Multivariate Analysis of a Hit List from a Spectral Search of a Polymer Mixture

Gregory M. Banik, Ph.D. and Marie Scandone
[1] Bio-Rad Laboratories, Inc., Informatics Division, Philadelphia, PA, USA

Spectroscopy 95922

Abstract

This application note describes a new method that combines cheminformatics tools with chemometrics tools for Principal Component Analysis (PCA) in an intuitive environment for performing multivariate analyses on spectral as well as chromatographic data. A new patent-pending technology Overlap Density Heatmaps (ODHs) [1] in Bio-Rad's KnowItAll® software allows the comparative visualization of large datasets of spectra or chromatograms and are used for visual data mining and analysis to assess the similarities and dissimilarities in large amounts of spectral, chromatographic, and other graphical data.

Currently, the use of PCA to analyze and visualize spectral hit lists generated from searching one or more reference databases is not a widely known technique. However, as more case studies and applications are introduced, it is expected that this technique will rapidly become more commonplace in the laboratory. In this case study, this new approach for spectroscopic analysis will be examined as applied to polymeric IR data using a combination of PCA and ODH technologies to analyze a query and the hit list resulting from an IR spectral search and to perform an overall analysis of a database.

Methods

A polymer mixture was used as a query spectrum and searched against the Sadtler IR-Monomers & Polymers (Comprehensive) Database [2], which contains over 11,000 spectra. Using the SearchIt™ application in 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 KnowItAll's AnalyzeIt™ MVP [3], an application that is a joint development between Bio-Rad Laboratories and Infometrix, Inc., a leader in chemometrics.

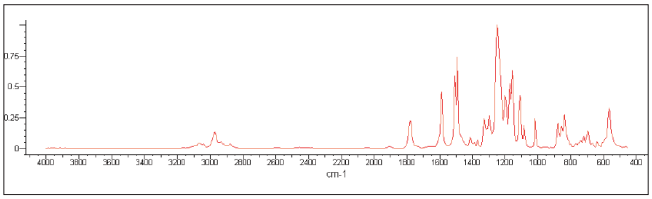

Fig. 1. Unknown Polymer Mixture

In AnalyzeIt 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 were 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 2n + 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 an approach to compensate 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 unknown query spectrum in Figure 1 resulted in a hit list of reference compounds comprised of 26 polysulfone spectra and 16 polycarbonate spectra plus a few miscellaneous compounds. PCA of the query and hit list spectra produced scores that show the spectra very nicely separated according to their type, as shown in the 2D Scores Plot (Figure 2) and the 3D Scores Plot (Figure 3). 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.

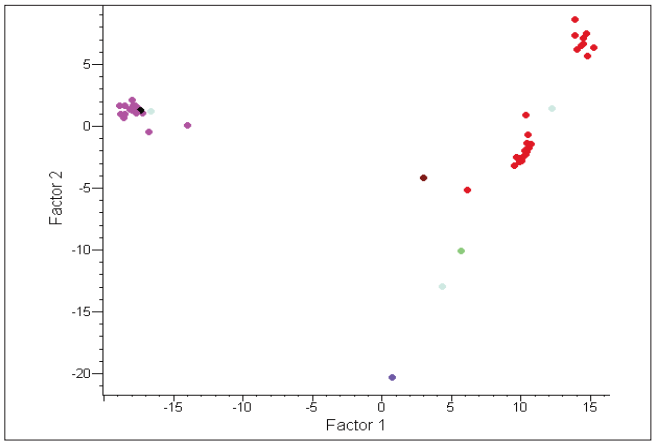

Fig. 2. 2D PCA Scores Plot ● - Query Spectrum; ● - Polycarbonate Hits; ● - Polysulfone Hits.

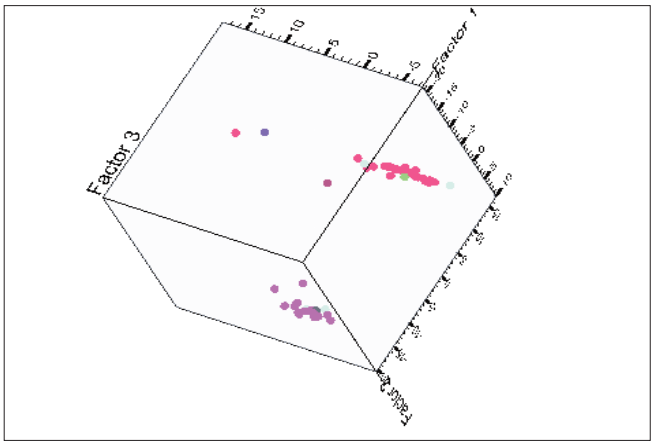

Fig. 3. 3D PCA Scores Plot ● - Query Spectrum; ● - Polycarbonate Hits; ● - Polysulfone Hits.

Using Bio-Rad's 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. The Overlap Density Heatmap of the query spectrum plus the polycarbonate hits shows the contribution of the polysulfone components in the query in purple (Figure 4). There is a peak at 1600 wavenumbers that can be contributed to the polysulfone in the query.

In the ODHM containing the polycarbonate hits (Figure 5) 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 ODH confirms that there is a high degree of similarity among these spectra: the heatmap presents an image that is predominantly red, indicating a high degree of commonality among the spectra (Figure 6).

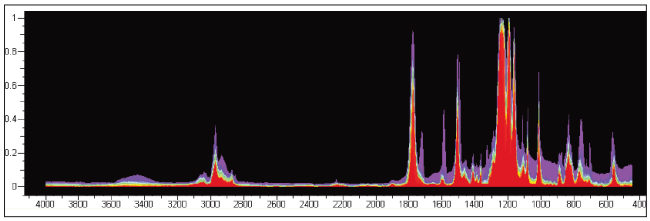

Fig. 4. Overlap Density Heatmap of query spectrum and the Polycarbonate hits from the Monomers & Polymers (Comprehensive) Database (OD Level= 0).

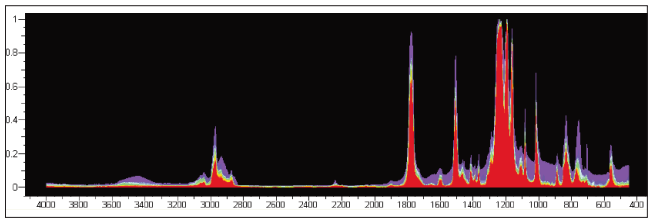

Fig. 5. Overlap Density Heatmap of Polycarbonate hits from the Monomers & Polymers (Comprehensive) Database (OD Level= 0).

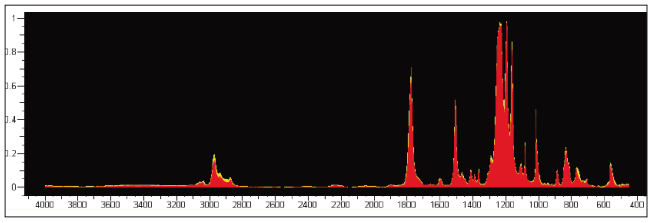

Fig. 6. Overlap Density Heatmap of Polycarbonate hits from the Monomers & Polymers (Comprehensive) Database (OD Level=75).

Similarly, the query spectrum plus the polysulfone spectra shows the difference contributed by the query spectrum containing polycarbonate (Figure 7). The peak that appears at 1800 wavenumbers can be contributed to the polycarbonate component of the query spectrum. The polysulfone hits display a high degree of overlap density (Figure 8). By selecting only the PCA scores corresponding to the polysulfone spectral hits, the corresponding ODH confirms that there is a high degree of similarity among these spectra: the heatmap presents an image that is predominantly red, indicating a high degree of commonality among the spectra (Figure 9).

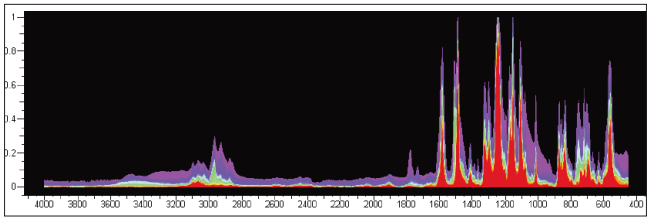

Fig. 7. Overlap Density Heatmap of query spectrum and the Polysulfone hits from the Monomers & Polymers (Comprehensive) Database (OD Level= 0).

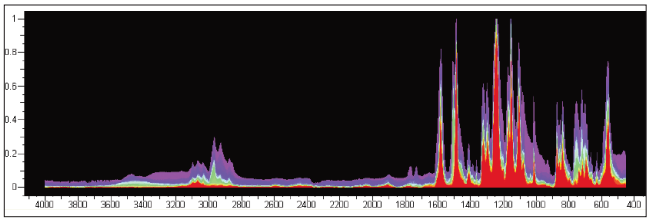

Figure 8. Overlap Density Heatmap of Polysulfone hits from the Monomers & Polymers (Comprehensive) Database (OD Level= 0).

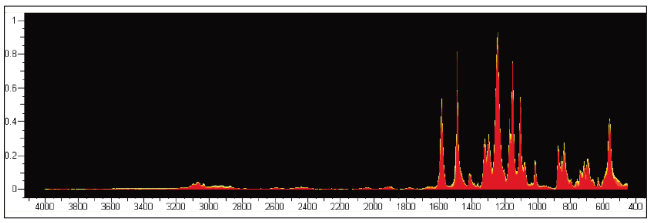

Fig. 9. Overlap Density Heatmap of Polysulfone hits from the Monomers & Polymers (Comprehensive) Database (OD Level=75).

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 ODH. This consensus spectrum is the visual representation of the spectral areas under the curve of the ODH. The top part of the heatmap is "traced" and becomes the Overlap Density consensus spectrum. This 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 Comprehensive Monomers and Polymers Database and confirming the components of the query (Figures 10 and 11).

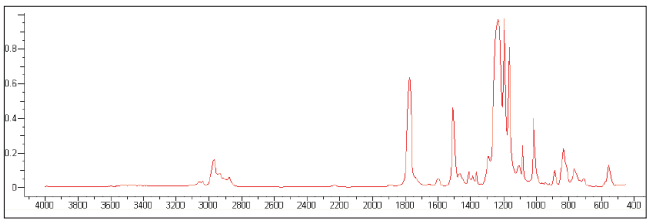

Fig. 10. Consensus spectrum of 16 polycarbonate spectra (OD Level =75, 48%).

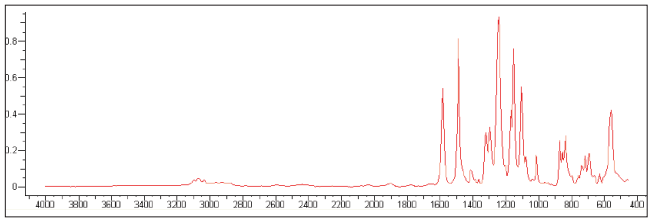

Fig. 11. Consensus spectrum of 26 Polysulfone spectra (OD Level = 75, 48%).

Viewing the ODH of all the hits generated from the initial search plus the query spectrum shows both similarity and dissimilarity among the components of the polymer mixture (Figure 12). An arbitrary scale (OD Scale) was created to define the extent of overlap for display in the objects being compared and displayed. When the OD Level = 0, it represents all levels of object overlap. When the OD Level = 100, it shows only the areas of highest overlap. When the OD Level = -100, it shows the unique areas or only the areas of lowest overlap.

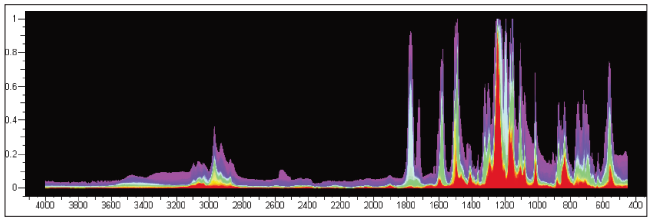

Fig. 12. Overlap Density Heatmap of all hits from the Monomers & Polymers (Comprehensive) Database plus the query spectrum (OD Level=0).

Conclusions

Principal Component Analysis (PCA) 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 Heatmaps (ODHs) 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 to analyze polymeric samples. It is more precise than spectral subtraction, which performs the point-by-point subtraction of one spectrum from another, and it is especially useful when analyzing mixtures or composite spectra.

References

[1] Overlap density heatmap application released as part of KnowItAll Informatics System 7.0, June 2006, Bio-Rad Laboratories.

[2] Sadtler "Monomers and Polymers Comprehensive Database" available as part of KnowItAll IR Spectral Library spectral database, Bio-Rad Laboratories.

[3] Principal component analysis performed with AnalyzeIt MVP KnowItAll application built using Infometrix Pirouette® IPAK technology.