Toward Diagnosis of Diabetes by NMR and Mulitvariate Analysis: Study of ^{1}H NMR Spectra of Human Serum in an Integrated Working Environment

Chen Peng,[1] Omoshile Clement,[1] Gregory Banik*,[1] Scott Ramos,[2] Tao Wang[3], and Bin Xia[3]

[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
[3] Beijing NMR Center, Peking University, Beijing 1088971, China

Abstract

Data analysis is one of the key steps involved in any metabolomics studies, and processing of NMR data is especially tedious and time-consuming since it involves multiple steps to transform and correct the data starting from the raw FIDs. In order to improve the efficiency of NMR-based metabolomics studies, we have developed KnowItAll® Informatics System, Metabolomics Edition as a consistent and integrated software environment that covers the entire process from processing raw NMR data —to multivariate analysis—to biomarker identification. Using two datasets of ^{1}H NMR spectra of mouse urine and human serum, we demonstrate the high efficiency of this integrated workflow for metabolomics data analysis (Figure 1).

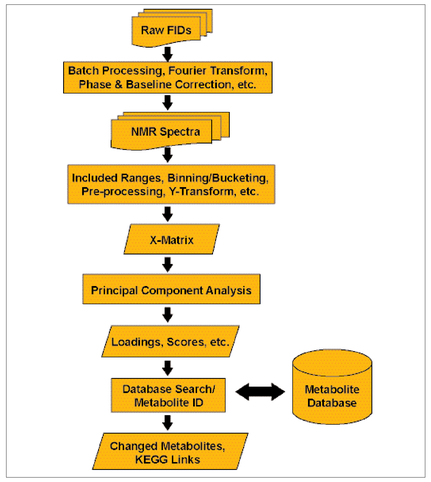

Figure 1. *The workflow of KnowItAll Metabolomics Edition for NMR-based metabolomics data analysis.*

Materials and Methods

Software Tools for Data Analysis

The data analysis was accomplished using KnowItAll version 7.5 on a PC Windows XP workstation. Several integrated applications in the KnowItAll environment were used, including ProcessIt™ NMR, MineIt™, AnalyzeIt™ MVP, and SearchIt™. Data transfer between the different applications was accomplished by simply clicking the "Transfer to" Bar in KnowItAll.

Dataset I: Diagnosis Studies of Diabetes

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^{0}C until assayed. Before the NMR experiment, each sample (150µl) was diluted with solvent solution (300µl H_2O, 50µl D_2O and 3µl DSS). All spectra were measured at a temperature of 298K on a Bruker Avance spectrometer operating at the proton frequency of 500.13 MHz. For each sample, 64 scans were collected into 8K complex data points with a spectral width of 8,012.8 Hz.

Dataset II: Mouse Hepatotoxicity Studies

Standardized rats were treated with low and high dose of hydrazine. Urine samples were collected at 24-hour intervals through 7 days, for 8 replicates each of 3 groups: controls, low dose and high dose. ^{1}H-NMR was run on urine of 192 samples on a Bruker spectrometer at 600.13 MHz and each processed spectrum contains 16K points.

Results and Discussion

Dataset I

The 37 raw FIDs were processed using the macro and batch processing functions of ProcessIt™ NMR and MineIt™ modules in the KnowItAll system. The macro contains commands for correction of DC offset, zero-filling, Lorentzian apodization, Fourier Transform, automatic phase correction, baseline correction, and reference setting. The processed spectra were automatically imported into a database, and the sample properties were manually added. The spectra were inspected and manually adjusted when necessary.

The PCA was done using the AnalyzeIt™ MVP module in the KnowItAll system. The spectral ranges of 10-5.15 ppm and 4.7-0.25 ppm were used. Baseline subtraction and vector length normalization were selected for Y-transform, and mean-centering was used for pre-processing. No binning was done. The first run showed that samples #8 and #37 were outliers and they were excluded from the second run, which led to the two clearly separated clusters that matched the samples' diabetic/non-diabetic origin (Figure 2).

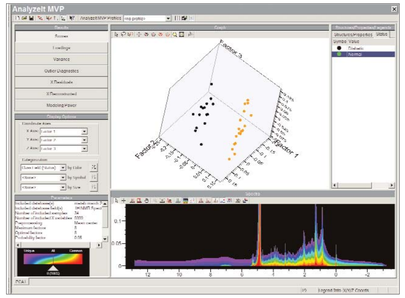

Figure 2*. The scores plot that shows a clean separation between samples of diabetic (black) and non-diabetic (orange - because they are currently selected) subjects and the ODHM of the selected non-diabetic spectra.*

Several methods were used to identify spectral areas that are most responsible for the variation between the two groups. As shown in Figure 3, by displaying the PC1 loadings plot from the PCA analysis along with the Overlap Density Heatmap (ODHM) of the original spectra, we identified the spectral points around 1.18 (A) and 1.30 ppm (B) that 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.

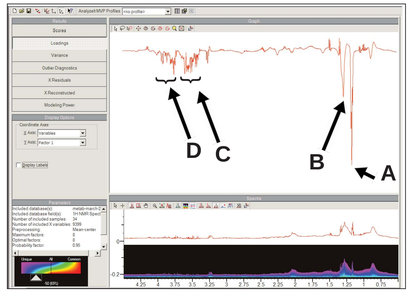

Figure 3. *PCA Loadings, Overlap Density Heatmap and OD consensus spectrum for visual identification of changed NMR peaks in the diabetic dataset.*

For each category of spectra, diabetic and non-diabetic, it is possible to generate a consensus spectrum with the Overlap Density Heatmap set at high commonality (e.g. OD level = 80). These two consensus spectra are then overlaid to better identify the spectral areas of highest variation between the two groups. For example, the peaks centered at 1.18 ppm, (See Figure 4) are unique to the diabetic group. They are located in the aliphatic area of the spectrum and may reflect changes in fatty acid composition.

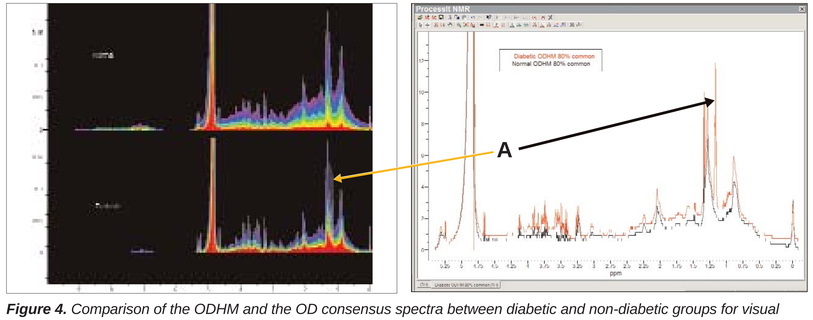

Figure 4. *Comparison of the ODHM and the OD consensus spectra between diabetic and non-diabetic groups for visual identification of changed NMR features.*

Finally the PC1 loadings plot was transferred to the SearchIt™ module as a query spectrum, and the peaks were picked and searched against a spectral database of 226 common metabolites (The ^{1}H and ^{13}C spectra were adapted from the Biological Magnetic Resonance Data Bank at University of Wisconsin, Madison). As illustrated in Figure 5, D-glucose and its derivatives are reported among the top hits. Each hit is provided with a link to the KEGG database, providing additional information on the metabolite composition and participation to metabolite pathways.

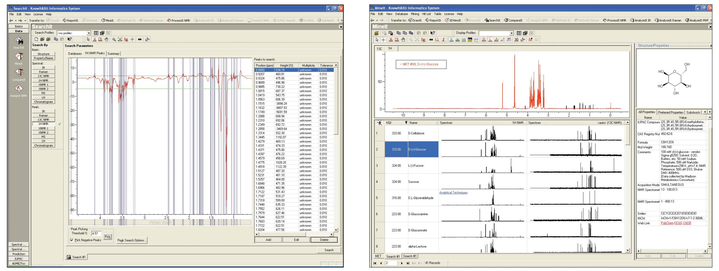

Figure 5. *The loadings plot along PC1 is tranformed into a query spectrum and its peaks are searched against a database of 226 common metabolites. D-glucose and its derivatives are reported among the top hits. Each hit is provided with a link to the KEGG.*

Dataset II

The frequency-domain spectra were batch imported to MineIt™ via a JCAMP filter. The full spectral resolution was retained (16K) without binning but only 10-6.2 ppm and 4.5-0.5 ppm regions were included to avoid water and urea regions. The spectra were vector length normalized and mean-centering was used as pre-processing method. Figure 6 shows the scores plots from the PCA of the 190 spectra (2 outliers excluded). It shows that samples treated with the higher dose diverge in a common direction, most control and low dose samples cluster in a single group, and a few low dose samples diverge in a direction very different from the high dose samples. In order to observe their position in scores space as function of time, we transferred the high dose spectra into MineIt™ as a hit list, and repeated the PCA on these 62 spectra using exactly the same parameters as for the whole dataset. In the resulting 3D scores plot (Figure 7), samples are colored by their corresponding time point (in hours), and the samples closest to the center of each group are connected to show the trajectory over time.

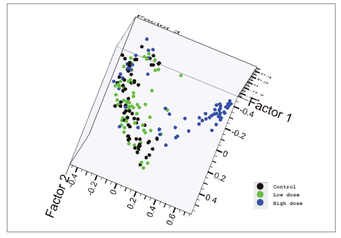

Figure 6*. PCA scores plot of the hepatotoxicity dataset. It shows that samples treated with the higher dose (blue) diverge in a common direction, most control (black) and low dose (green) samples cluster in a single group, and a few low dose samples diverge in a direction very different from the high dose samples.*

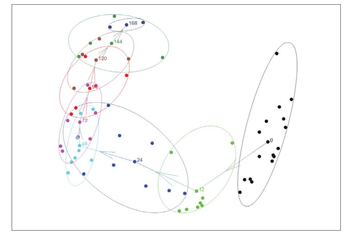

Figure 7*. The PCA scores plot of the high dose samples and the trajectory of the groups over time (0-168 hours after the dose). The ellipses show the dispersion of each group of samples in the space.*

Conclusion

The KnowItAll Informatics System, Metabolomics Edition provides valuable tools that are useful to almost every aspect of the data analysis involved in metabolomics studies. Among them, the batch processing of NMR, ODHM for multiple spectral visualization and analysis, the fully integrated pre-processing and PCA and the post-PCA data interpretation tools are especially useful. It is demonstrated that the data interpretation for metabolomics studies can be more efficient in this seamlessly integrated working environment of the KnowItAll system.

The two case studies show that combining the capabilities of NMR processing and database management with principal component analysis gives the analyst the capability of evaluating trends in NMR quickly and efficiently.

- PCA of the metabolomics data of the serum samples provides a reliable way to diagnose diabetes.
- PCA of the hepatotoxicity data of the urine samples illustrates an efficient means of identifying underlying biochemical events. Study of toxins or drugs in development can benefit substantially.

Acknowledgement

The hepatotoxicity data were kindly offered by Dr. Andy Nicholls, GlaxoSmithKline, London

References

1. Anthony, M.L.; Beddell, C.R.; Lindon, J.C. and Nicholson, J.K.; Studies on the comparative toxicity of S-(1,2-dichlorovinyl)-L-cysteine, S-(1,2-dichlorovinyl)-homocysteine and 1,1,2-trichloro-3,3,3-trifluoro-1-propene in the Fischer 344 rat., *Arch Toxicol.*, 69, 99-110 (1994).
2. Antti H.; Bollard M.E.; Ebbels T.; Keun H.; Lindon, J.C.; Nicholson, J.K. and Holmes E.; Batch satstistical processing of ^{1}H NMR-derived urinary spectral data, *J. Chemometr.*, 16, 461-468 (2002).
3. Gartland, K.P.R.; Sanins, S.M.; Nicholson, J.K.; Sweatman, B.C.; Beddell, C.R. and Lindon, J.C.; Pattern recognition analysis of high resolution ^{1}H NMR spectra of urine: A nonlinear mapping approach to the classification of toxicological data, *NMR Biomed.*, 3, 166-172 (1990).
4. Kyoto Encyclopedia of Genes and Genomes: Kanehisa, M.; A database for post-genome analysis. Trends Genet. 13, 375-376 (1997). Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids. Res. 28, 27-30 (2000). Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.F.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; and Hirakawa, M.; From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354-357 (2006). Reference by Bio-Rad to KEGG does not imply ownership or endorsement by either party.
5. Nicholson, J.K. and Wilson, I.D. High resolution proton NMR spectroscopy of biological fluids, *Prog. NMR Spectrosc.*, 21, 449-501 (1989).
6. Nicholson, J.K.; Timbrell, J.A.; and Sadler, P.J.; Proton NMR spectra of urine as indicators of renal damage: Mercury nephrotoxicity in rats, *Mol. Pharmacol.*, 27, 644-651 (1985).
7. Nicholson, J.K.; O'Flynn, M.; Sadler, P.J.; Macleod, A.; Juul, S.M. and Sonksen. P.H.; Proton NMR studies of serum, plasma and urine from fasting normal, and diabetic subjects, *Biochem. J.*, 217, 365-375 (1984).
8. Nicholson, J.K. Higham; D. Timbrell, J.A. and Sadler, P.J.; Quantitative ^{1}H NMR urinalysis studies on the biochemical effects of acute cadmium exposure in the rat., *Mol. Pharmacol.*, 36, 398-404 (1989).