An Integrated Software Environment for NMR-Based Metabolomics Data Analysis - Leap to Biomarker Identification from PCA

*Michelle D'Souza, Ph.D., Regis Grenier, Gregory M. Banik, Ph.D., Chen Peng, Ph.D.

Bio-Rad Laboratories, Inc., Informatics Division, Two Penn Center Plaza, Suite 800, 1500 John F. Kennedy Blvd, Philadelphia, PA 19102

Introduction

There has been a maturation in computer technologies that convert the instrumental measurements of patient samples to a collection of spectra and then prepare the spectra for sophisticated multivariate analysis tools. The next steps, recognition of altered concentrations of metabolites and identification of the affected metabolic pathways, remain a challenge. We have developed two technologies that combined with an integrated informatics system to accomplish this critical step.

Using this integrated informatics approach, raw NMR spectra are processed, and the processed NMR spectra are transferred to an integrated chemometrics data analysis tool. The outcome is a Principal Component Analysis (PCA) of the spectra, which includes a graphical display called a Scores Plot. In the Scores Plot, each NMR spectrum from a metabolomics experiment is represented as a single point in the plot. Similar spectra will appear nearer to one another in a Scores Plot, and dissimilar spectra will appear further away from one another. A related diagnostic graphical display called a Loadings Plot indicates the regions of the NMR spectra that describe the most variance in the NMR data set. The Scores Plot is related to the Loadings Plot in that the Loadings Plot describes the spectral regions that are highly variable in the experimental samples and therefore where in the Scores Plot each point will appear.

A Loadings Plot can be used as a query to search a reference metabolite database to identify a list of putative metabolites whose concentration is changing throughout the course of the experiment. Alternatively, the reference metabolite database can be projected into the same space as the PCA Loadings Plot and filtered to identify the putative metabolites visually in the Loadings Plot. In this second approach, the list of filtered metabolites can be used to create a rank-ordered list of putative biological pathways based on the co-occurrence of the filtered metabolites in each pathway. Direct internet links from the rank-ordered list to the KEGG database allow the researcher to explore each pathway in more detail.

Methodology

1) PCA

The Principal Component Analysis (PCA) was run with the AnalyzeIt™ MVP package, IPAK,[1] integrated in KnowItAll software platform.[2]

2) Peak Search

Peak Search compares lists of NMR chemical shifts (Peak Tables) from a database to the Peak Table created for a query spectrum or a Loadings Plot. The query Peak Table is configurable in that parameters for the tolerance and the minimum number of peaks to match can be set by the user. Both positive and negative peaks can be selected since Loadings Plots can have positive or negative values, but only the peak locations (ppm) are used. The closeness of the match between the query Peak Table and the database Peak Tables is indicated with a Hit Quality Index (HQI) rating scale. Database hits are sorted in the descending order of HQI.

3) Database Projection

Based on the concept first published by Dieterle, etc.,[3] NMR spectra from a reference metabolite database are projected onto the PCA Scores Plot of the NMR spectra from the original experiment using the Loadings Plot from the original experiment:

$$T_{Met} = X_{Met} \times P \quad [1]$$

where X_{Met} is the matrix of the spectra of the metabolites. P is the (transposed) Loadings Plot matrix from the original experiment PCA. The resulting T_{Met} is a [m x n] matrix, where m is the number of metabolite spectra and n is the maximum number of factors. The T_{Met} can be plotted and overlaid on top of the Scores Plot from the original experiment PCA.

If a standard metabolite significantly contributes to the separation of the samples, the metabolite is projected to have large values on factor axes. The non-significant metabolites cluster around the origin (0, 0, 0). This allows the user to visually recognize the significant metabolites. The location of a projected spectrum along a certain factor axis depends on spectrum and loadings correlation and the Y-scale of the spectrum. Y-scaling options can be applied to a spectrum containing many peaks to avoid a "false negative" and to a spectrum of no peaks at all to avoid it becoming a "false positive."

The X-residuals (the difference between the original spectra and the reconstructed spectra) of the metabolites indicate how reliable the projection results are for the particular samples. Compounds generate big residuals can be filtered out.

4) Ranked Links to KEGG Pathways

The KEGG pathways are linked to the filtered standard compounds. They are ordered by occurrences among filtered compounds. A user can click on a pathway to go to the KEGG online system.[4]

5) NMR Metabolism Standard Database and Samples

A standard metabolite NMR database using data publicly available from University of Wisconsin was constructed. Currently, it contains 279 metabolite NMR spectra.[5]

The blood sample metabolite NMR spectra were contributed by Professor Xia of Beijing University. It contains samples from 14 diabetic patients and 20 non-diabetic people.

6) Process

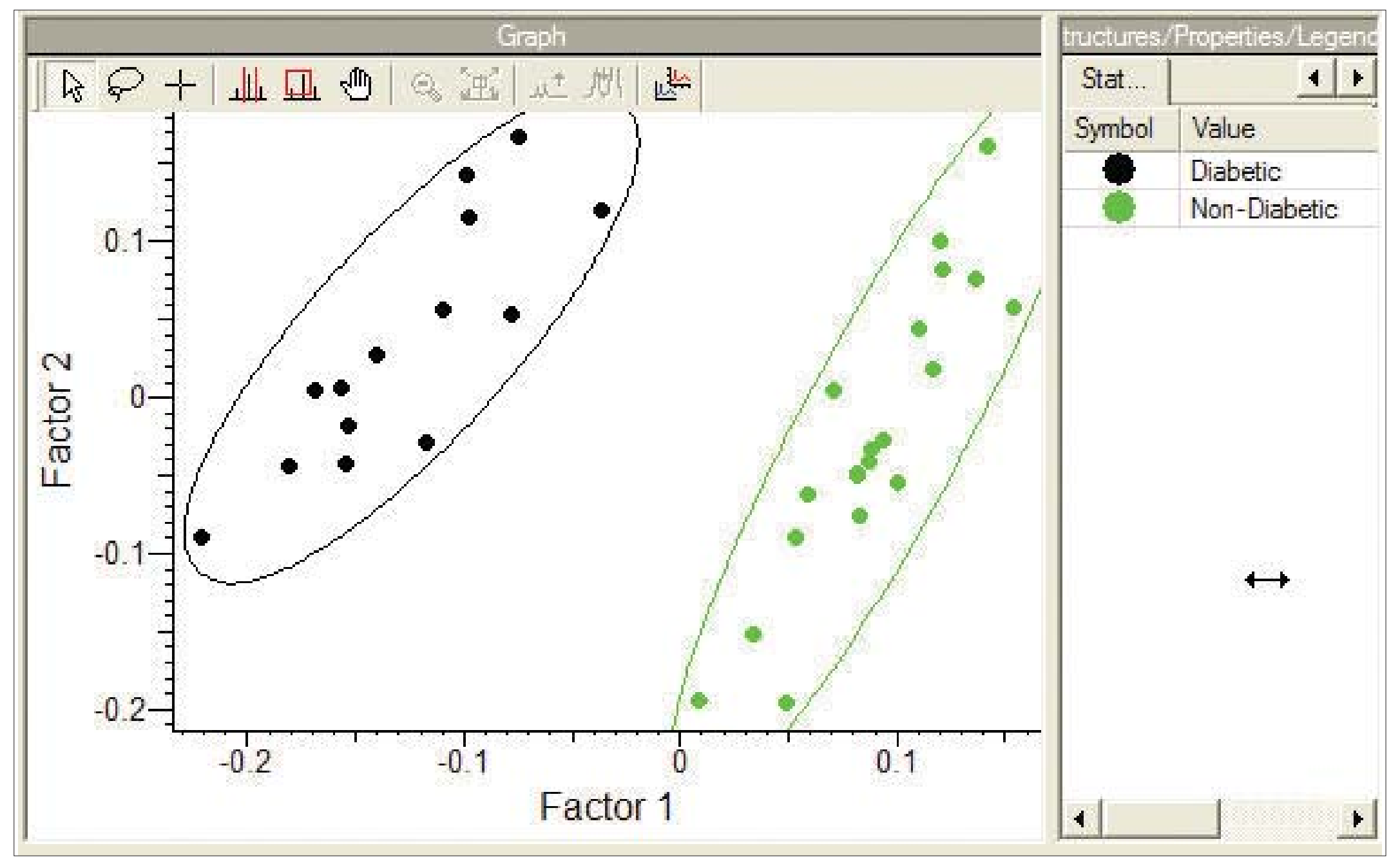
Each raw instrument signal was processed into a NMR peak spectrum with water peak removed at 4.75 - 5.5 pm region. The 0.5 - 8.5 ppm region of sample spectra were analyzed by multivariate analysis. Each spectrum was transformed by subtracting its baseline value and dividing by "sample 2-norm" transformation. "Mean-centering" was performed in spectral pre-processing.

The Loadings Plot of "factor 1" was used for the peak search against the standard metabolite database. Compounds with high HQI are most likely to be the biomarkers. The standard metabolites are projected on to the sample PCA space. Possible biomarkers would shift towards (or located close to) the diabetic patients cluster.

Result and Discussion

1) PCA Class Separation

The PCA was able to separate these two classes (Figure 1). As one can see, the two classes "Diabetic" and "Non-Diabetic" are well separated alone "factor 1" axis.



Symbol	Value
●	Diabetic
●	Non-Diabetic

Figure 1. Diabetic patient blood samples PCA. The green spots are non-diabetic population; the black spots are diabetic population.

2) PCA Loadings Plot

The Loadings Plot of the above PCA resembles a NMR spectrum, but the peaks actually signify the class separation contributed by the sample spectra.

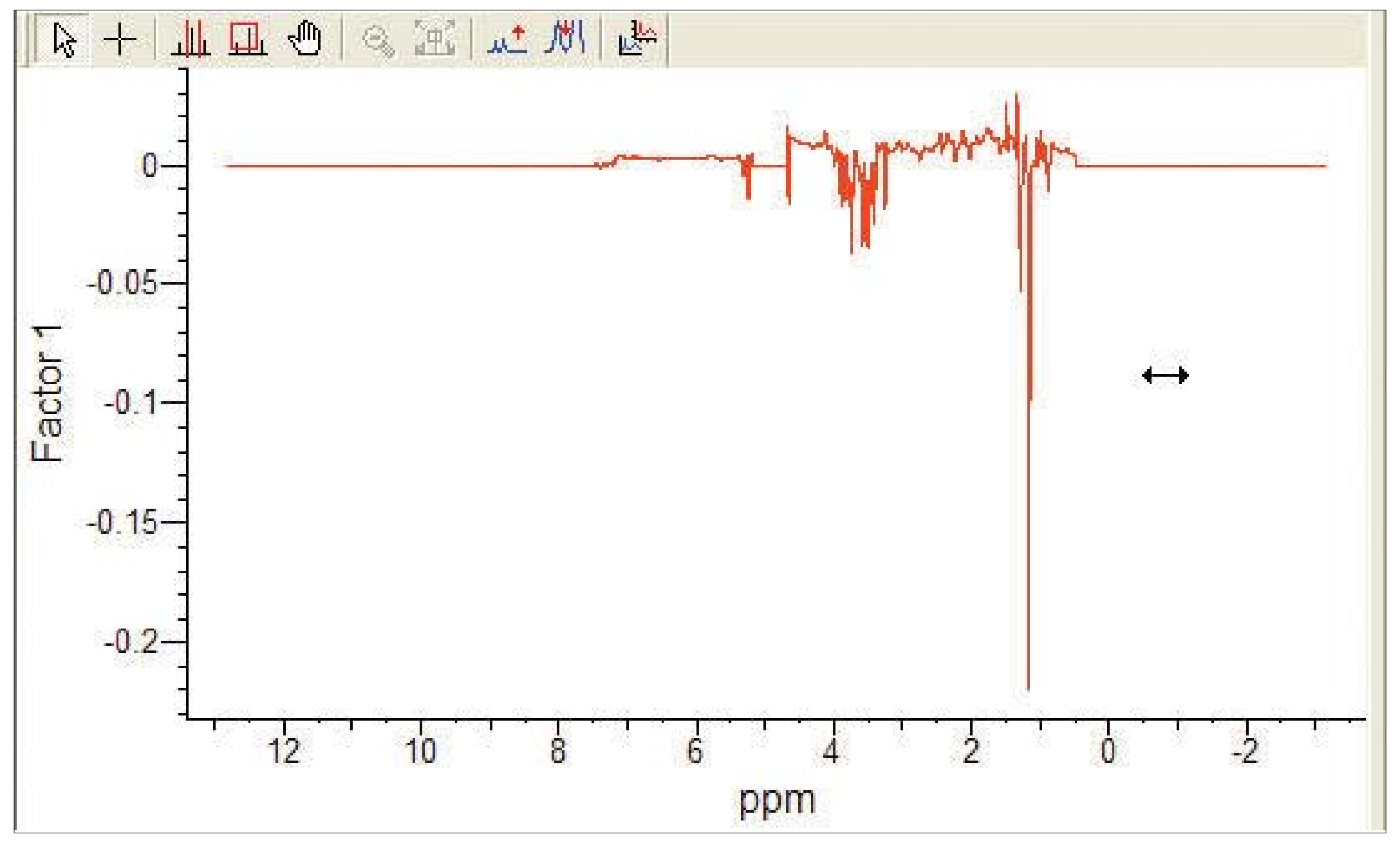


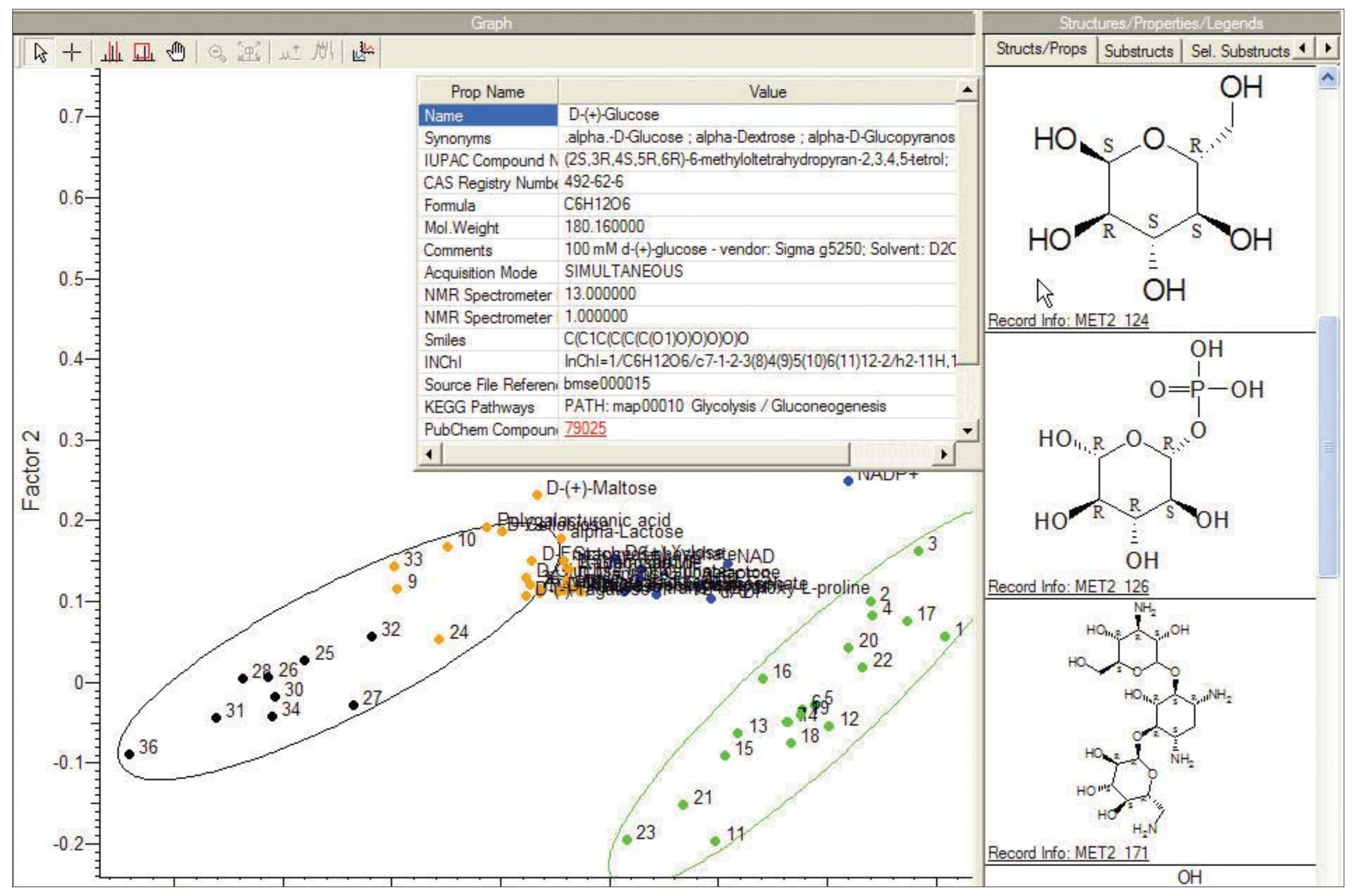
Figure 2. Sample spectra PCA Loadings Plot.

The Loadings Plot peaks are selected to search against the standard metabolite database. The peak intensity threshold was 0.0126 to filter out minor peaks. Negative peaks were included. We required at least four peaks from a standard spectrum to match the query. We found the top eight hits are: D-Cellobiose, D-(+)-Glucose, Melibiose, L-(-)Arabitol, N-Acetyl-D-glucosamine, alpha-D-Glucose -1-phosphate, D-Glucuronate, Sucrose. Except L-(-)Arabitol which is a false positive, all seven top hits contain sugar structure fragment. This exercise confirms with observation that high blood sugar level is an indication of diabetic disease. Therefore, peaks search of a sample PCA Loadings Plot is a viable way to find compounds that are biomarkers.

The KEGG pathways are linked to the hit compounds. A user can access the metabolism pathways with a single click.

3) Database Projection

We projected the standard metabolite database onto the PCA space. The compound filtration criterion is: top 10% by distance from the origin (0, 0, 0), which suggests how much a compound contributes to the separation. The result is show in Figure 3, where green spots are non-diabetic population, and the black spots are diabetic population. The blue spots represent standard compounds whose NMR spectra are projected into this PCA space.

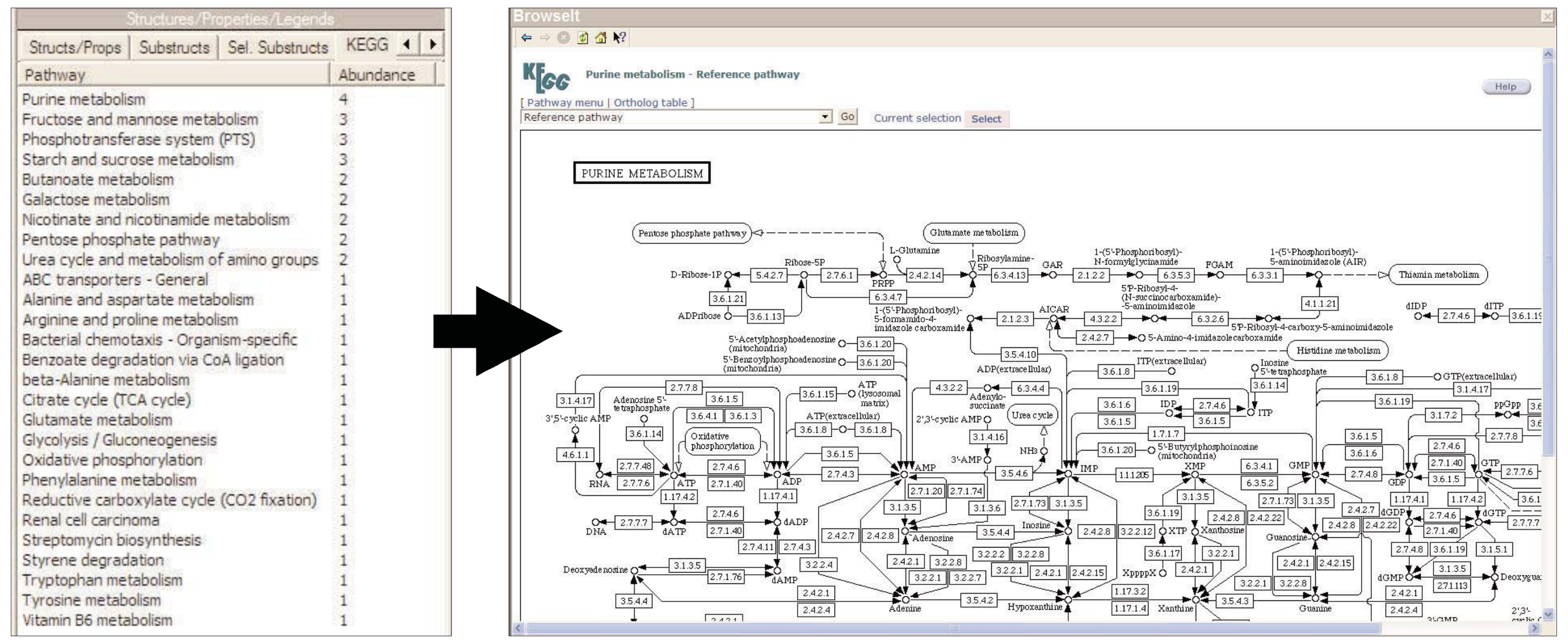


Prop Name	Value
Name	D-(+)-Glucose
Synonyms	.alpha.-D-Glucose ; alpha-Dextrose ; alpha-D-Glucopyranos
IUPAC Compound N	(2S,3R,4S,5R,6R)-6-methyloltetrahydropyran-2,3,4,5-tetrol;
CAS Registry Numbe	492-62-6
Formula	C6H12O6
Mol.Weight	180.160000
Comments	100 mM d-(+)-glucose - vendor: Sigma g5250; Solvent: D2C
Acquisition Mode	SIMULTANEOUS
NMR Spectrometer	13.000000
NMR Spectrometer	1.000000
Smiles	C(C1C(C(C(C(O1)O)O)O)O)O
INChI	InChI=1/C6H12O6/c7-1-2-3(8)4(9)5(10)6(11)12-2/h2-11H,1
Source File Referen	bmse000015
KEGG Pathways	PATH: map00010 Glycolysis / Gluconeogenesis
PubChem Compoun	79025

Figure 3. Sample spectra PCA Loadings Plot.

Compounds containing sugar structure fragment are clustered around the black (diabetic) circle as shown in Figure 3. Examing the top 10% compounds filtered, we found them mostly consistent yet slightly different to those suggested by peak search. Therefore, database projection is an effective method to explore the possible NMR-based biomarkers for diseases.

The KEGG pathways linked to the filtered standard compounds are listed by occurrences. This allows the user to prioritize the metabolic pathways to explore. A user can click on a pathway to go to the KEGG online system.



Pathway	Abundance
Purine metabolism	4
Fructose and mannose metabolism	3
Phosphotransferase system (PTS)	3
Starch and sucrose metabolism	3
Butanoate metabolism	2
Galactose metabolism	2
Nicotinate and nicotinamide metabolism	2
Pentose phosphate pathway	2
Urea cycle and metabolism of amino groups	2
ABC transporters - General	1
Alanine and aspartate metabolism	1
Arginine and proline metabolism	1
Bacterial chemotaxis - Organism-specific	1
Benzoate degradation via CoA ligation	1
beta-Alanine metabolism	1
Citrate cycle (TCA cycle)	1
Glutamate metabolism	1
Glycolysis / Gluconeogenesis	1
Oxidative phosphorylation	1
Phenylalanine metabolism	1
Reductive carboxylate cycle (CO2 fixation)	1
Renal cell carcinoma	1
Streptomycin biosynthesis	1
Styrene degradation	1
Tryptophan metabolism	1
Tyrosine metabolism	1
Vitamin B6 metabolism	1

Figure 4. KEGG Pathways online system.

Conclusion

It is important to obtain a PCA where classes of samples are clearly separable. The subsequent peak search or database projection are viable methods to discover compounds which contribute to the class separation. This system completes the steps to go from a PCA to biomarker(s) identification and finally, to suggest the possible metabolic pathways.

Reference

1. IPAK™, version 4.0. Infometrix, Inc., Bothell, WA. www.infometrix.com.
2. KnowItAll®, version 7.8. Bio-Rad Laboratories, Inc., Philadelphia, PA. www.knowitall.com.
3. Dieterle, F.; Ross, A.; Scholetterbeck, G.; Senn, H. Metabolite Projection Analysis for Fast Identification of Metabolites in Metabonomics. Application in an Amiodarone Study. Anal. Chem., 2006, 78 (11), 3551 -3561.
4. KEGG: Kyoto Encyclopedia of Genes and Genomes. Available online at: www.KEGG.com.
5. Available online at: www.bmrb.wisc.edu/metabolomics/metabolomics_standards.html.

Acknowledgement

We wish to thank our partners who contributed data and technology: Dr. Tao Wang and Professor Bin Xia of Beijing NMR Center, Peking University; Dr. Ian Lewis and John Markley of Department of Biochemistry, University of Wisconsin-Madison; Scott Ramos and Brian Rohrback of Infometrix, Inc.