Biomarker Identification Combining Multivariate Analysis of NMR Spectra with an Innovative Spectral Data Analysis Approach in an Integrated Working Environment

Regis Grenier, Yann Bidault, Chen Peng, Ph.D. and Gregory M. Banik, Ph.D. Bio-Rad Laboratories, Inc., Informatics Division, 3316 Spring Garden St, Philadelphia PA 19104, USA
Scott Ramos, Brian Rohrback, Infometrix Inc., 10634 E. Riverside Dr., Suite 250, Bothell, WA 98011, USA
Tao Wang, Bin Xia, Beijing NMR Center, Peking University, Beijing 100871, China

Abstract

An integrated informatics solution, compatible with multiple instrument formats, covering the full range of steps from raw data processing to biomarker identification, has been long-anticipated by researchers performing metabonomics studies. Such a solution now exists, and we will demonstrate its successful application to the analysis of ^{1}H NMR spectra of human serum samples from 37 diabetic and non-diabetic subjects. This integrated informatics approach also includes an innovative patent-pending spectral Overlap Density Heatmap (ODH) tool. By relating statistical findings to spectral observation in a supervised approach, ODH complements Principal Component Analysis (PCA) and offers meaningful direction for biomarker identification.

The integrated informatics approach to metabolomics research includes the following steps:

- Batch processing of raw FIDs, phasing, baseline correction, and cross-spectral alignment.
- Batch import of the processed spectra.
- Include/exclude spectral ranges, optional binning and bucketing, pre-processing, Y-transforms, and PCA.
- Visualization as an Overlap Density Heatmap, a novel method to quantitatively evaluate the similarity/dissimilarity among multiple overlaid spectra.
- Comparative use of loadings plots resulting from PCA and peaks resulting from Overlap Density Heatmaps as search queries against a database of known metabolites.
- Link to the KEGG database for metabolic pathways of identified metabolites.

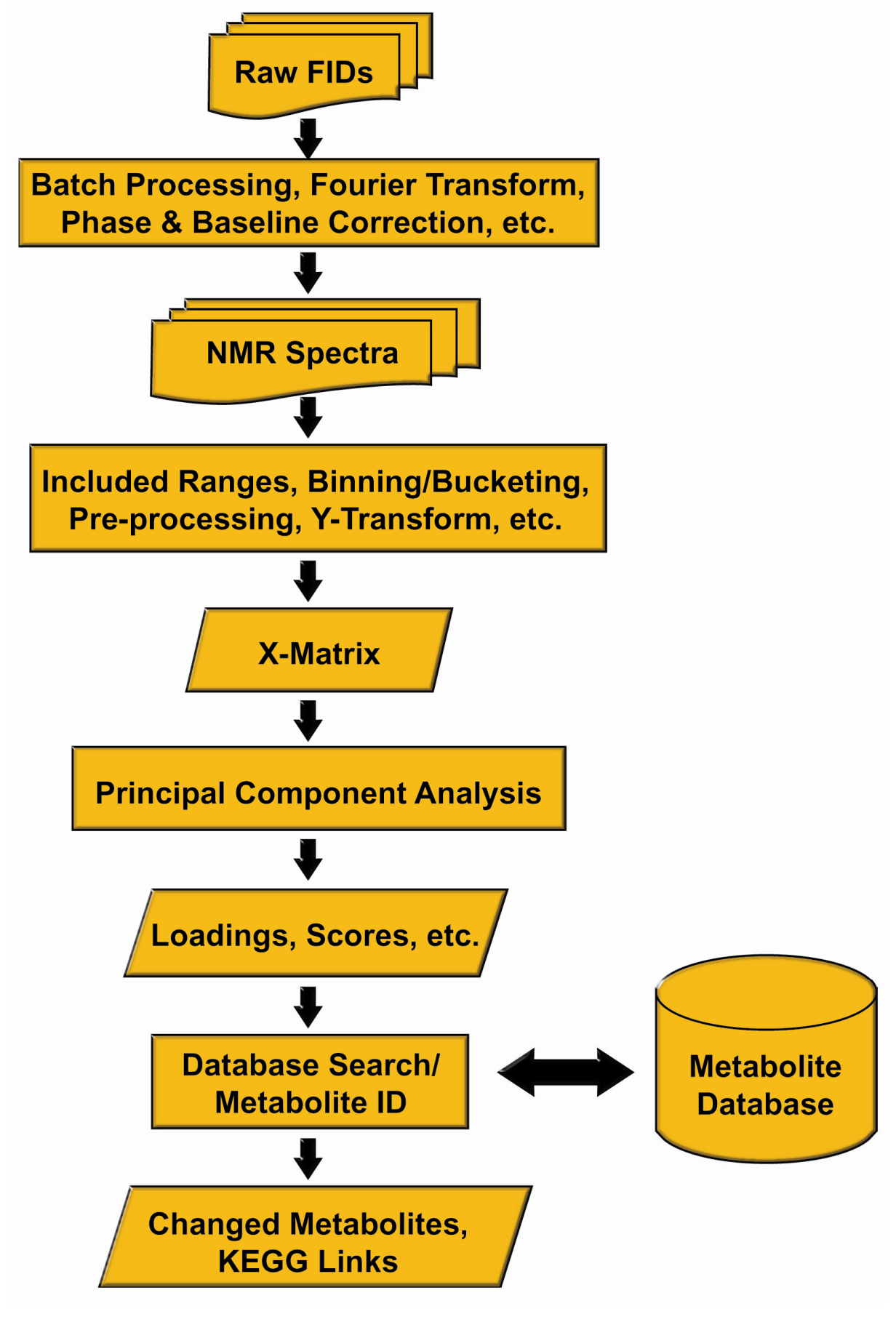

Raw FIDs
Batch Processing, Fourier Transform, Phase & Baseline Correction, etc.
NMR Spectra
Included Ranges, Binning/Bucketing, Pre-processing, Y-Transform, etc.
X-Matrix
Principal Component Analysis
Loadings, Scores, etc.
Database Search/ Metabolite ID
Metabolite Database
Changed Metabolites, KEGG Links

Figure 1 — Biomarker Identification Workflow.

Materials and Methods

Data

1. 37 blood samples were collected from 17 diabetic and 20 non-diabetic patients (origin: Prof. Xia, Beijing NMR Center). After clotting, ^{1}H-NMR spectra were generated from serum as follows:
    - BRUKER Avance-500 spectrometer, 500.13 MHz
    - 64 scans collected for each sample, 8K data points
2. The NMR FIDs were imported into Bio-Rad Laboratories' KnowItAll® Informatics System, Metabolomics Edition.
3. The NMR Metabolite Database from the University of Wisconsin Madison was used for metabolite identification.

Spectrum Processing

The 37 FIDs were batch processed by the ProcessIt™ NMR module in the KnowItAll platform. The parameters used for processing are listed in Figure 2, and were applied as a macro function to all the spectral data.

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 baseline is the highest. This method works well for spectra with only positive peaks, a relatively flat baseline, and does not require that the spectra have many well-isolated peaks–which is usually impractical in metabolomics studies.

#	Function name	Parameters
1	DC Offset	
2	Zero Fill	ZF Factor =2
3	Exponential WF	LB=0.5
4	Default Fourier	
5	Auto Phase	Method =GoodLook
6	Baseline Correction	Spline
7	Set Reference	Ref Point=13125, PPM= 0

Figure 2 — NMR spectra batch processing parameters.

Data Pre-Processing and Principal Component Analysis

A chemometrics component, Infometrix' Pirouette®, has been integrated within Bio-Rad's KnowItAll platform for performing PCA. 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 1st 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. These settings are displayed in Figure 3.

For the purpose of this study, we used the spectral data points as input to the PCA (no binning, full resolution). The software, however, offers a full range of binning-bucketing options such as fixed-width bucketing, Intellibucket™–variable width binning (setting boundaries at local minima by means of an Overlap Density Heatmap Consensus Spectrum using a "looseness" factor e.g., bin widths of 0.04 +/- 0.02 ppm)–, and finally AFNS–Automated Filtering of NMR Spectra, a novel method that selects spectral features based on their statistical significance and then smoothes the spectral points using their optimized filter widths. AFNS uses a rolling binning algorithm with multiple bin widths and ANOVA-based filtering as a means of identifying significant features in complex spectra.

Function Name	Parameters
Pre-processing	Mean-Centering
Maximum factors	8
Binning/Bucketing/Exdude range	Include range: 10.0-0.5 ppm Exdude range: 5.5-4.5 ppm No binning
Y-transforms	Subtract: variable 1 Divide By: sample 2-norm

Figure 3 — PCA data preprocessing parameters.

Overlap Density Heatmap

Overlap Density Heatmap (ODH) is a novel technology, designed by Bio-Rad (patent pending), which allows researchers to visually examine and evaluate spectral differences or commonalities. As opposed to statistical analysis methods such as PCA and other multivariate analysis techniques, ODH is applied to the full resolution spectra, with no binning or bucketing data preprocessing. Compared to a conventional overlay display of multiple spectra, ODH allows one to quickly identify the highly common areas (in red) and less common areas (in violet) in each group, and hence provides a better technique to overview multiple spectra (Figure 4). In order to overcome possible peak misalignment issues, which constitute a common problem in metabonomics studies, several global and local peak alignment options exist in the KnowItAll system.

ODH can be utilized either in a non-supervised or in a supervised manner. In a non-supervised approach, all the sample spectra are selected regardless of their class of origin. By moving the ODH selector toward "Dissimilarity", one will highlight the peak areas of highest variability. Peaks of highest interest for class separation are then identified in a supervised way by examining the ODH at OD level = 0 and its related consensus spectrum for each class considered.

Results and Discussion

Identifying spectrum areas of highest variability with ODH

We first utilized ODH in a non-supervised manner—all samples (i.e. non-diabetic and diabetic) were selected—with an ODH Level = -36 that corresponds to an area under the curve (AUC) of 80% relative to the total AUC at OD level = 0. Four areas of high variability—noted (a), (b), (c), and (d) in Figure 6—are identified.

ODH is then used in a supervised way by generating a consensus spectrum (OD level = 0) for each class: diabetic and non-diabetic. The peaks that best point to class separation are the peaks in the carbohydrate area between 4.04 and 3.37 ppm (A), and in the aliphatic area with peaks centered at 1.30 (B), 1.18 (C) and 0.856 (D) (Figure 7). Note that C is unique to the diabetic class, while the other areas reflect changes in abundance between non-diabetic and diabetic samples.

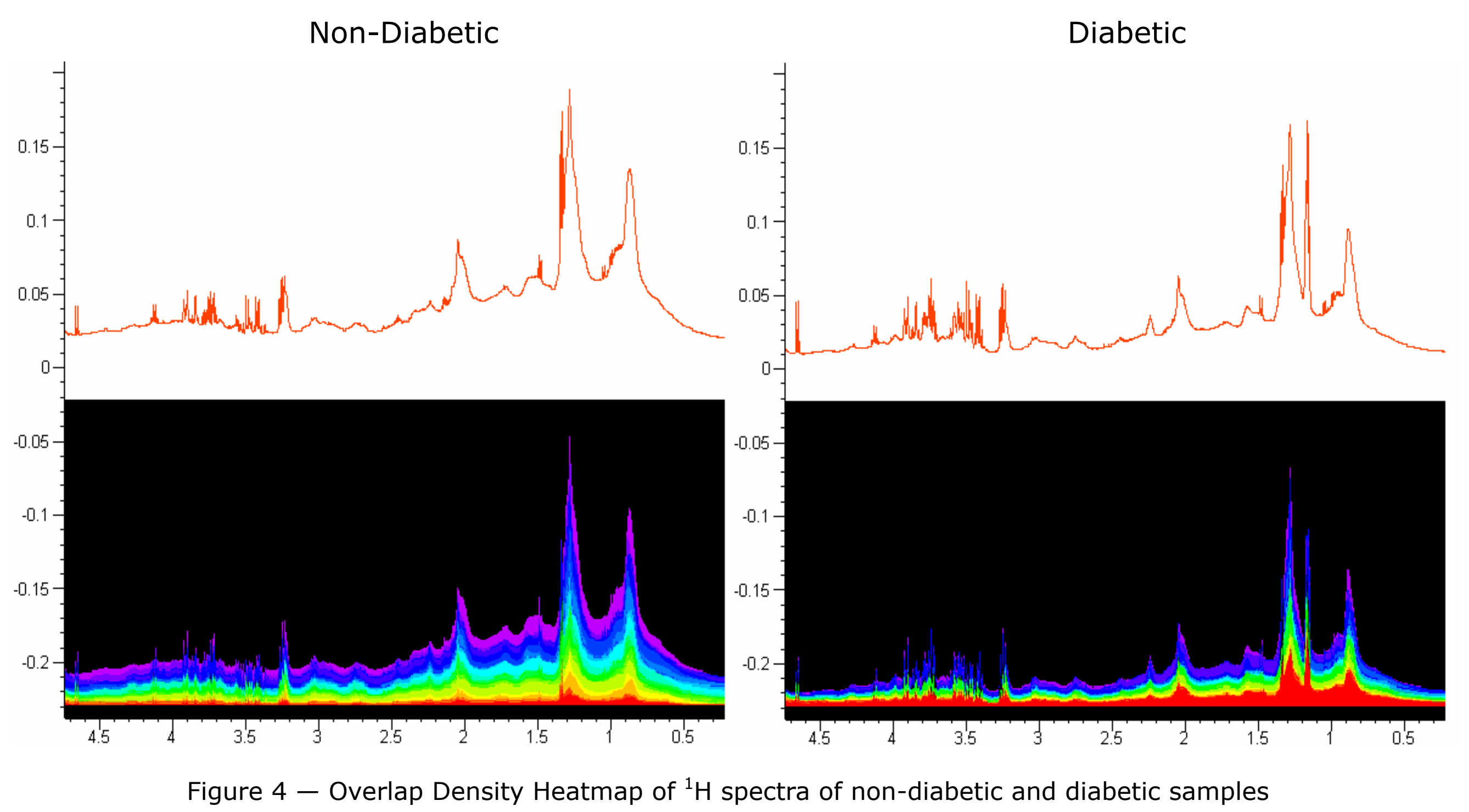


Figure 4 — Overlap Density Heatmap of ^{1}H spectra of non-diabetic and diabetic samples with their respective consensus spectrum on top (ODH Index set at 0 for each class).

Whilst the non-supervised approach helped determine areas of highest variation within all samples, regardless of their class, the supervised approach confirmed those peaks that were most responsible for class separation, providing additional information on the nature of the variability: which peaks are unique to a given class (i.e. (C) unique to diabetic) and which ones are present in both classes, but with significant variation of intensity (i.e. (A), (B) and (D) in one class vs. the other one).

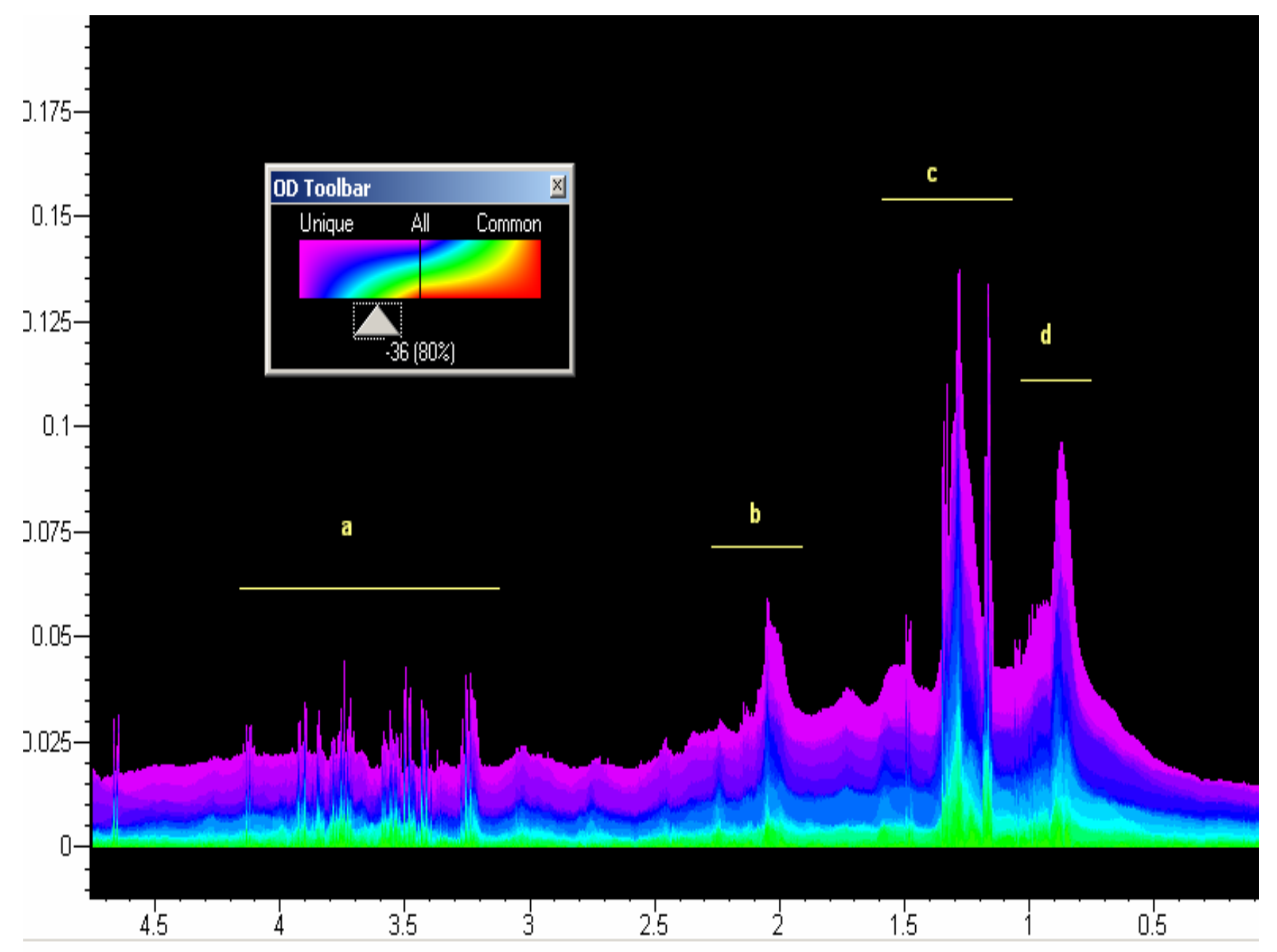


Figure 6 — ODH non-supervised appraoch: identification of areas of high variability in the 4.75-0.25 ppm range.

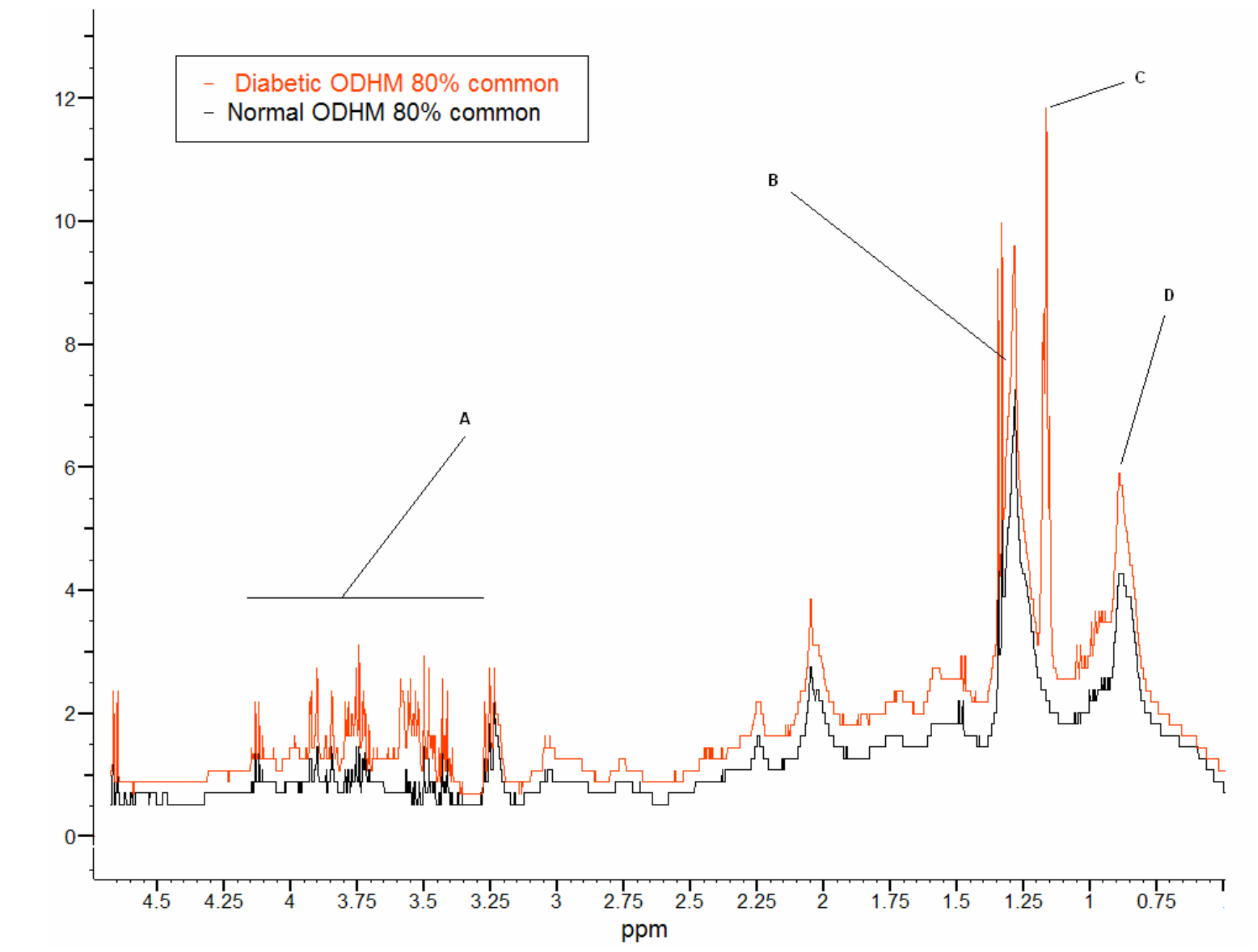


Figure 7 — ODH supervised approach: peaks responsible for best class separation in the 4.75-0.25 ppm range.

Principal Component Analysis

A PCA is performed according to the data pre-processing settings indicated in Figure 3. The scores plot displays a very good clustering of the diabetic vs. non-diabetic samples. (Figure 8). The loadings plot expresses loading values (peak position in ppm) on the X-axis and the intensity of their contributions to the factor variance on the Y-axis (Figure 9).

There is a high level of agreement with the findings of the ODH studies, presented in the former section, that is made visible by examining the loadings plot. Note the importance of peaks or peak areas A, B, C, and D (Figure 11).

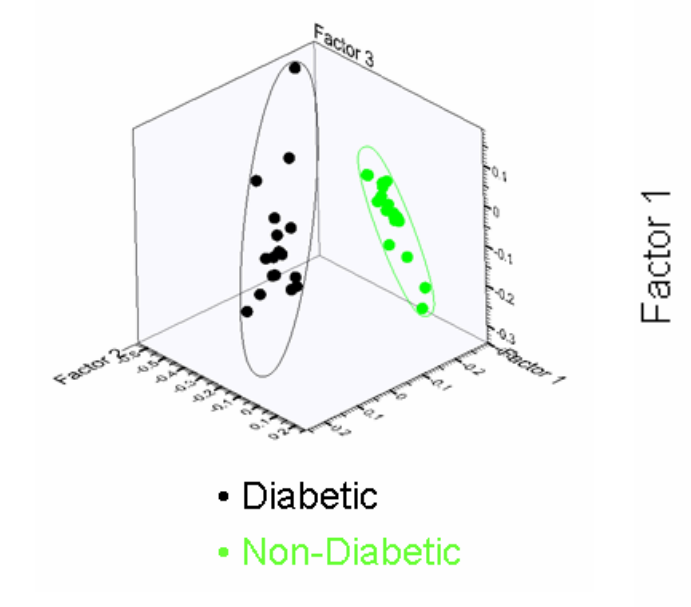


Figure 8 — PCA scores plot.

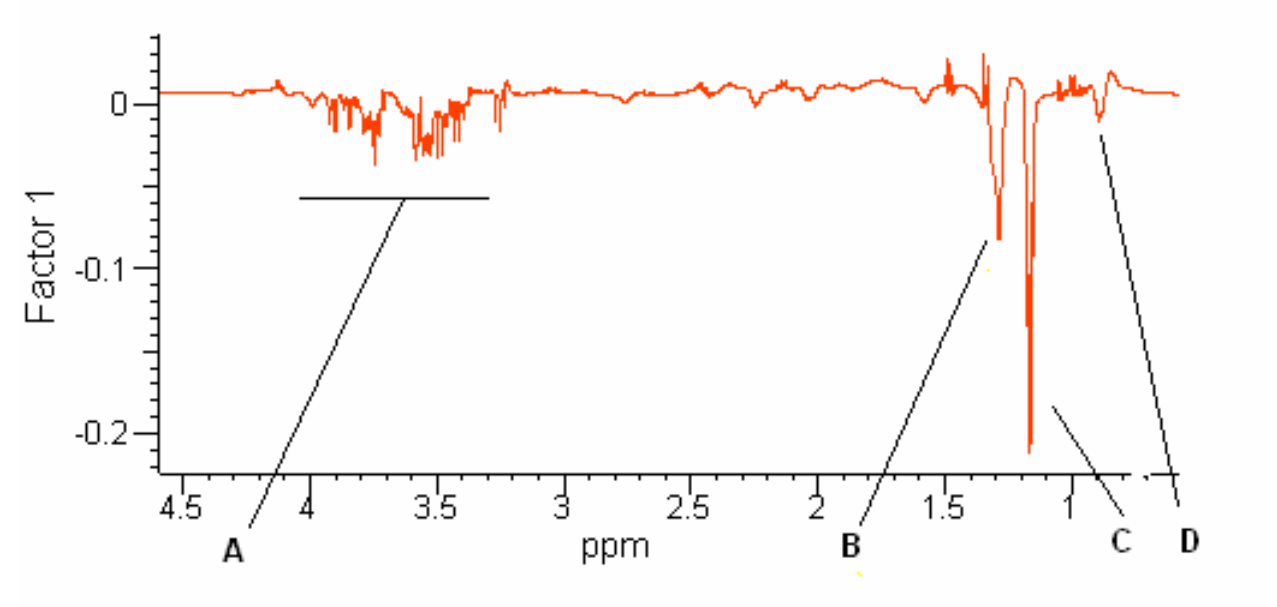


Figure 9 — PCA loadings plot.

From query spectra to metabolite database searching

Both ODH and PCA allow researchers to reconstruct a query spectrum that can be searched against databases of metabolites, either by subtracting the non-diabetic ODH consensus spectrum from the diabetic one (Figure 10) or by transforming the loadings plot into a query spectrum. Both PCA and ODH query spectra are very similar (Figure 11), all peaks—whether negative or positive—being taken into account in the query spectrum confirm the results described above.

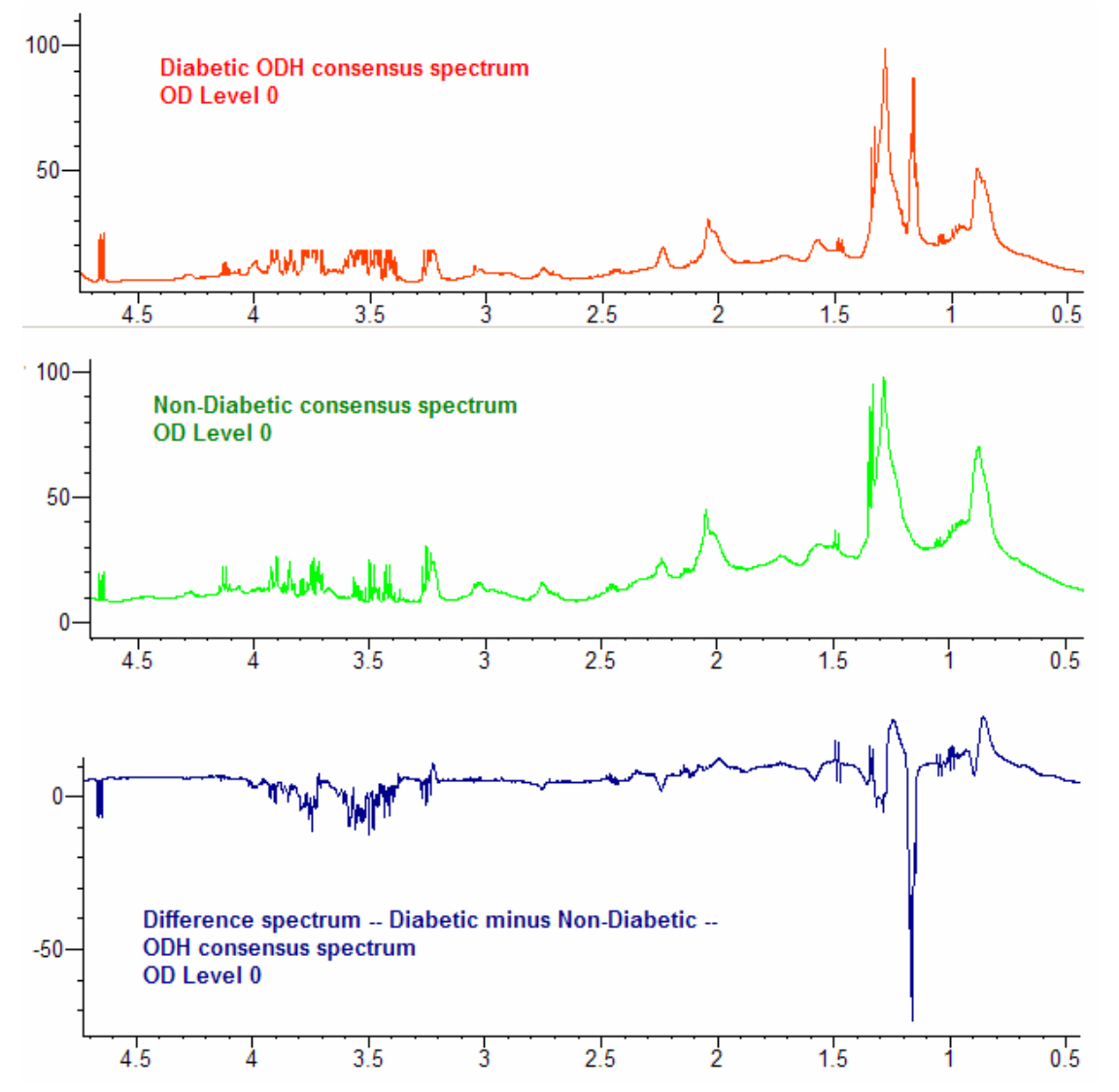


Figure 10 — Constructing an ODH query spectrum.

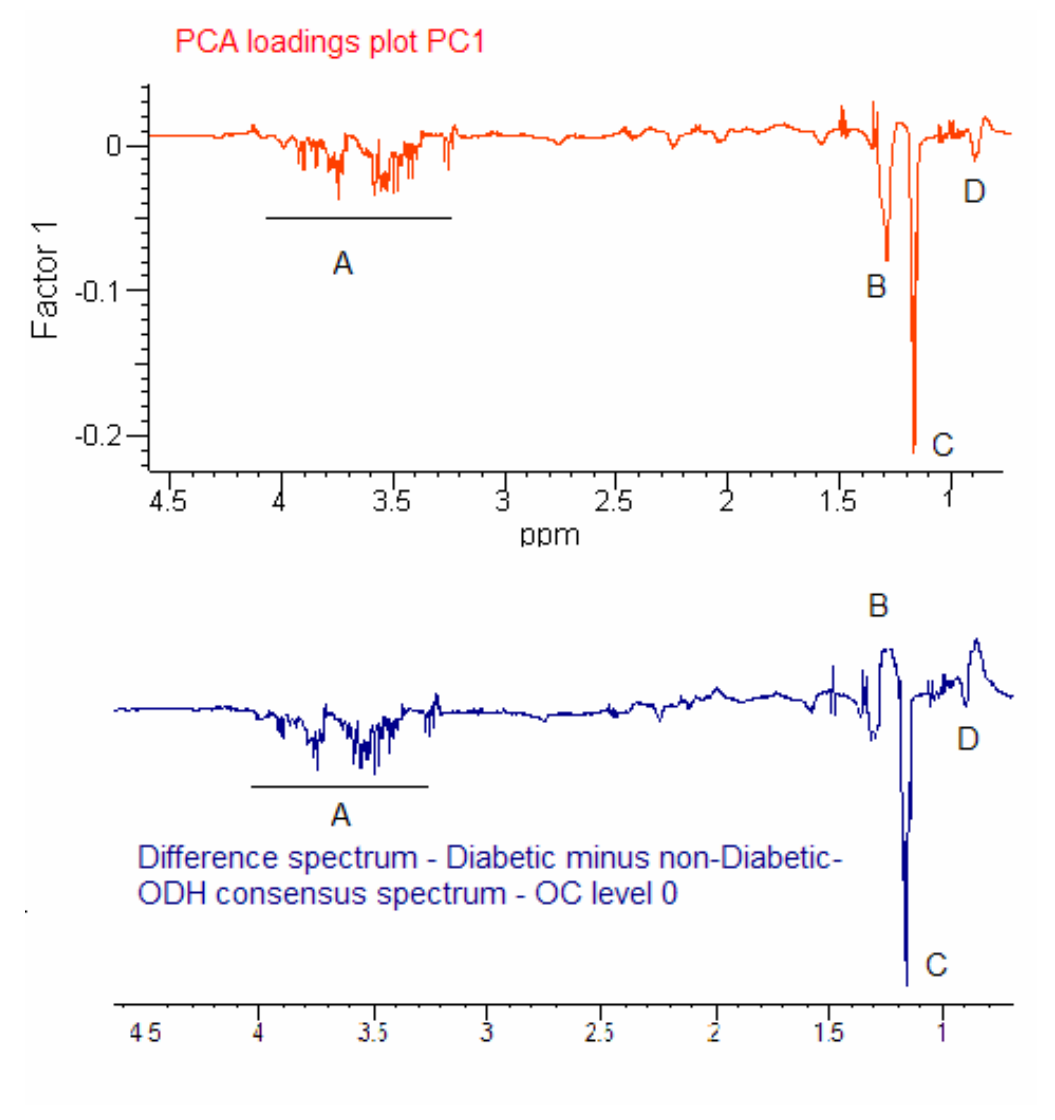
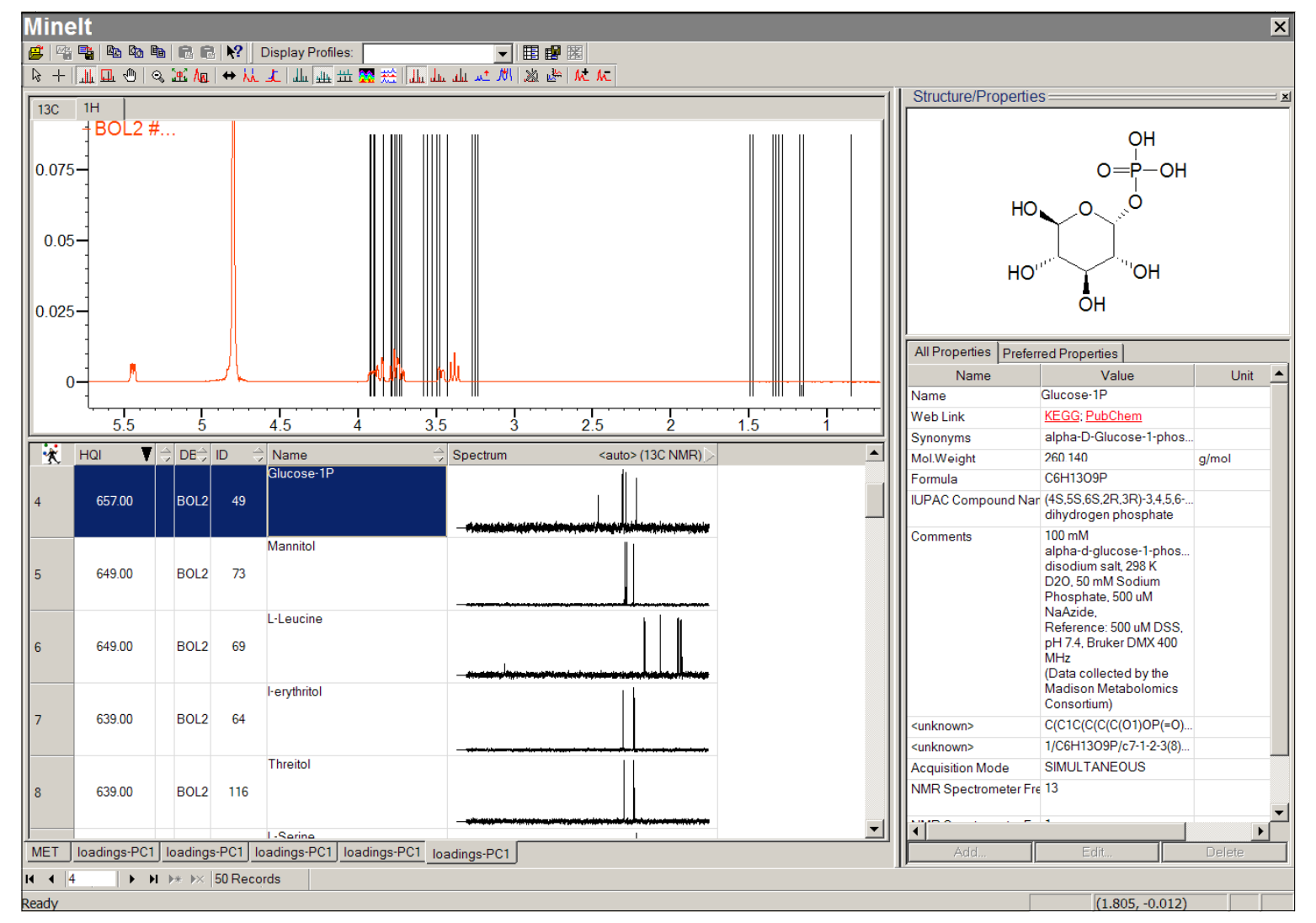


Figure 11 — Comparing ODH and PCA query spectra.



	HQI	DE	ID	Name
4	657.00	BOL2	49	Glucose-1P
5	649.00	BOL2	73	Mannitol
6	649.00	BOL2	69	L-Leucine
7	639.00	BOL2	64	Erythritol
8	639.00	BOL2	116	Threitol

Name	Value	Unit
Name	Glucose-1P	
Web Link	KEGG; PubChem	
Synonyms	alpha-D-Glucose-1-phos...	
Mol.Weight	260.140	g/mol
Formula	C6H13O9P	
IUPAC Compound Nar	(4S,5S,6S,2R,3R)-3,4,5,6-...dihydrogen phosphate	
Comments	100 mM alpha-d-glucose-1-phos... disodium salt, 298 K D2O, 50 mM Sodium Phosphate, 500 uM NaAzide, Reference: 500 uM DSS, pH 7.4, Bruker DMX 400 MHz (Data collected by the Madison Metabolomics Consortium)	
<unknown>	C(C1C(C(C(O1)OP(=O)...	
<unknown>	1/C6H13O9P/c7-1-2-3(8)...	
Acquisition Mode	SIMULTANEOUS	
NMR Spectrometer Fre	13	

Figure 12 - Query spectrum and returned hitlist with records linked to KEGG.

The query spectrum can then be searched against a database of ^{1}H NMR spectra. In this study, we utilized a database provided by the University of Wisconsin Madison, converted in the KnowItAll *.sdb file format in order to make it fully searchable. The hits are ranked by similarity to the query spectrum order. To further link these findings to useful biomarkers, hits resulting from searches are linked to the KEGG database, providing additional information on the metabolite composition and participation in metabolite pathways. An example of such a hit is given in Figure 11 where the 3.9-3.4 ppm range of the spectrum-converted loadings returned glucose-1P as one of the major hits.

Conclusions

This NMR-based metabolomics study demonstrates the potentials of utilizing a software platform that combines the full range of applications from raw data analysis to biomarker identification in one single environment. The well-established PCA and the novel Overlap Density Heatmap technology provided within the KnowItAll® Informatics System complete each other for the identification of key metabolites or biomarkers: ODH provides a visualization tool that "talks" to the chemist and allows direct examination of the data, either in a supervised or non-supervised way. Such an advanced tool can help define the best spectrum processing and data pre-processing options, and lead researchers to identify spectrum areas on which to focus in their analyses. PCA, on the other hand, provides easier automation and reproducibility, as well as easier access to time evolutions via the display of trajectories in toxicity studies.

References

1. 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.* **1984**, *217*, 365-375.
2. 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.* **1985**, *27*, 644-651.
3. Nicholson, J.K. and Wilson, I.D. High resolution proton NMR spectroscopy of biological fluids. *Prog. NMR Spectrosc.* **1989**, *21*, 449-501.
4. 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.* **1990**, *3*, 166-172.
5. 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.* **1994**, *69*, 99-110.
6. 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.* **1989**, *36*, 398-404.
7. Antti H., Bollard M.E., Ebbels T., Keun H., Lindon, J.C., Nicholson, J.K. and Holmes E. Batch stastistical processing of ^{1}H NMR-derived urinary spectral data, *J. Chemometr.* **2002**, *16*, 461-468.
8. Kyoto Encyclopedia of Genes and Genomes: Kanehisa, M.; A database for post-genome analysis. *Trends Genet.* **1997**, *13*, 375-376. Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2000**, *28*, 27-30. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, Hirakawa, M.; From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354-357. Reference by Bio-Rad to KEGG does not imply ownership or endorsement by either party.