Overlap Density Heatmaps: A Novel Method For Visualizing and Evaluating The (Dis)Similarity Among Multiple Spectra

Ty Abshear,[1] Omoshile Clement*,[1] Chen Peng,[1] Gregory Banik,[1] and Scott Ramos[2]

[1]Bio-Rad Laboratories, Inc., Informatics Division, 3316 Spring Garden Street, Philadelphia, PA 19104 USA
[2]Infometrix, Inc., Suite 250, 10634 E. Riverside Drive, Bothell, WA 98011 USA

Abstract

In this study, we introduce a novel tool, the Overlap Density Heatmap (ODH), for the visualization and quantitative evaluation of the (dis)similarity in massive amounts of spectral or chromatographic data. An ODH displays the common and unique features of overlapped objects through color coding areas depicting levels of overlap. By changing the OD scale, one can choose to display only those features of a certain level of commonality (ODC) or uniqueness (ODU), and one can generate their respective consensus spectrum.

ODHs can be used in a wide variety of applications, and has particular relevance to the quantitative evaluation of metabolomics spectral data. As an example, in a study of the ^{1}H NMR spectra of human serum samples from 37 diabetic and non-diabetic subjects, we generated OD consensus spectra at ~80% ODC of the normal and diabetic samples, and obtained a difference spectrum of the two, at an offset of 1.0. This difference spectrum identified diagnostic peak regions common to all diabetic or all normal patient samples. Using the difference spectrum as a search against the whole dataset revealed clear separation between both patient populations. Further, we then deployed the difference spectrum as a search query against a database of known metabolites. The hits retrieved contained metabolites with sugar moieties (largely monosaccharide and disaccharides), further supporting the versatility of applying the ODH technique to metabolite identification. In a separate metabolomics study, we deployed PCA technique to the analysis of the 37 biological samples. Good class separation between both patient populations was obtained. The observed loadings plot, highlighting important diagnostic peak positions which may characterize certain implicated metabolites, was compared to the difference spectrum from ODH and found to share many similar features. When used as search queries, both spectra (loadings plot and ODH-based difference spectrum) retrieved metabolites that favor sugar-based moieties (*vide infra*). These findings demonstrate that ODH can provide an unbiased approach to enhancing the multivariate analysis and interpretation of NMR-based metabolomics data.

Materials and Methods

Data Preparation

37 Proton NMR raw FIDs (Bruker) resulted from the analysis of human serum samples from diabetic and normal patients.[1] The 37 FIDs were batch processed using the macro function in the KnowItAll® ProcessIt™ NMR module and imported to a database in the KnowItAll MineIt™ module.[2]

Overlap Density Heatmap

The Overlap Density Heatmap (ODH)[2] is a new patent-pending technology from Bio-Rad that allows the visual examination and evaluation of spectral differences or commonalities. Compared to conventional overlay display of multiple spectra, OD heatmaps allow researchers to quickly identify the highly common areas (depicted in red) and the less common areas (depicted in violet) in each group, and hence provide a better technique to the overview of multiple spectra (see Figure 1). Moving the ODH slider to the right, displays the most common areas of the spectra (red), while moving it to the left will display the areas of highest uniqueness (violet).

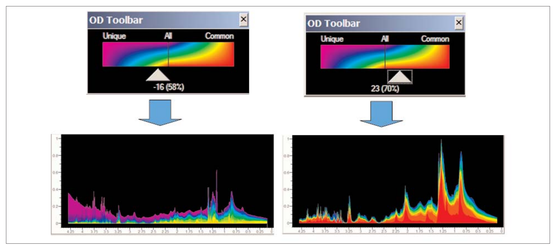

Figure 1 - ODH display of unique (left) and similar (right) spectral overlays

Principle Component Analysis

Principal Components Analysis (PCA) was performed with the KnowItAll AnalyzeIt™ MVP module.[2] The spectral regions of 4.5-0.5 ppm were used for the computation while excluding the strong water peaks (around 4.80 ppm) and other baseline regions. Prior to PCA, each spectrum was transformed by subtracting by its baseline value (the value of the 1st point in the region of 4.5-0.5 ppm) and dividing by sample 2-norm (i.e., vector length normalization). Mean centering and 3 PCs were used in pre-processing. Default setting of 0.04 ppm width in binning/bucketing was used.

Searching Metabolite Databases

The KnowItAll software can transform ODH spectral data as well as PCA loadings into a query spectrum that can be searched against a database of known metabolites. Peak searches, specific of spectrum area of high interest, can lead to the identification of changes in metabolite composition from one group to the other. Using these two approaches provide greater understanding of the utility of both tools as aid to metabolomics research. These are demonstrated in this study.

Results and Discussion

Application of ODH in Metabolomics Research

In the first example, Overlap Density Heatmap (ODH) was used to analyze ^{1}H NMR spectral data of blood samples from 23 normal and 14 diabetic patients. It was necessary to remove outliers from the spectral data, hence an OD level of ~15-20 representing ~80-85% of the areas under the curves (i.e., eliminating up to 20% outliers/noise) was used. Consensus spectra at this OD level of similarity were generated for the normal (23) as well as diabetic (14) samples. Next, a difference spectrum of the two consensus spectra was generated in order to identify diagnostic peak regions that may be important in each patient population (see Figure 2).

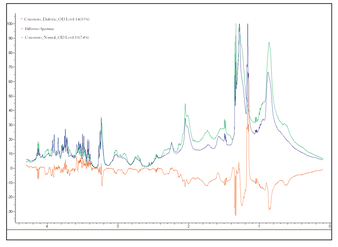

Figure 2 - ODH-based difference spectrum (^{1}H NMR) of diabetic and normal patient samples

The difference spectrum (Consensus_Diabetic-minus-Consensus_Normal) was tested against the full dataset in order to evaluate its ability to separately identify each patient population. Performing a Euclidean Distance spectral search on any of these peak areas (1.165 ± .1, 3.551 ± .1, 3.744 ± .1) with any random record from the dataset against the whole dataset yielded a clean 100% separation between the normal and diabetic classes. In other words, by picking a diabetic spectrum as the query, it is possible to retrieve all other diabetic spectra at the top of the hitlist (highest hit quality index (HQI)). Similarly, using a normal patient spectrum as the search query also retrieved all other normal spectra at the top of the hitlist. Finally, the difference spectrum with several diagnostic peak positions (1.153, 1.165 ppm, 3.53 ppm, 3.538 ppm and 3.585 ppm) was used as a search query against a 131-compound metabolite database in order to identify metabolites implicated in this disease type. A total of 50 metabolites were retrieved, many of which are sugar-based mono- and disaccharides, with the top-scoring hit being D-Fructose-6-phosphate. The results are shown in Figure 3.

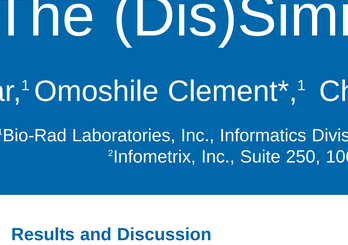

Figure 3 - Hitlist based on difference spectrum (Figure 3) search against a metabolite database

PCA Spectral Processing in Metabolomics Research

After processing the NMR FIDs as described in the Materials and Methods section, the PCA analysis results in the identification of two very well clustered groups of samples. The two clusters match the diabetic/normal patient samples (see Figure 4).

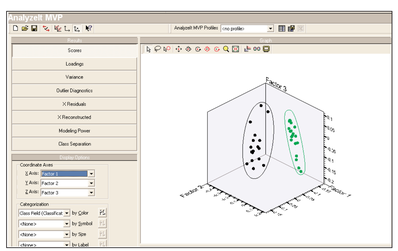

Figure 4 - PCA plot of spectral analysis of diabetic (green dots) and normal (black dots) samples

Identifying Spectrum Areas of Highest Variability

In combination with displaying the loadings from the PCA analysis, it is possible to very clearly identify those areas in the spectra that are most responsible for the variation between the two groups, thus providing useful insights for biomarker identification. Figure 5 displays a 2D loadings plot of spectral peaks vs PC2. It is clearly shown that spectral points at 1.153, 1.165 and 1.175 ppm (A), and between 3.52 - 3.59 ppm (B) contribute significantly to PC2.

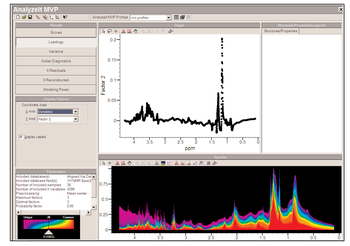

Figure 5 - 2D loadings plot highlighting diagnostic peak regions for potential implicated metabolites in the diabetic sample population

Correlating ODH to PCA in Metabolomics Study

We have demonstrated how a difference spectrum between OD display of diabetic and normal samples can yield diagnostic information about pertinent peak positions implicating certain metabolites undergoing changes biologically; we will now demonstrate how the OD display spectrum correlates with results obtained from chemometric analysis based on PCA for classifying differences between two states (e.g., diseased and non-diseased). As described earlier in the "Materials and Methods" section, PCA was applied to the 37 samples to yield well partitioned clusters representing each patient class (Figure 4). The 2D loadings plot derived from this analysis was compared to the ODH-based difference spectrum (see Figure 6). As shown in Figure 6, there is good correlation between both spectra derived from two different approaches. Furthermore, using the loadings plot as a search query with the diagnostic peak positions listed above against the metabolite database yielded 50 hits, as obtained for the search based on the ODH difference spectrum (see Figure 3). The hitlist retrieved from the PCA loadings plot also highlighted many sugar-based moieties, with D-Fructose-6-phosphate tied for highest scoring hit (Figure 7).

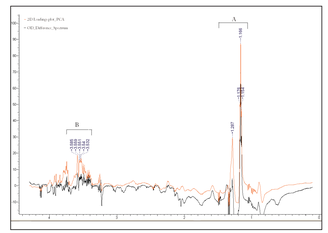

Figure 6 - Overlay of ODH-based difference spectrum and PCA-based loadings plot showing similarity in peak positions identified as important in the metabolomics study

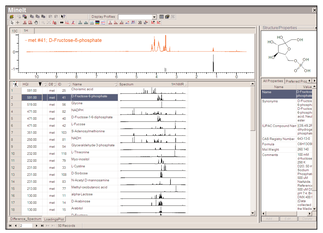

Figure 7 - Hitlist based on loadings plot (Figure 6) search against a metabolite database

Conclusions

We have shown that our new and patent-pending Overlay Heat Density (ODH) technique has wide applications in spectral visualization, analysis, and optionally metabolomics research. In combination with chemometric technique, the ODH technology offers a very powerful approach to the study of metabolite identification and characterization. Incorporation of Infometrix's chemometric program (Pirouette®) into the KnowItAll platform offers a fully-integrated NMR-based metabolomics investigation to researchers. Such an integrated platform opens the door to multi-technique metabolomics study which should provide more options for researchers in this area.

References

1. Data provided by Professor Bin Xia, Beijing NMR Center, Peking University, Beijing 1088971, China.
2. KnowItAll Release v7.5, scheduled for release in August 2006. Further information on the KnowItAll program can be obtained from our website at http://www.knowitall.com.
3. Biological Magnetic Resonance Data Bank (BMRB), a database of 1H and 13C NMR spectra of 131 metabolites, available from the University of Wisconsin, Milwaukee, USA. http://www.bmrb.wisc.edu/