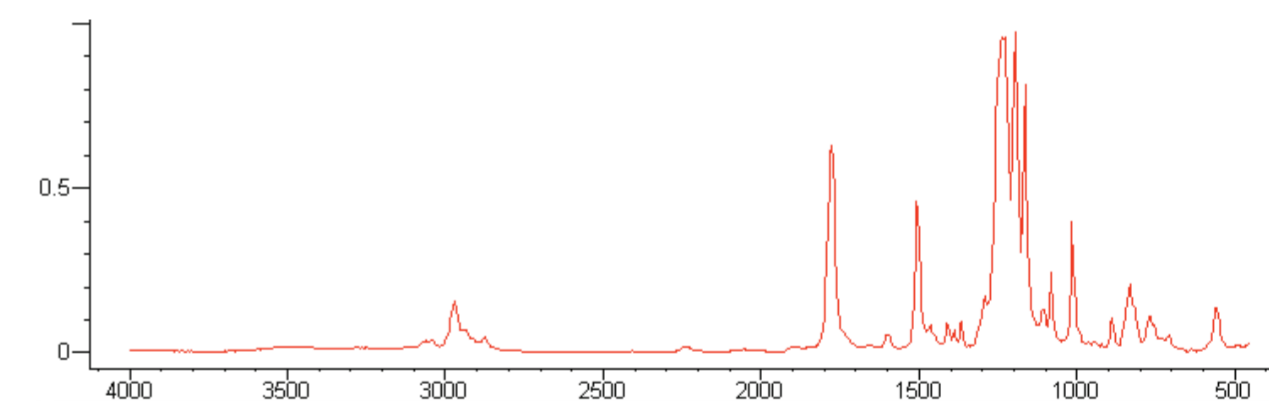


Figure 6. Consensus OD spectrum of 16 polycarbonate spectra (OD Level = 75 or 48%).

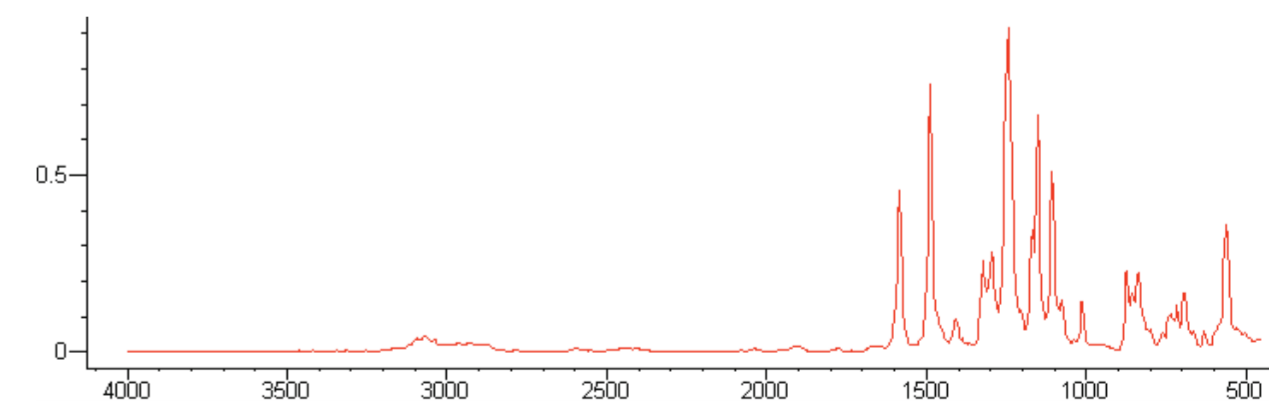


Figure 7. Consensus OD spectrum of 34 polysulfone spectra (OD Level = 75 or 27%).

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. This technique is an excellent tool to identify components in mixtures and can be used effectively in the polymer industry.