application note

Toward Diagnosis of Diabetes by NMR and Multivariate Analysis

Gregory M. Banik, Ph.D. Bio-Rad Informatics,
Two Penn Center Plaza, Suite 800, 1500 John F. Kennedy Blvd.,
Philadelphia, PA 19102-1737   USA

Tao Wang, Bin Xia, Ph.D., Beijing NMR Center,
Peking University,
Beijing, 100871, China

Scott Ramos, Infometrix, Inc.
Suite 250, 10634 E. Riverside Dr.,
Bothell, WA 98011

Abstract

Nuclear magnetic resonance spectroscopy (NMR) is becoming a key tool for understanding the metabolic processes in living systems. Among its many applications, advanced spectroscopic techniques are combined with multivariate statistical approaches to provide diagnostic information for diseases and to identify the changes in the metabolic pathways [1,2]. This study demonstrates the potentials of this approach by the multivariate analysis of the ^{1}H NMR spectra of serum samples from diabetic and healthy people.

Method

Thirty-seven blood samples were collected from seventeen diabetic patients and twenty healthy people; then they were allowed to clot in plastic tubes for about two hours at room temperature. Aliquots of serum were collected from the blood and stored at -80°C until assayed.

Before the NMR experiment, each sample (150µl) was diluted with solvent solution (300µl H_2O, 50ul D_2O and 3µl DSS). All spectra were measured at a temperature of 298K on a BRUKER Avance-500 spectrometer operating at the proton frequency of 500.13 MHz using pulse sequence ZGPR (RD-90-t_1-90- acquisition, RD being a relaxation delay of 1.5s during which the water resonance is selectively irradiated). For each sample, 64 scans were collected into 8K complex data points with a spectral width of 8012.8Hz.

The whole data analysis process, from NMR spectral processing to principal component analysis to metabolite database search, were all done using several of the integrated application modules in the KnowItAll® Informatics System (Metabolomics Edition).

The raw spectra (FIDs) were automatically processed and saved into a database using the macro-based batch processing function of the KnowItAll® Informatics System ProcessIt™ NMR

and MineIt™ applications. The macro included correction of DC offset, zero-filling to 16,192 points, Lorentizan apodization of 0.2 Hz, Fourier Transform, phase correction, baseline correction with automatic base point detection and Spline fitting, and referencing (with the same data point at the DSS peak set to 0 ppm). After the batch processing, some of the imperfectly phased spectra were re-phased and baseline corrected again manually and saved back to the database. Status information (normal or diabetic) was added manually for each sample into the database.

The principal component analysis (PCA) was run with the AnalyzeIt™ MVP application. The spectral regions of 10-5.15 ppm and 4.75 - 0.5 ppm were used for the computation in order to exclude the strong water peaks and other baseline regions. Prior to PCA, each spectrum was transformed by subtracting by its baseline value (the value of the first point in the region of 10-5.15 ppm) and dividing by sample 2-norm (i.e., vector length normalization). Mean-centering was used in pre-processing.

While we mainly used the spectral datapoints as input to PCA analysis, we also compared the results using a conventional way of binning/bucketing (either with a fixed width of 0.04 ppm per bin or using the IntelliBucket™ method where the bin width is automatically adjusted based on the Overlap Density Heatmap (ODH) consensus spectrum). The integral of the bins were used as their Y-values, and they are subtracted by the integral of the first bin (the one closest to 10 ppm) to correct the baseline; next, the bins were scaled by dividing by sample 2-norm. Mean-centering was used in pre-processing.

The post-PCA analysis was done using the SearchIt™ and ProcessIt™ NMR applications. A ^{1}H and ^{13}C NMR spectral library of over 225 standard metabolites [3] were searched to identify changed metabolites using queries such as a loading plot and a difference spectrum between the OD consensus spectra of the individual classes.

Discussion

The great variability among the spectra can be seen using the overlay display of some of the spectra (Figure 1). The strong water peaks at about 4.8 ppm make the spectral range between 5.15 to 4.75 ppm inaccessible and is hence excluded from the analysis. It is also noted these spectra showed little local peak shifts, except for the water and DSS peaks. By aligning the spectral globally using the peaks between 4-3 ppm, or by simply setting the reference to a certain data point (instead of to the top of the DSS peak), it effectively eliminated the cross spectra misalignment.
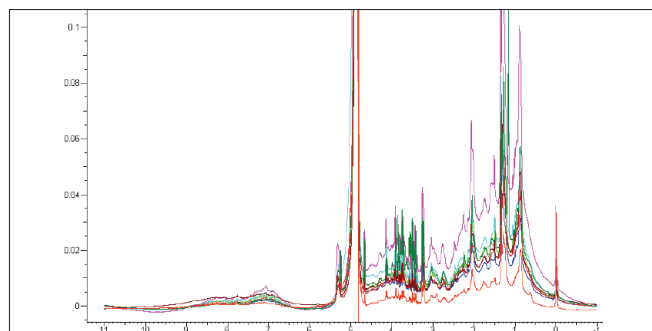

Figure 1. *Overlay of eight metabolomics 1H NMR spectra.*

Subsequent PCA analysis was focused on the regions of 10 - 5.15 ppm and 4.75 - 0.5 ppm. First, we used all 37 samples. The PCA scores (PC1 vs. PC2) are shown in Figure 2 where clear distinctions between the diabetic and normal samples are evident.
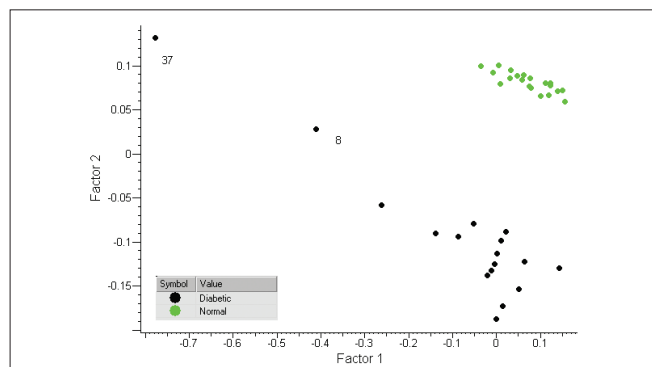

Figure 2. *PCA scores of the 37 1H NMR spectra of the serum samples of diabetic (black) and non-diabetic (green) patients. From this plot it is evident that samples 8 and 37 are outliers.*

From this plot it is evident that the samples 8 and 37 are outliers. After removing both of these samples from the PCA, we generated the following scores plot (Figure 3).
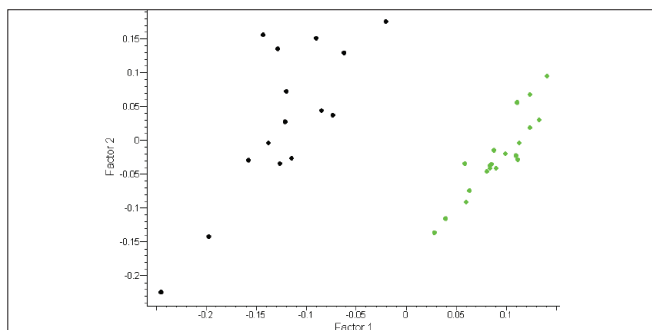

Figure 3. *PCA scores of the 35 1H NMR spectra of the serum samples from diabetic (black) and non-diabetic (green) patients after samples 8 and 37 were excluded as outliers. Similar to Figure 1, PC1 and PC2 provides a clear differentiation between the two groups of samples.*

The loadings plots of PC1 (Figure 4) provides insight to the main spectral differences that lead to the differentiation of the two groups. Note that such a loadings plot has a point-to-point correspondence to the actual spectral data points, and hence is convenient to compare with the original spectra directly. (See further discussion regarding Figure 6.) Furthermore, it is used as a query spectrum and the peaks are searched against a library of over 225 common metabolites. The top hit, with peaks around 4-3.4 ppm matched, was D-glucose (Figure 5).
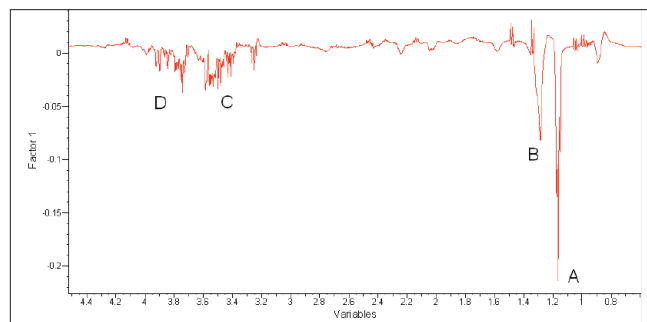

Figure 4. *PCA loadings of PC1, which shows that spectral points around 1.18 (A) and 1.30 ppm (B) contribute most significantly to the first principal component. Peaks between 3.37-3.68 ppm (C) and those between 3.71-4.04 ppm (D) also contribute significantly to PC1. The peaks in the loadings plot are searched against a spectral library of 225 metabolites and D-glucose is returned as the top hit.*
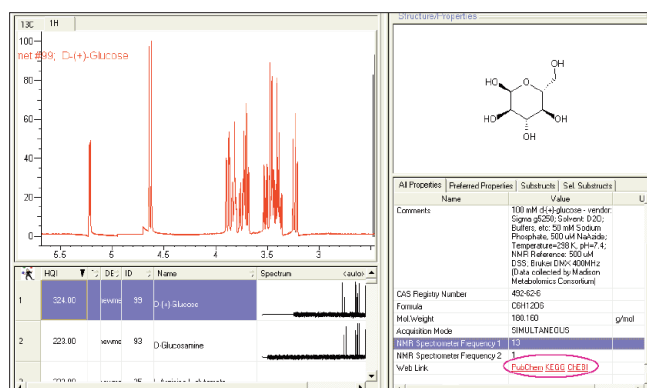

Figure 5. *The loading plot of PC1 (Fig. 4) is searched against a spectral library of 225 common metabolites (tolerance = 0.01 ppm, minimum number of peaks to matach = 8), and D-Glucose is returned as the top hit. Note that from the Property Pane, links are provided for browsing other properties of the hit, such as its relevant metabolic pathways from the KEGG database.*

Other than the conventional loading plots, the Overlap Density Heatmap display functions of the KnowItAll® Informatics System provide novel approaches to evaluate spectral differences and, conversely, similarities. Figure 6 shows the ODHs of all spectra in both groups (Above: normal; below: diabetic). Compared to conventional overlay displays of multiple spectra, the OD heatmap allows one to quickly identify the highly common areas (in red) and less common areas (in violet) within each group of spectra being examined, and hence provide a better technique to analyze multiple spectra at once.
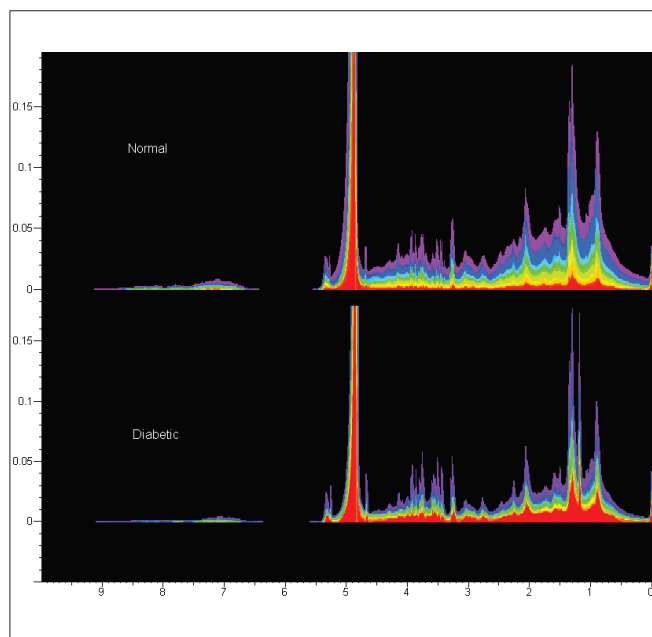
Figure 6. *Overlap Density heatmap of spectra in both groups (Above: normal; below: diabetic). From red to violet are spectral areas with high to low levels of overlap.*

Figure 7 shows the OD heatmap consensus spectra (a mathematically reconstructed spectrum created from the maximum spectral y-values at each spectral x-value from a heatmap) retaining 80% common features among the spectra of each group. The black and red curves correspond to the normal and diabetic groups, respectively. Comparing the normal curve to the diabetic curve, it is easy to see that the most significant difference between the two groups is the newly emerged peaks between 1.19-1.14 ppm (A), and the new peaks between 3.60-3.51 ppm (B) and 3.80-3.73 ppm (C). These observations agree with those from the loading plots. The difference spectrum is generated between the two OD consensus spectra and is searched against the 225 metabolites, and again D-glucose is returned as one of the top hits.
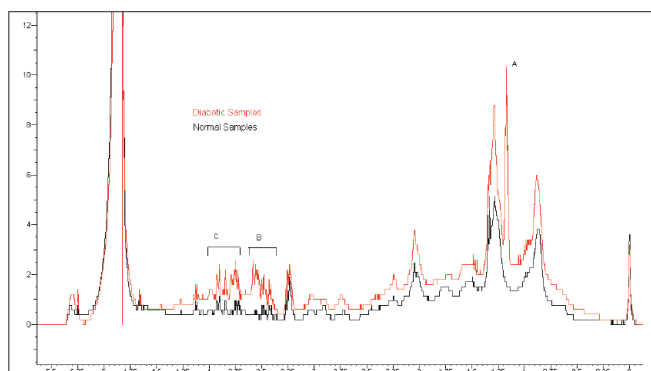

Figure 7. *Overlap Density heatmap profiles showing 80% of the common features of each group (Black: the normal group. Red: the diabetic group). Comparing the two profiles allows one to easily identify the major differences between the groups. The difference spectrum (no shown) is searched against a spectral library of 225 metabolites and D-glucose is returned as the top hit.*

Based on the published peak assignments of human serum spectrum [4], these changed chemical shifts can be assigned to the CH_3 groups (0.91 ppm) and $(CH_2)_n$ groups (1.26 and 1.30 ppm) of the fatty acid side chains in lipids, in particular LDL and VLDL; 3-hydroxy butyrate isobutyrate (1.18 ppm),

lactate (1.34 and 1.38 ppm) and sugar, glycerol, and amino acid CH (corresponding to the region near 3.5ppm).

In most of the published applications, the spectra were sub-sampled into bins usually with a width of 0.04 ppm. To compare the results with or without binning, we repeated the previous experiment with an IntelliBucket™ algorithm, which first divides the spectra into bins with binning in the same spectral regions using a fixed width of 0.04 ppm, and then optimizes the boundaries of the bins within a variation range of ± 0.02 ppm. Referring to the OD consensus spectrum of 80% commonality of all the spectra, the edges of a bin are adjusted so that the nearby local minimum, if any, is used. This produced 235 bins with various bin width. As demonstrated in Figure 8, the scores plots show very similar clustering of the samples as shown in Figure 3, where binning was not done. However, the loadings plot of PC1 (Figure 9) is of much lower resolution than the one without binning (Figure 4), and hence it is harder to map the bins to the changed chemical shifts.
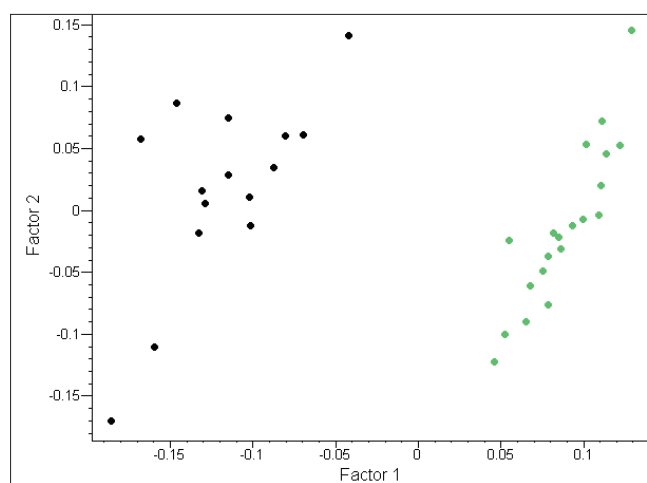
Figure 8. *PCA scores of the 35 ^{1}H NMR spectra of the serum samples from diabetic (black) and non-diabetic (green) patients after samples 8 and 37 were excluded as outliers. Prior to the analysis, the spectral regions were divided into 235 bins of width of 0.04 ppm.*
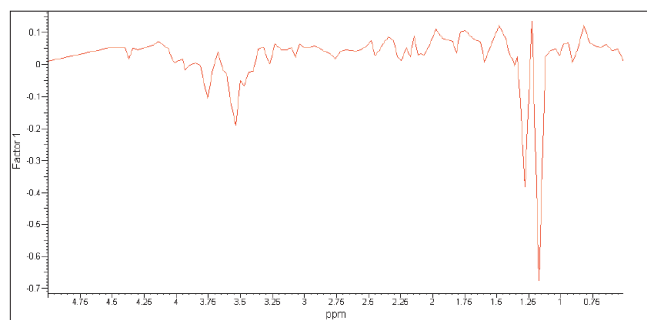
Figure 9. *PCA loadings of PC1 from the analysis with binning. Compared to Figure 4, this loadings plot has a much lower resolution and gives less information about the changed chemical shifts.*

Conclusions

This technical note demonstrates the potentials of applying NMR-based metabolomics in disease diagnostics and in order to identify the changes of metabolic pathways. While further elaboration can be made in the analyses from this initial study in the future and continued improvements will be made in future generations of the software, the following conclusions can be currently drawn:

1. The macro-based batch processing of metabolomics NMR spectra using KnowItAll® ProcessIt™ NMR and MineIt™ significantly improves the efficiency of spectral processing and management, which is usually both tedious and time consuming.
2. Principal Component Analysis (PCA) of the metabolomics NMR data of the serum samples using KnowItAll® Informatics System's AnalyzeIt™ MVP application provides a reliable way to diagnose diabetes.
3. The traditional PCA loadings plots and the novel OD heatmap profiles generated using the KnowItAll® platform provide different approaches to identifying the differences between the spectra, which opens the door to further identifying key metabolites or biomarkers.
4. Searching a spectral library of common metabolites provides a helpful method for identifying changed metabolites.
5. When spectral misalignment is not serious, it is preferable to run the PCA analysis of the metabolomics NMR data at the datapoint resolution, rather than using the traditional binning and bucketing.

References

1. F.F., Brown, I.D. Campbell, P. W. Kuchel, D. L. Rabenstein, Human erythrocyte metabolism studies by ^{1}H spin echo NMR. *FEBS Lett*. **82** (1977), 12-16.
2. J. K. Nicholson, L. C. Lindon, C. E. Holmes, 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, **29** (1999), 1181-1189.
3. The raw spectral data were adapted from Biological Magnetic Resonance Data Bank at the University of Wisconsin at Madison, http://www.bmrb.wisc.edu/.
4. K. S. Solanky, N. J. C. Bailey, B. M. Beckwith-Hall, A. Davis, S. Bingham, E. Holmes, J. K. Nicholson, A. Cassidy, Application of biofluid ^{1}H nuclear magnetic resonance-based metabonomics techniques for the analysis of the biochemical effects of dietary isoflavones on human plasma profile. *Anal. Biochem*., **323** (2003), 197-204.

BIO-RAD

Bio-Rad Laboratories, Inc.