

## Overlap Density Heatmaps: A Novel Approach to Metabolomics Research

Omoshile Clement\*,<sup>1</sup> Ty Abshear,<sup>1</sup> Chen Peng,<sup>1</sup> Gregory Banik,<sup>1</sup> and Scott Ramos.<sup>2</sup>

<sup>1</sup> Bio-Rad Laboratories, Inc., Informatics Division, 3316 Spring Garden Str., Philadelphia, PA 19104, USA.

<sup>2</sup> Infometrix, Inc., Suite 250, 10634 East Riverside Dr., 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. This new spectral visualization tool was applied to a study of the <sup>1</sup>H NMR spectra of human serum samples from 37 diabetic and non-diabetic subjects. OD consensus spectra of the normal and diabetic samples were generated, and a difference spectrum of the two was determined. This difference spectrum identified diagnostic peak regions which can distinguish diabetic from 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. Only D-glucose, a known biomarker for diabetes, was retrieved. The result illustrates the versatility of applying the ODH to identifying the requisite metabolite/biomarker for a disease type such as diabetes.

In a complementary study, we deployed Principal Component Analysis (PCA) to the analysis of the same 37 samples. Good class separation between both patient populations was obtained. The PCA loadings plot, which highlights important peak positions, and may also implicate metabolites, was compared to the difference spectrum from ODH and found to share many similar features. When used as a search query, the loadings plot spectrum retrieved a single hit, (D-glucose), from the metabolite database, similarly as was found for the ODH-based search query. These findings demonstrate that ODH can provide an unbiased approach to enhancing the multivariate analysis and interpretation of NMR-based metabolomics data.

### Introduction

A new spectral visualization and analysis tool - the Overlap Density Heatmap (ODH)<sup>1</sup> - is described in this study and applied to the study of biomarker identification in a set of diseased and control biofluid samples. 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.

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.<sup>2</sup> Peak searches, specific of spectral area of high interest, can lead to the identification of changes in metabolite composition from one group to the other. Using these two approaches promises greater understanding in areas such as metabolomics research, and their utility and complementarity are demonstrated in this study.

### Materials and Methods

#### Data Preparation

37 Proton NMR raw FIDs (Bruker) resulted from the analysis of human serum samples from diabetic and normal patients.<sup>1</sup> The 37 FIDs were batch processed using the macro function in the KnowItAll ProcessIt NMR module.<sup>3</sup> All spectra were acquired at 298K on a Bruker Avance-500 spectrometer. For each sample, 64 scans were collected into 8K complex data points with a spectral width of 8012.8 Hz.

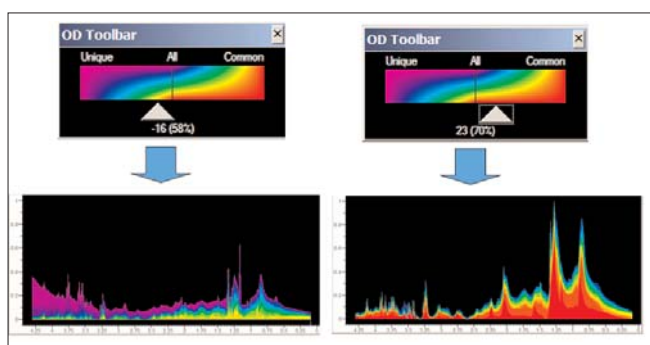
#### Spectrum Processing

The 37 FIDs were processed in the KnowItAll ProcessIt NMR module.<sup>1</sup> Processing parameters are presented in Figure 1 below. These parameters are recorded as a macro function and then applied to the whole set of FIDs in batch mode. The GoodLook™ autophasing algorithm, developed by Bio-Rad, is

a method that systematically optimizes the phase parameters until the integration of the peaks above the base line is the highest. This method works well for spectra with only positive peaks and a relatively flat baseline as is the case for the spectra in this study.

### Overlap Density Heatmaps

The Overlap Density Heatmap (ODH)<sup>1</sup> 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 common areas (depicted in red) and the uncommon areas (depicted in violet) in each group, and hence provide unique insight to the overview of multiple spectra (see Figure 1). By moving the ODH slider to the right, the user can select the more common areas of the spectra (red), while moving it to the left will display areas of greater uniqueness (violet).



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

### Principal Component Analysis

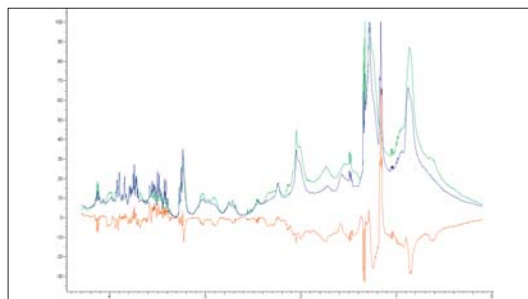
Principal Component Analysis (PCA) was performed with the Analyzelt™ MVP module<sup>1</sup>, a joint product between Bio-Rad and Infometrix incorporating Infometrix' Pirouette® technology for multivariate analysis. The spectral region of 10-0 ppm (excluding the solvent peaks at 4.5-5.0 ppm) were used for the computation. Prior to PCA, each spectrum was transformed by subtracting by a baseline value (10), and dividing by the value of the first point in the region of 10 - 0ppm, and dividing by the Sample 2-norm (i.e., vector normalization). Mean-centering was used in pre-processing. No binning/bucketing was used in the study.

## Results and Discussions

### Application of ODH in Metabolomics Research

The Overlap Density Heatmap (ODH) was used to analyze <sup>1</sup>H NMR spectral data of blood samples from 23 normal and 14 diabetic patients. To minimize the influence of outliers in the spectral data, an OD level of ~15-20, representing ~80-85% of the areas under the curves was used. Consensus spectra at this OD level of similarity were generated for normal (23) as well as diabetic (14) samples. Next, a difference spectrum of the two consensus spectra was generated which may be important for distinguishing the two patient populations (see Figure 2).

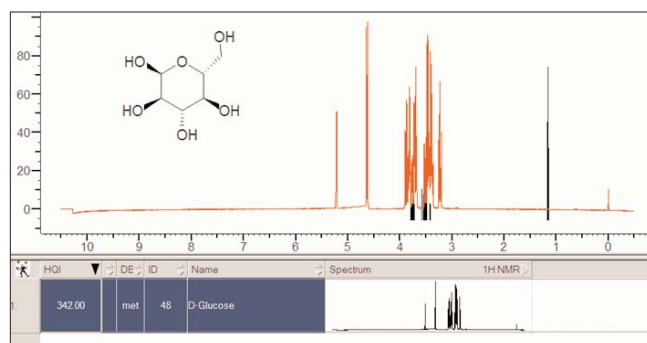
The difference spectrum (Consensus\_Diabetic-minus-Consensus\_Normal) was tested against the full dataset in order to evaluate its ability to separately identify each patient population. Performing a Euclidean Distance spectral search on



**Figure 2.** ODH-based difference (red) spectrum (<sup>1</sup>H NMR) of diabetic (black) and normal (green) patient samples.

any of these peak areas ( $1.165 \pm .1$ ,  $3.55 \pm .1$ ,  $3.74 \pm .1$ ) 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.

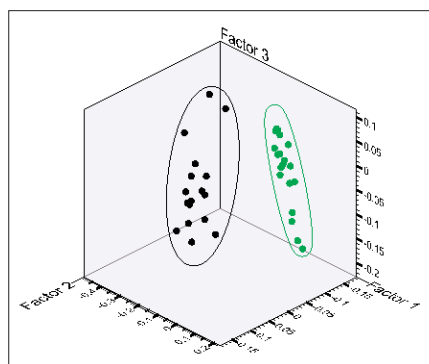
Finally, the difference spectrum with several diagnostic peak positions was used as a search query against a 131-compound metabolite database in order to identify metabolites implicated in this disease type. One hit was retrieved - D-glucose (see Figure 3), suggesting that this approach yield results that correctly correlates the known biomarker for diabetes - glucose.



**Figure 3.** D-Glucose retrieved as hit using an OD consensus spectrum (difference spectrum, see Figure 2) search of a metabolite database for biomarker identification.

### PCA Spectral Processing in Metabolomics Research

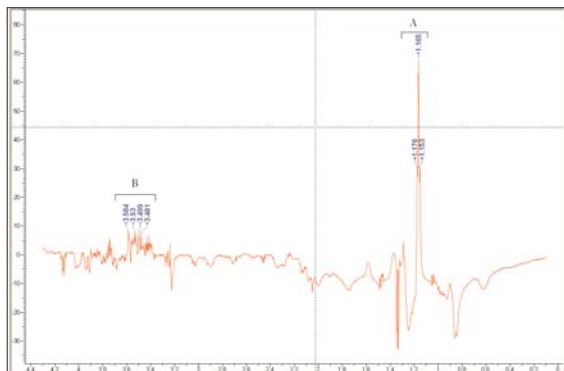
After processing the NMR FIDs as described in the "Materials and Methods" section the PCA analysis results in two very well clustered groups of samples. The two clusters correspond to the diabetic and normal patient samples (see Figure 4).



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

## Identifying Spectrum Areas of Highest Variability

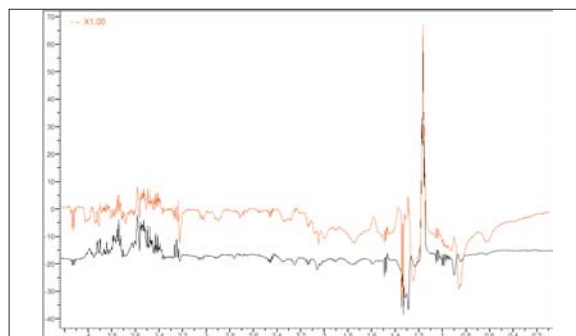
In combination with displaying the loadings from the PCA analysis, it is possible to very clearly identify those areas in the spectra that are most responsible for the variation between the two groups, thus providing useful insights for biomarker identification. Figure 5 shows a 2D loadings plot of spectra 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 of spectral peaks vs. PC2 showing regions diagnostic for biomarker identification.

## Correlating ODH to PCA in Metabolomics Study

We have demonstrated how a difference spectrum generated from an OD heatmap of diabetic and normal samples can yield diagnostic information about pertinent peak positions, thus implicating possible metabolites. We have also shown how such an OD-derived spectrum can correlate with results obtained from chemometric analysis, based on PCA, for characterizing differences between two states (e.g., diseased and non-diseased). PCA yielded well-separated clusters representing each patient class (Figure 4). The loadings derived from this analysis were compared to the ODH-based difference spectrum (see Figure 6) and good correlation was found between both profiles. Furthermore, search queries using diagnostic peak positions found in the PCA loadings yielded a single hit - D-glucose as found for the study based on ODH consensus spectrum (see Figure 3).



**Figure 6.** Overlay of ODH-based difference spectrum and PCA-based loadings plot showing similarity in peak positions.

## Conclusions

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

## References

1. These tools were available in the KnowItAll Release 7.0 (June 2006). Further information on the KnowItAll program can be obtained from our website at <http://www.knowitall.com>.
2. Biological Magnetic Resonance Data Bank (BMRB), a database of <sup>1</sup>H and <sup>13</sup>C NMR spectra of 131 metabolites, available from the University of Wisconsin, Milwaukee, USA. <http://www.bmrb.wisc.edu/>
3. Data provided by Professor Bin Xia, Beijing NMR Center, Peking University, Beijing 1088971, China.

**BIO-RAD**

**Bio-Rad  
Laboratories, Inc.**

**Informatics Division**  
[www.knowitall.com](http://www.knowitall.com)

**China**

Phone: +1 215 382 7800 • E-mail: [informatics.china@bio-rad.com](mailto:informatics.china@bio-rad.com)

**Europe**

Phone: +44 20 8328 2555 • E-mail: [informatics.europe@bio-rad.com](mailto:informatics.europe@bio-rad.com)

**Japan**

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

**Rest of World**

Phone: +1 215 382 7800 • E-mail: [informatics.row@bio-rad.com](mailto: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](mailto:Informatics.usa@bio-rad.com)