

Evaluation of the Orthogonal Projection on Latent Structure Model Limitations Caused by Chemical Shift Variability and Improved Visualization of Biomarker Changes in ^1H NMR Spectroscopic Metabonomic Studies

Olivier Cloarec,^{*,†} Marc E. Dumas,[†] Johan Trygg,[‡] Andrew Craig,[†] Richard H. Barton,[†] John C. Lindon,[†] Jeremy K. Nicholson,[†] and Elaine Holmes[†]

Biological Chemistry, Biomedical Sciences Division, Faculty of Medicine, Imperial College London, South Kensington, London SW7 2AZ, U.K., and Research Group of Chemometrics, Institute of Chemistry, Umeå University, Umeå, Sweden

In general, applications of metabonomics using biofluid NMR spectroscopic analysis for probing abnormal biochemical profiles in disease or due to toxicity have all relied on the use of chemometric techniques for sample classification. However, the well-known variability of some chemical shifts in ^1H NMR spectra of biofluids due to environmental differences such as pH variation, when coupled with the large number of variables in such spectra, has led to the situation where it is necessary to reduce the size of the spectra or to attempt to align the shifting peaks, to get more robust and interpretable chemometric models. Here, a new approach that avoids this problem is demonstrated and shows that, moreover, inclusion of variable peak position data can be beneficial and can lead to useful biochemical information. The interpretation of chemometric models using combined back-scaled loading plots and variable weights demonstrates that this peak position variation can be handled successfully and also often provides additional information on the physicochemical variations in metabonomic data sets.

Recently, postgenomic technologies, including transcriptomic and proteomic methods, have been developing as a means of generating high information content biological assays for disease diagnoses or for evaluating the beneficial or adverse effects of pharmaceuticals.¹ Metabonomics is an emerging complementary postgenomic technology and involves the determination of the levels and time dependence of low molecular mass endogenous metabolites in biofluids such as urine, plasma, or in tissues, as a result of some pathophysiological effect. Most studies on mammalian systems have used ^1H nuclear magnetic resonance

spectroscopy (^1H NMR) for data generation,² but more recently, liquid chromatography hyphenated with mass spectrometry³ has been employed. In other types of biological systems, GC/MS has been used extensively.⁴ Megavariate pattern recognition methods are then used to extract relevant biological information from these complex spectroscopic data, and unsupervised or supervised chemometrics methods, such as principal component analysis (PCA) or discriminant analysis by projection on latent structures (PLS-DA), are often used. Metabonomics has already been used widely to effectively describe the metabolic variation associated with different strains of animal,⁵ toxicity,⁶ disease, or therapeutic intervention.⁷

Individually, biofluid ^1H NMR spectra (typically measured at 500 or 600 MHz) are numerically complex, as they contain many thousands of intensity variables (typically 64k) providing information on hundreds of endogenous metabolites. Furthermore, because of the number of samples required for robust data analysis, particularly for human-related studies where environmental and genetic factors are diverse, the amount of data presented for analysis can be huge, and therefore, many hundreds of megabytes are often needed to define even one data set.

One major problem that has to be addressed by the chemometrics methods is the natural biological variability in metabolite concentrations across a large cohort of samples. In addition, the NMR peaks corresponding to certain molecules, such as citrate,

* To whom correspondence should be addressed. Tel: +442075943114. E-mail: o.cloarec@imperial.ac.uk.

[†] Imperial College London.

[‡] Umeå University.

(1) Lindon, J. C.; Holmes, E.; Nicholson, J. K. *Anal. Chem.* 2003, 75, 384A–391A.

(2) Nicholson, J. K.; Lindon, J. C.; Holmes, E. *Xenobiotica* 1999, 29, 1181–1189

(3) Plumb, R. S.; Stumpf, C. L.; Gorenstein, M. V.; Castro-Perez, J. M.; Dear, G. J.; Anthony, M.; Sweatman, B. C.; Connor, S. C.; Haselden, J. N. *Rapid Commun. Mass Spectrom.* 2002, 16, 1991–1996

(4) Raamsdonk, L. M.; Teusink, B.; Broadhurst, D.; Zhang, N.; Hayes, A.; Walsh, M. C.; Berden, J. A.; Brindle, K. M.; Kell, D. B.; Rowland, J. J.; Westerhoff, H. V.; van Dam, K.; Oliver, S. G. *Nat. Biotechnol.* 2001, 19, 45–50.

(5) Brindle, J. T.; Antti, H.; Holmes, E.; Tranter, G.; Nicholson, J. K.; Bethell, H. W. L.; Clarke, S.; Schofield, P. M.; McKiligin, E.; Mosedale, D. E.; Grainger, D. J. *Nat. Med.* 1999, 8, 1439–1444

(6) Bollard, M. E.; Stanley, E. G.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. *NMR Biomed.* In press

(7) Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Schlotterbeck, G.; Senn, H.; Niederhauser, U.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Chem. Res. Toxicol.* 2002, 15, 1380–1386.

can be subject to peak position variation from sample to sample even though all the NMR chemical shifts are registered against that of a standard compound. The chief cause of such variability is pH differences between samples since ^1H NMR chemical shifts are dependent on the ionization state of such molecules. Other environmental effects can also influence chemical shifts, including metal ion concentrations, metabolite–protein binding, and chemical exchange phenomena.

For these two reasons, methods that reduce the amount of data down to manageable proportions have been used. One of the most common methods involves signal integration within spectral regions or “bins”, and these have generally corresponded to a typical spectral width of 0.04 ppm.^{8,9} This data reduction method also has the property of largely concealing any peak position variation and thus providing more robust modeling. However, the loss of resolution can be a drawback because the interpretation of derived chemometric models in terms of identified metabolic biomarkers cannot easily be obtained from the reduced data but requires reexamination of the real NMR spectra to identify the metabolites responsible for the observed patterns.

An alternative approach has been to use automatic peak alignment algorithms to resolve the problem of peak position variation in ^1H NMR spectra, allowing the use of the full spectral resolution for pattern recognition. Stoyanova et al.¹⁰ removed the positional noise using PCA to determine the misalignment across a series of biofluid NMR spectra. Methods involving the application of a genetic algorithm to align segments of spectra have also been used.^{11,12}

Here, the influence of typical peak position variation on the robustness of pattern recognition models that are based on ^1H NMR spectra of biofluids is evaluated. For this purpose, positional noise was introduced into a set of simulated NMR spectra and pattern recognition was carried out both with and without data reduction on these data sets using a new chemometrics approach orthogonal (O)-PLS-DA as an alternative to classical PLS-DA methods.¹³

The impact of data preprocessing can be very important in carrying out such pattern recognition procedures. The use of mean-centered variables provides principal component (PC) loadings, which are easily interpretable due to their covariance structure and their similarity to a spectrum, and moreover, the varying metabolites can be directly identified from the loadings. However, usually all the metabolites do not have the same variation and the same range of intensities in the NMR spectra. Therefore, the interpretation can be distorted because some metabolites with apparent covariation in the loadings are not really responsible for the discrimination between the different groups or classes. In contrast, autoscaling (mean-centering and univariate scaling) gives the same weight to all the spectral variables

because of their now equal variance, and therefore, the loadings only show the variables, which really impact on the discrimination between classes. However, these loadings are difficult to interpret because their relative values are strongly distorted by the variance scaling procedure. A third method is often therefore used. This is Pareto scaling, which divides each variable by the square root of its standard deviation. The advantages of this technique are the loadings are less distorted and the dominating influence of high variance variables is reduced, but this approach also still has the limitation that the loadings are distorted and the high variance variables have relatively greater weight in the modeling. Here a new approach is presented in which the loadings from autoscaled models are plotted after back transformation with the respective weight of each variable. This then allows a direct interpretation of such loadings as pseudo-NMR spectra.

Finally, an experimental metabonomic study comparing data-reduced and full ^1H NMR spectra is presented, using as an example, the biochemical profiling of mercury(II) chloride nephrotoxicity. Sprague–Dawley rats were treated using a single dose of HgCl_2 , and biochemical effects of this toxin on urinary composition were observed by high-resolution ^1H NMR spectroscopy over a 9-day period.^{8,14}

EXPERIMENTAL SECTION

Treatments and Sample Preparation. A single ip dose of either vehicle (0.9% saline) alone or HgCl_2 was administered to male Sprague–Dawley rats ($n = 5$ per group) at a dose level of $0.75 \text{ mg}\cdot\text{kg}^{-1}$ as described earlier.⁸ Each animal was housed individually in a metabolic cage, and urine samples were collected continuously at various time intervals (predose and 0–8, 8–24, 24–32, 32–48, 48–72, 72–96, 96–120, 120–144, and 144–168 h after treatment). Urine samples were centrifuged at 3000 rpm for 10 min in order to remove particulate contaminants, and the samples were stored at -40°C pending NMR spectroscopic analysis.

To minimize any gross variation in the pH of the urine samples, $200 \mu\text{L}$ of a buffer solution ($0.2 \text{ M Na}_2\text{HPO}_4/0.2 \text{ M NaH}_2\text{PO}_4$, pH 7.4) was mixed with $400 \mu\text{L}$ of urine in a microcontainer. The resulting solution was left to stand for 10 min and then centrifuged at 10 000 rpm to remove any precipitate. However, it is impossible to ensure stable and identical pH values for urine samples, and some pH variation will always be observed across a set of samples. A total of $500 \mu\text{L}$ of the supernatant was placed in a 5-mm-o.d. NMR tube (Wilmad 507PP). A field-frequency lock was provided by adding $100 \mu\text{L}$ of $^2\text{H}_2\text{O}$ to the sample in the NMR tube. Histopathological examination of the kidneys showed clear evidence of proximal tubular necrosis.⁸

^1H NMR Spectroscopy. For each sample, a ^1H NMR spectrum was measured at 600.13 MHz on a Bruker DRX-600 spectrometer. The water resonance was suppressed using the first increment of a NOESY pulse sequence with irradiation during a 3-s relaxation delay and also during the 100-ms mixing time. Typically, 64 free induction decays (FIDs) were collected into 64k data points using a spectral width of 7002.8 Hz, an acquisition time of 4.68 s, and a total pulse recycle time of 7.68 s. Prior to Fourier transformation, an exponential line-broadening factor of

(8) Holmes, E.; Nicholson, J. K.; Nicholls, A. W.; Lindon, J. C.; Connor, S. C.; Polley, S.; Connelly, J. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 245–255.

(9) Spraul, M.; Neidig, P.; Klauck, U.; Kessler, P.; Holmes, E.; Nicholson, J. K.; Sweatman, B. C.; Salman, S. R.; Farrant, R. D.; Rahr, E.; Beddell, C. R.; Lindon, J. C. *J. Pharm. Biomed. Anal.* **1994**, *12*, 1215–1225.

(10) Stoyanova R.; Nicholls, A. W.; Nicholson, J. K.; Lindon, J. C.; Brown, T. R. *J. Magn. Reson.* **2004**, *170*, 329–335.

(11) Forshed, J.; Schuppe-Koistinen I.; Jacobsson S. P. *Anal. Chim. Acta* **2003**, *487*, 189–199.

(12) Lee G.-C.; Woodruff, D. L. *Anal. Chim. Acta* **2004**, *513*, 413–416.

(13) Trygg, J. *J. Chemom.* **2002**, *16*, 283–293.

(14) Holmes, E.; Bonner, F. W.; Sweatman, B. C.; Lindon, J. C.; Beddell, C. R.; Rahr, E.; Nicholson, J. K. *Mol. Pharmacol.* **1992**, *42*, 922–930.

0.3 Hz was applied to the FIDs, which were then zero-filled to 128k.

To use as much relevant data as possible in each NMR spectra, complete spectra in the chemical shift range δ 0–10 were used for pattern recognition. All spectra were phase-corrected and referenced to the CH₃ resonance of creatinine set at δ 3.05 ppm. A baseline correction was also applied to each spectrum using a polynomial curve fit.¹⁶ This was carried using Matlab (Version 6.5, The Mathworks inc, Natick, MA) after importation of 22 000 points/spectrum using spline cubic interpolation to estimate the maximum of the peak. The effect of variations in the presaturation of the water signal was removed by zeroing the intensity values between δ 4.6 and 4.9.

To make a comparison of the use of real NMR spectra and reduced spectra, the original “binning” methodology was also applied,⁶ and here each NMR spectrum was segmented into 250 chemical shift regions of 0.04 ppm.

Finally, to take account of large variations in urine concentration, using both reduced and full resolution data sets, all spectra were then normalized to a total integrated intensity of 100 units.

Simulation of NMR Spectra. A biofluid NMR spectrum can be considered as a linear combination of spectra of pure single-molecule components weighted by their respective contributions, which are generally directly proportional to the concentration of the molecules in the sample. Therefore, for one sample s , a spectrum I_s can be simulated by the following function of the frequency (δ):

$$I_s(\delta) = \sum_{j=1}^{nc} c_{s,j} Sp_j(\delta) + \epsilon(\delta) \quad (1)$$

where $c_{s,j}$ and Sp_j are respectively the concentration and the reference spectra of molecule j , ϵ is the signal not related to any molecules (e.g., baseline drift, instrumental noise), and nc is the number of molecules in the mixture.

The NMR spectrum of an individual chemical species comprises a set of resonances dependent upon the chemical structure, and this can be represented as a linear combination of peaks (nominally of various widths) corresponding to singlets or multiplets according to the neighboring chemical environment. However, in principle in an ideal situation, each peak can be simulated by a Lorentzian line shape. Therefore, it is possible to approximate an NMR spectrum for a single chemical species j by a linear combination of n_j Lorentzians,¹⁵

$$Sp_j(\delta) = \sum_{i=1}^{n_j} \frac{H_{i,j}\omega}{4(\delta - \delta_{i,j})^2 + \omega} \quad (2)$$

where $H_{i,j}$ is the relative intensity of one of the n_j peaks, ω^2 is the full width at half-height of the peaks, and $\delta_{i,j}$ is the position of the peak i .

For real substances, there can be a shift in the position of the peaks for each sample due to differences in solution properties

such as pH. As a first approximation, it is assumed that the shift is equivalent for each resonance for a given molecule but that the magnitude can be different for each sample. This shift ($\Delta\delta_{s,i}$) can be included in the following equation

$$Sp_j(\delta) = \sum_{i=1}^{n_j} \frac{H_{i,j}\omega}{4(\delta - \delta_{i,j} + \Delta\delta_{s,j})^2 + \omega} \quad (3)$$

Combining eqs 1 and 2 and assuming that ω is constant for all of peaks in all of the molecules, eq 4 describes the variation of the biofluid NMR spectra:

$$I_s(\delta) = \sum_{j=1}^{nc} c_{s,j} \sum_{i=1}^{n_j} \frac{H_{i,j}\omega}{4(\delta - \delta_{i,j} + \Delta\delta_{s,j})^2 + \omega} \quad (4)$$

To simulate a data set of biofluid NMR spectra, the following parameters are needed: $c_{s,j}$ is the relative concentration of each molecule j for each sample s and the full set of these parameters defines the discrimination between the classes. $\Delta\delta_{s,i}$ for each sample and each molecule, and these parameters can be defined either randomly or according to the sample classes if the influence on the discrimination between the classes is to be investigated; $\delta_{i,j}$, $H_{i,j}$, and ω are defined choosing chemical compounds of interest; finally, normally distributed random noise is added to all the variables of the simulated spectra in order to mimic instrumental noise.

To make simulated spectra realistic, several molecules were chosen as chemical compounds of interest because of their known peak position variation or overlap. These were citrate, taurine, and trimethylamine *N*-oxide (TMAO). The last two molecules have been chosen not only because their respective signals are subject to positional noise but also because they overlap, a common feature of real NMR spectra. These spectra were generated to simulate full resolution spectra with 3200 data points for 1 ppm. An example of a simulated data set is presented Figure 1. In this example, positional noise has only been added to the citrate peaks.

A number of data sets have been generated based on two distinct sample classes easily separated in the absence of any positional noise. These were used to investigate the following situations for class separation: (i) no position variation; (ii) discrimination based on a compound subject to peak position variation; (iii) discrimination based on the peak position variation, but no concentration differences; (iv) discrimination based on the peak position variation and on a compound subject to peak position variation; (v) discrimination where the peaks from a compound overlap with another, which is subject to peak position variation; (vi) discrimination based on a compound, where all the compounds present are subject to peak position variation.

Pattern Recognition. The multivariate pattern recognition method used in this paper is O-PLS,^{13,17} where the variation in X (the NMR spectra) and Y (the descriptive or class variable) is

(15) Gunther H. *NMR Spectroscopy*; John Wiley & Son: Stuttgart, 1980.

(16) Ebbels, T. M. D.; Lindon, J. C.; Nicholson, J. K.; Holmes, E. Methods for spectral analysis and their applications: spectral replacement. U.S. Patent 20010029380 20011220 [US2002145425], 2002.

(17) Trygg, J.; Wold, S. *J. Chemom.* **2003**, *18*, 53–64.

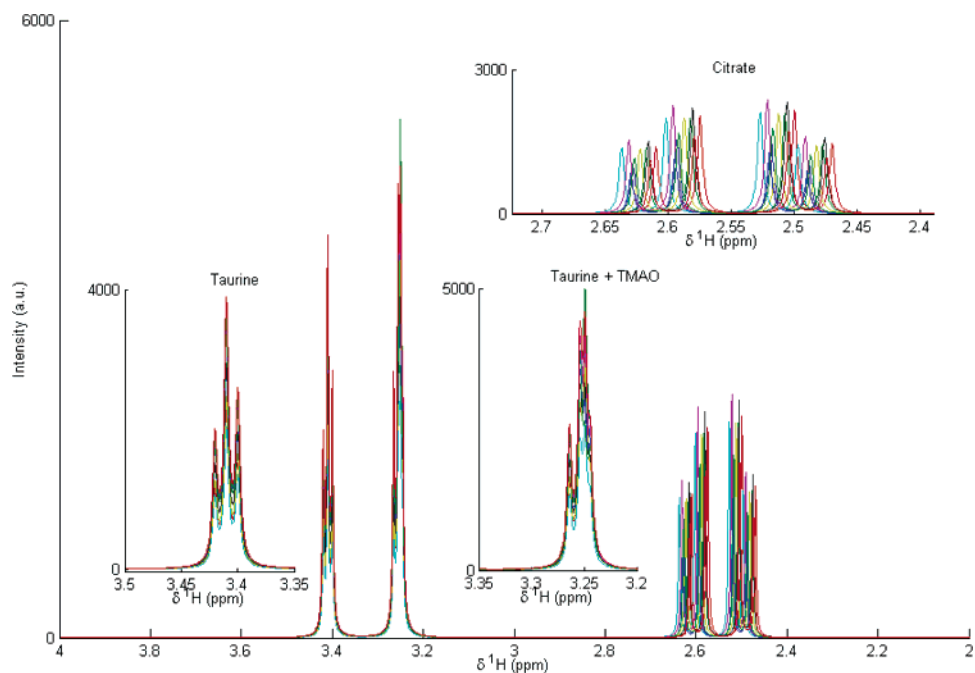


Figure 1. Example of simulated ^1H NMR spectra based on citrate, taurine, and TMAO.

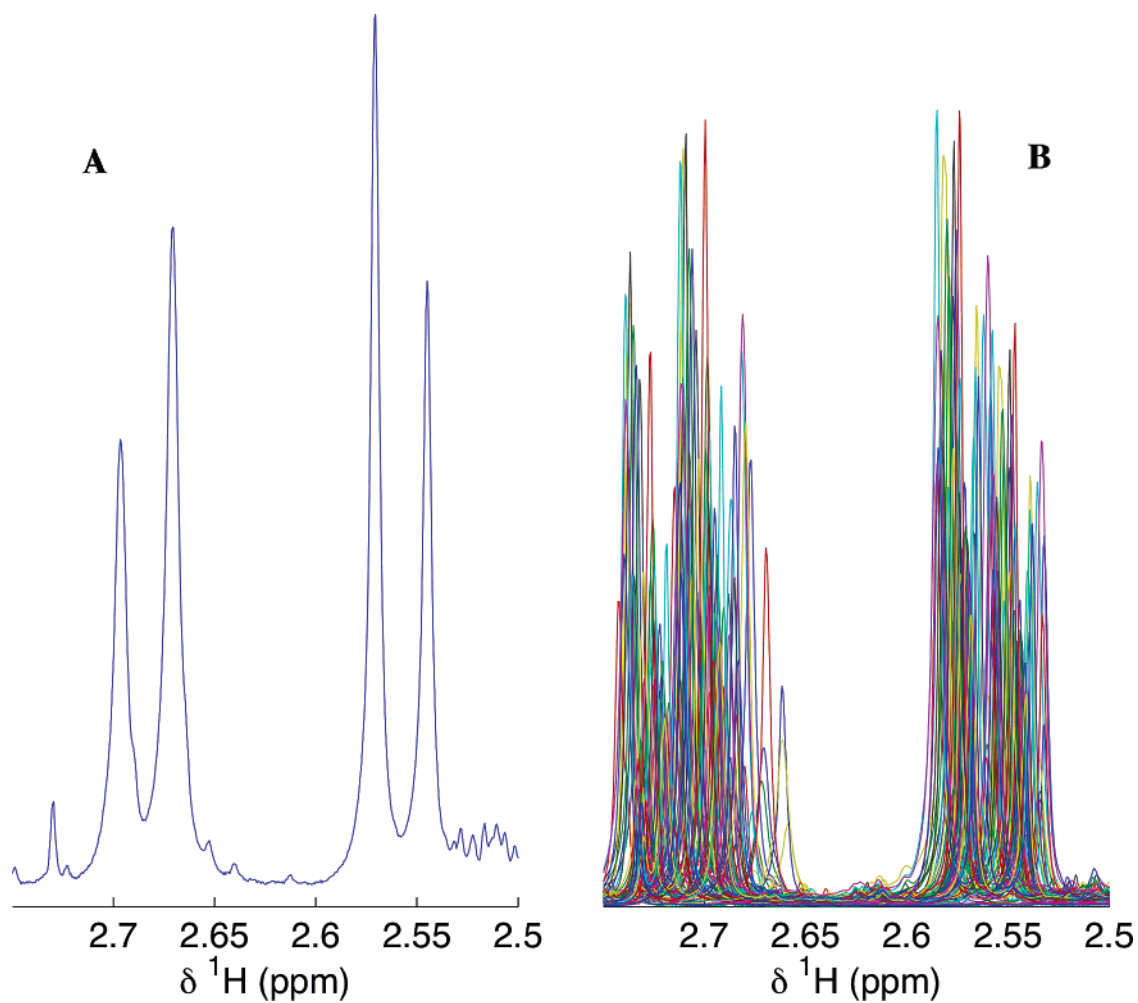


Figure 2. Example of peak position variation in the citrate region: (A) single NMR spectrum; (B) superimposition of 62 NMR spectra.

separated into three parts. The first part contains the variation-common in X and Y , and the last two parts contain the specific

variation for X and Y , respectively. The specific variations for X and Y are so-called “structured noise”.^{13,17} An O-PLS model is thus

Table 1. Summary of the Simulated Data Sets and of Their Respective O-PLS Models

simulation	molecule	relative concn		data-reduced O-PLS model		full-resolution O-PLS model	
		mean ^a	variance	no. of components	Q_Y^2	no. of components	Q_Y^2
1	citrate	75	100				
	taurine	5 (1), 7.5 (2)	4	2	0.78	2	0.79
	TMAO	5	4				
citrate ^b	5 (1), 7.5 (2)	4					
2	taurine	5	4	2	0.72	3	0.49
	TMAO	5	4				
	citrate ^c	5	4				
3	taurine	5	4	5	0.05	3	0.65
	TMAO	5	4				
	citrate ^c	5	4				
4	taurine ^c	5	4	3	0.66	3	0.74
	TMAO	5 (1), 7.5 (2)	4				
	citrate ^c	5 (1), 7.5 (2)	4				
5	taurine	5	4	2	0.76	3	0.72
	TMAO	5	4				
	citrate ^c	5	4				
6	taurine ^b	5	4	2	0.66	3	0.51
	TMAO ^b	5 (1), 7.5 (2)	4				
	citrate ^b	5	4				

^a (1) and (2): respective concentration in groups 1 and 2. ^b Molecule prone to peak position variation not related to the group differences. ^c Molecule prone to a peak position variation differing between the groups.

written as follows:

$$\text{model of } \mathbf{X}: \mathbf{X} = \mathbf{T}\mathbf{W}^t + \mathbf{T}_{\text{Yosc}}\mathbf{P}_{\text{Yosc}}^t + \mathbf{E}$$

$$\text{model of } \mathbf{Y}: \mathbf{Y} = \mathbf{T}\mathbf{C}^t + \mathbf{F}$$

$$\text{prediction of } \mathbf{Y}: \hat{\mathbf{Y}} = \mathbf{T}\mathbf{C}^t$$

\mathbf{T} represents the score matrices for \mathbf{X} and \mathbf{Y} , and \mathbf{W} and \mathbf{C} are the joint orthonormal loading matrices, respectively. \mathbf{E} and \mathbf{F} are the respective residual matrices for \mathbf{X} and \mathbf{Y} . \mathbf{T}_{Yosc} is the score matrix orthogonal \mathbf{Y} , and \mathbf{P}_{Yosc} is the corresponding loading. The superscript t means that the matrix is transposed.

The O-PLS method provides similar prediction to PLS (projection to latent structures). However the interpretation of the models is improved because the structured noise is modeled separately from the variation common to X and Y . Therefore, the O-PLS loading and regression coefficients allow for a more realistic interpretation than PLS, which models the structured noise together with the correlated variation between X and Y . Furthermore, the orthogonal loading matrices provide the opportunity to interpret the structured noise. More details on the differences between the PLS and O-PLS algorithms have been described previously by Trygg et al.¹³ To test the validity of models against over-fitting, the cross-validation parameter Q_Y^2 , was computed:¹⁷

$$Q_Y^2 = 1 - \frac{\sum (\mathbf{T}\mathbf{C}^t - \mathbf{Y})^2}{\sum \mathbf{Y}^2}$$

For the applications presented here, each line of the \mathbf{X} matrix is a NMR spectrum corresponding to one sample and each column of \mathbf{Y} defines a class (or group) whose values are dummy variables. The method can therefore be defined as O-PLS-DA.

Postprocessing and Interpretation. A mean-centered multivariate data set has the mean of each variable within the set subtracted from its value for each sample. An autoscaled multivariate data set is a mean-centered multivariate data set with each value of each variable divided by the standard deviation of the variable. Moreover, in the case of two classes, the value of a O-PLS-DA loading for a variable corresponds to the correlation coefficient between the variable and the classes descriptor.

The method proposed in this paper consists of computing a model based on autoscaled data and then in a first step back-transforming the loadings by multiplying all values by their respective standard deviation. In a second step, the back-transformed loading is plotted using for each point a color corresponding to the weight value in the model that represents the correlation of the X variable with Y . The scale range of colors is set using the maximum and the minimum of the autoscaled model weight. The interpretation of the loading is therefore straightforward for the spectroscopist because the resulting plot provides a loading with the same shape as that of a spectrum (covariance), but on the same plot, the important variables for the discrimination between the classes (correlation) are highlighted by the color code.

In the following application, this method was applied to O-PLS-DA loadings but it can be also used with PLS loadings.

Computer and Software. NMR processing and pattern recognition were carried out using a Power Mac G5 with dual 64-bit 2-GHz processors and 2 GB of synchronous dynamic random access memory. NMR processing and pattern recognition routines were written in-house in the MATLAB 6.5 environment (The Mathworks Inc.).

RESULTS AND DISCUSSION

Observation of Positional Noise. The ¹H NMR spectra of the rat urine samples were acquired as described above and previously.⁸ The samples were chosen such that the differences

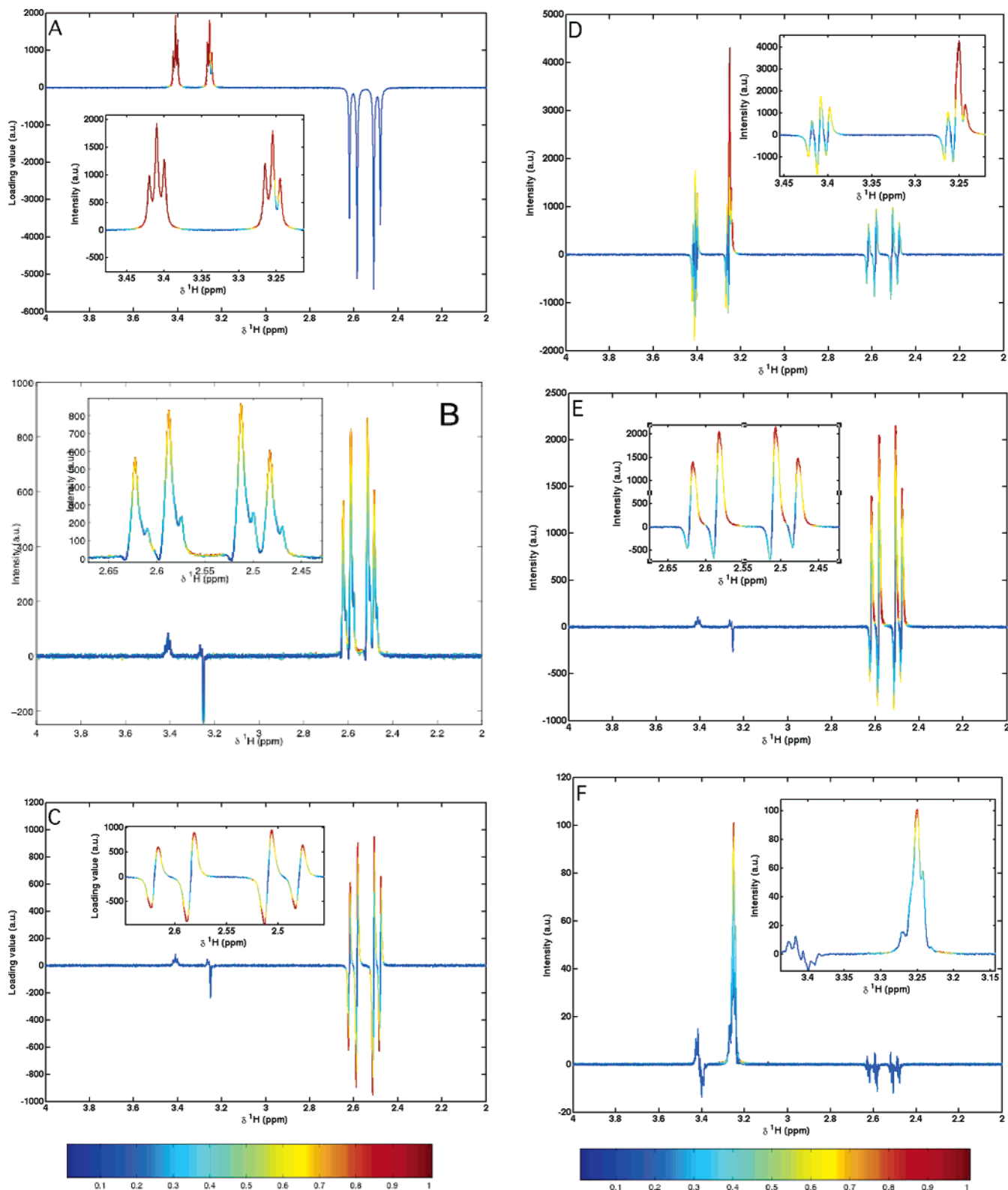


Figure 3. O-PLS loadings for the different simulation: (A) discrimination without peak position variation, (B) discrimination based on a compound subject to peak position variation, (C) discrimination based on the peak position variation, (D) discrimination based on the peak position variation and on a compound subject to peak position variation, (E) discrimination based on a compound overlapping with another which is subject to peak position variation, and (F) discrimination based on a compound, where all the compounds present are subject to peak position variation. The color scale corresponds to the UV model variable weights.

between different time points of urine collection were small and did not affect any classification. After data acquisition, the spectra of the 62 rat urine samples were separated into two groups of 31 spectra, corresponding to the control and treated groups.

Citrate is known to be one of the compounds that have the largest peak position variation, mainly due to pH and metal ion variation. To illustrate this variation, Figure 2A shows the citrate region of a single spectrum of rat urine indicating the usual AB

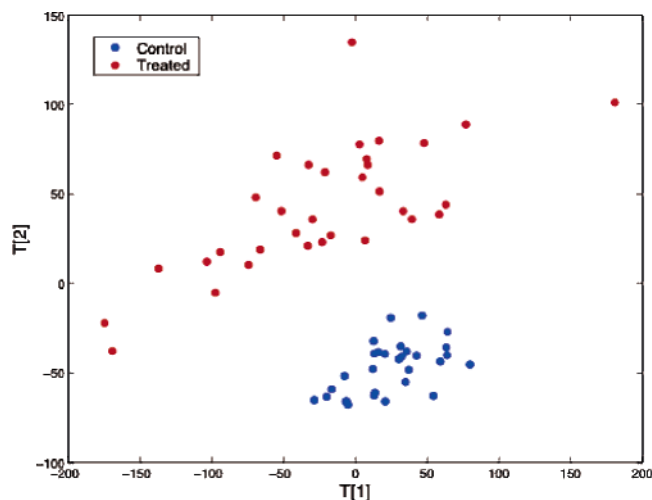


Figure 4. PCA scores and loadings map, representing the distribution of the HgCl₂-treated and control rat urine sample.

Table 2. Summary of the O-PLS Models for the Discrimination between Treated Rat Urine Samples for Both Full and Reduced Data Sets

data type	comp	R ² Xcorr	R ² Xyo	R ² X	R ² Y	Q ² Y
reduced	1	0.34	0	0.34	0.87	0.68
reduced	2	0.32	0.27	0.60	0.93	0.84
full	1	0.24	0	0.24	0.93	0.81
full	2	0.22	0.11	0.33	0.97	0.86

spin pattern. Figure 2B shows the same range for the superimposed 62 rat urine spectra, and it can be seen that the range of variation of the citrate resonances covers ~0.05 ppm. It is obvious that the impact of this peak position variation on pattern recognition models must be important. First, when the number of samples is sufficient, this positional variation might be modeled explicitly and therefore this would allow the use of the full-resolution spectra data set. In this case, detection of low-concentration metabolites, which can have their resonance contributions obscured when combined in a bin with a higher concentration metabolite resonance. Second, the loading and O-PLS coefficients can be difficult to interpret without understanding the impact of the peak position variation on them. For these reasons, a simulation study based on the variation of only three compounds that are subject to peak position variation may provide the information necessary for accurate interpretation of more complex data sets such as in a full toxicological study.

Simulated Spectra Data Set. The aim of the simulation is to understand the influence of the positional noise on chemometric modeling, in particular the O-PLS model (coefficients and prediction ability), in order to assess the possibility of using the full resolution of ¹H NMR spectra in metabonomic studies instead of the reduced data format previously used mainly to tackle the problem of peak position variation. Each simulated NMR spectrum covers the region between δ 2–4, and for each simulation, two groups of 100 spectra each were generated (Table 1). The concentration column shows the composition of the artificial mixtures, and when two values are present, they correspond to the different concentration in the two groups. Two O-PLS models have been computed for each simulation, the first one after data segmentation (i.e., integration into 0.04 ppm width bins), and the

Table 3. Confusion Matrix for the External Validation of the O-PLS Model

		predicted	
		control	Hg treated
actual	control	8	0
	Hg treated	0	8

second one, using the full-resolution spectra. The comparison has been done on the basis of the cross-validation parameter, Q². Although this parameter can be difficult to interpret by itself, especially for discriminant analysis, changes in Q² can be applied to model parameter optimization (i.e., the number of PCA, O-PLS, or PLS components). In this simulation, Q² has been used to compare both reduced- and full-resolution spectra O-PLS models for each case of peak position variation.

The first simulation does not contain any peak position variation. It was generated to show how, using the weight of the variables in combination with a mean-centered model O-PLS loading, it is possible to highlight and identify the significant metabolites directly from the representation of the loading. Here, the concentration of citrate and its variability is much greater than those of taurine and TMAO. The resulting O-PLS loading is presented in Figure 3A, and its interpretation is easy for the spectroscopist because the shape of the different peaks is the same as those in a real NMR spectrum. The taurine resonances can also be recognized easily in this plot. The TMAO does not seem to have any influence on the loading shape because the taurine spectra appear virtually without distortion. Only a localized region at δ 3.27 (blue spot) shows where the TMAO peak overlaps with those from taurine. However, the influence of citrate on the loading shape is important, since high variance introduces a bias in the model if it is not taken into account during the modeling. This example shows the efficiency of the back-scaled loading method in highlighting the real metabolites responsible of the discrimination between the two groups. Finally, because no positional noise was involved in this simulation, both data-reduced and full-resolution models Q² have almost the same value.

For the second simulation, the discrimination is based on a molecule prone to a peak position variation but which is not related to the discrimination between the two groups. The prediction ability of the full-resolution model is reduced by the peak position variation, and an increased number of components is necessary to obtain an optimum Q². For this simulation, the data-reduced O-PLS model shows better characteristics (Q² and number of components) because the binning of the data reduces the peak position variation. However, using the full resolution, the remaining discrimination is enough to enable interpretation and Figure 3B presents the loading derived from this simulation. Despite the fact that the shape of the citrate spectrum is strongly distorted, it is still possible to recognize the AB structure of the spectrum for this molecule (δ 2.54 and δ 2.66). This example shows that, although random positional noise reduces the discrimination between the groups, it is still possible to interpret the model.

When the peak position variation is related to the discrimination, but the concentration of the molecule is not (simulation 3), the resulting O-PLS model provides good prediction abilities. The corresponding loading (Figure 3C) shows the characteristic

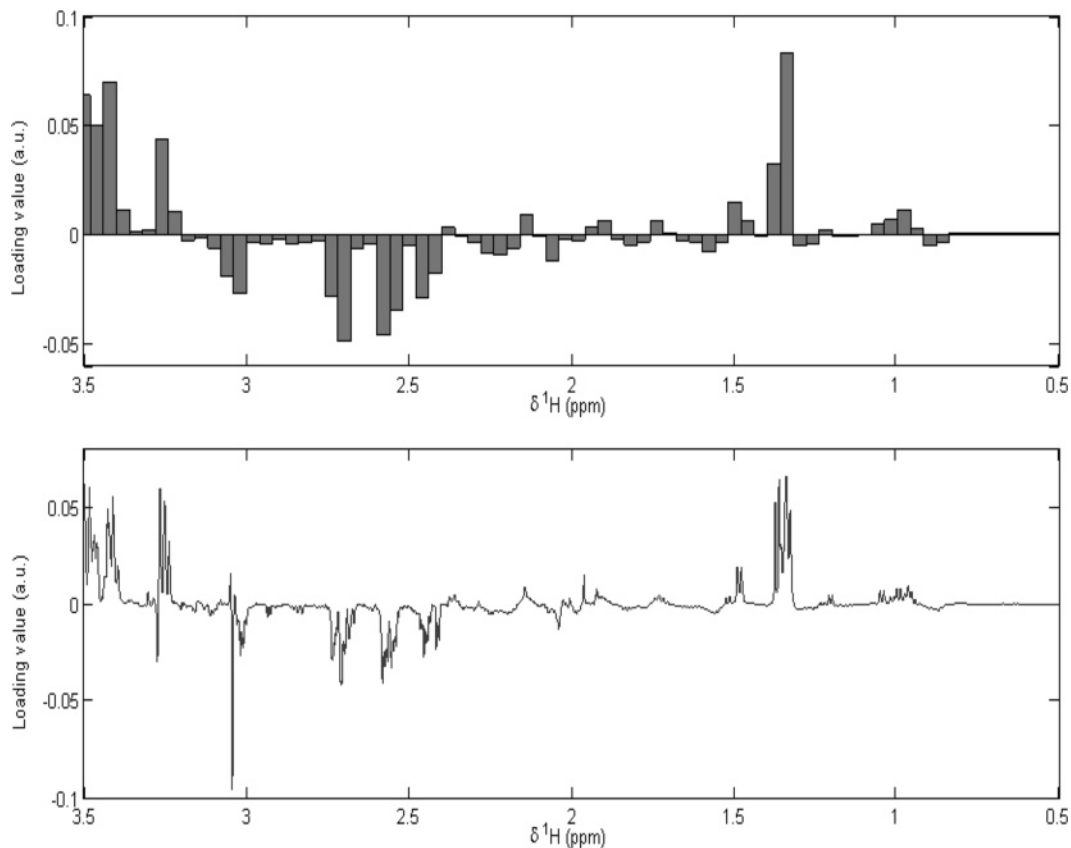


Figure 5. Comparison of the O-PLS loadings obtained with data-reduced and full-resolution NMR spectra.

dispersion-phase signature of chemical shift variation for the citrate peak, where the shape is similar to a first-derivative curve of a peak. Both positive and negative parts of the citrate signal intensity are equivalent, and if the integral of the citrate area is computed, it represents only 3% of the mean intensity of this molecule resonance, showing in this way that the concentration of citrate does not seem to be important in the discrimination between the two groups. Furthermore, the colors representing the weight of the variable on the discrimination are equivalent for both the negative and the positive parts of this signal. This confirms that the citrate concentration does not have any influence on the discrimination. The fact that the peak position variation is different for the two classes means that there exists between the two groups a difference not related to the concentrations of metabolites that give rise to the resonance peaks but rather to the environment of these metabolites. Those environmental differences can have different origins such as a global or specific ionic composition of the medium. In this case, the peak position variation is not only a nonproblem but a supplementary source of information for the characterization of a biofluid. The reduced-data model is unable to discriminate between the groups because the data binning removes the chemical shift variation information by summing over the entire region where the peak position varies. For this reason, although the information related to this peak position variation is important, being indicative of a higher level of calcium for example, it would not be modeled when using the reduced data.

Simulation 4 presents a more complex case where both peak position variation and concentration are responsible for the discrimination between the groups. The predictivity of both O-PLS models is comparable even though using full-resolution data leads

to a better Q^2 . The loading corresponding to the full-resolution model shows a shape similar to the previous simulation for the citrate and taurine peaks (Figure 3D). However, due to the TMAO concentration variation between the two groups, the TMAO peak is present in the loading and the weights of the corresponding variables thus contribute the most to the observed discrimination. For the reduced-data model, the loading does not have enough resolution to identify the singlet of TMAO because the contributions of TMAO and taurine are integrated into the same bin.

For the fifth simulation, the concentration and the peak position of the citrate was varied according to the discrimination between the two groups. In this case, the characteristics of both data-reduced and full-resolution models are very close even though the number of components for the full-resolution model is higher. The shape of the corresponding O-PLS loading presents again a characteristic dispersion-phase signature for the citrate region. However, the color and the intensities show together the impact of the concentration on the model (Figure 3E). The peak position variation still has a nonnegligible influence but according to the color is less important to the discrimination. For the reduced-data model, the information on the peak position variation is lost, as was the case for simulation 3.

The last simulation involves the case where all the compounds are subject to peak position variations independent of the group and the discriminant compounds, where the TMAO is overlapped with taurine. Here, on the base of the O-PLS loading, the different peaks are more difficult to assign. However, assuming some prior knowledge of the position of the peaks, the attribution of TMAO as the discriminant metabolite can be deduced.

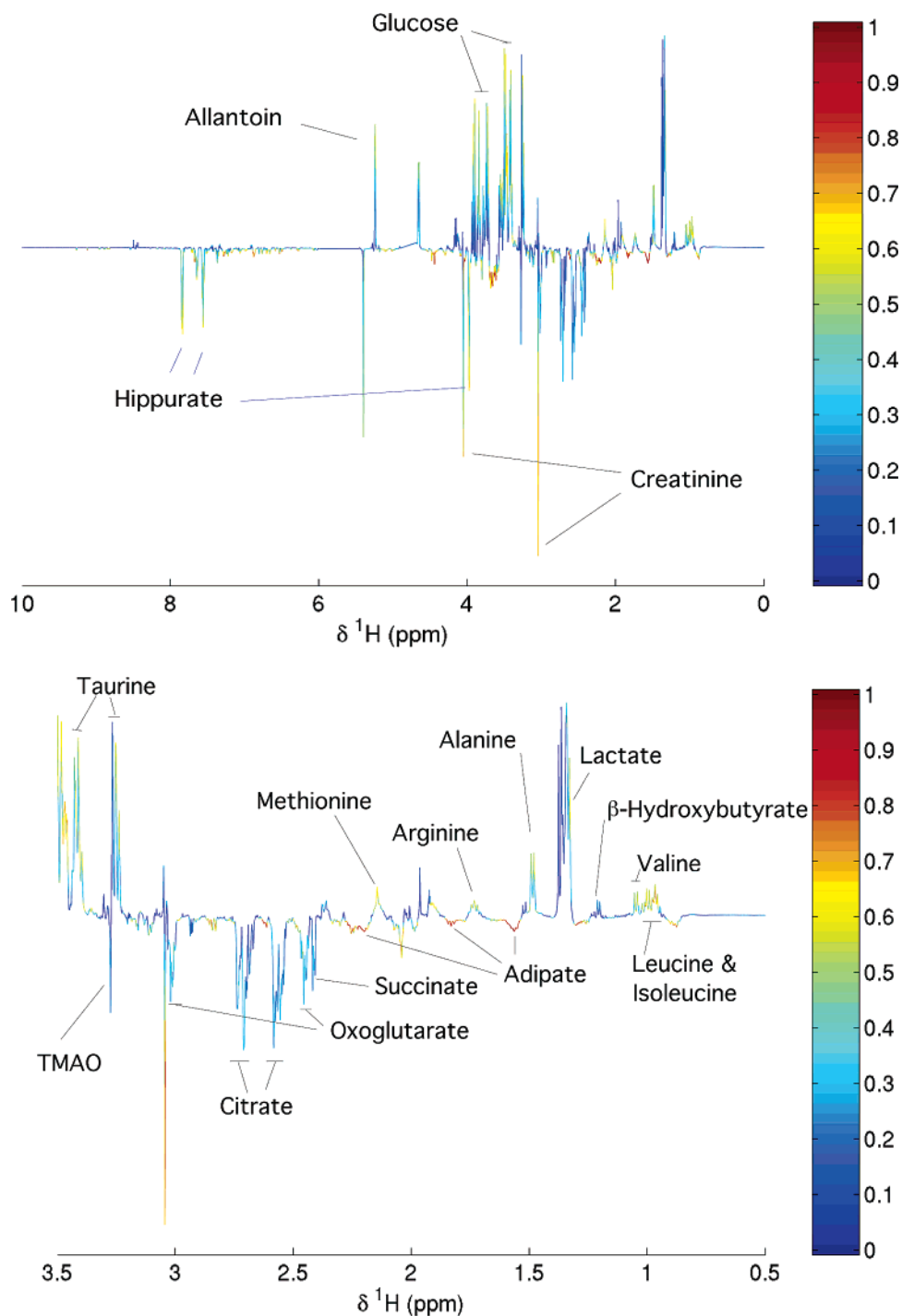


Figure 6. O-PLS loading related to the discrimination between HgCl_2 -treated and control rat urine.

These simulations show that it should be possible to use the full NMR spectra of biofluids instead of reduced spectra in order to carry out pattern recognition. Although the interpretation sometimes requires more effort, the procedure involving the back-transformation of the loading and its representation according to the weight of the variable in the models provides simplified interpretation and allows high-resolution pattern recognition. Other simulations have been carried out involving group sizes of 50, 20, 10, and 5. These showed that for 20 or more samples per group the peak position variation can be modeled properly, allowing interpretation of the O-PLS models. However, for only five samples per group and when the discrimination is based only

on a compound subject to a strong peak position variation, the model loses its prediction abilities and its interpretability. Finally, for 10 samples, the quality of the models really depends on the peak position variation intensity, especially when the discrimination is based on peaks subject to position variation. However, interpretation of metabolomic studies on so small a number of samples is discouraged because of the inherent biological variation, even if the NMR data are reduced or not.

An Example from a Toxicological Study. The PCAs of the ^1H NMR spectra from the rat urine samples, both with and without data reduction, show a clear discrimination between both treated and control groups (Figure 4). Unfortunately, the loading inter-

pretation is difficult because of the large number of variables. For this reason, a representation of the individual loadings as spectra can be more useful for interpretation by spectroscopists. The PCA scores and loadings are computed according to the main variation in the **X** matrix, and for this reason, the loadings are not fully related to the discrimination between the groups and so their interpretation cannot be straightforward. It is therefore necessary at least to combine all loadings according to the direction between the centers of gravity of the two groups to obtain a discriminant loading, which is the basis of principal component discriminant analysis (PC-DA). However, the properties of the O-PLS-DA method provide a more comprehensive description of the discrimination between classes.

The predictive components of an O-PLS model define a subspace of variation in *X* related to *Y*, for discriminant analysis between *k* classes, the predictive components define a subspace of *k* - 1 dimensions. Therefore, for our two groups of urine samples, one loading is enough to interpret the difference between the two classes. It is equivalent to a combination of principal components, but this loading is computed with the full data set and is more accurate than the discriminant loading obtained with the PC-DA. Furthermore, the structured noise can also be modeled with this method, which is not the case with PC-DA. The O-PLS model summaries for the reduced and nonreduced ¹H NMR data set are provided in Table 2.

The use of 7-fold cross-validation indicated that two O-PLS components were appropriate for the both models, a predictive one and a **Y**-orthogonal one. The prediction abilities of the models estimated through the cross-validation are excellent for both reduced and nonreduced ¹H NMR data. Moreover, an external validation step has been also carried out using 25% of the spectra¹⁶ as a prediction set and 75% as a training set, and the results are displayed as a confusion matrix (Table 3), confirming the excellent prediction abilities of the O-PLS models built on this data set. This shows, in this case, that the peak position variation does not affect the prediction ability of the O-PLS models. For the nonreduced data set, the O-PLS model also shows that 22% of the variation of **X** after autoscaling is related to **Y** and only 33% of the variation of **X** after autoscaling is needed to obtain the best prediction ability through the cross validation.

The comparison between the O-PLS loadings obtained from full spectra and data-reduced spectra clearly demonstrates the benefit of using as high a spectral resolution as possible (Figure 5). For example, the complex of peaks around δ 1.3 ppm is reduced to only two variables in the reduced spectra, which can lead to imprecise or incomplete interpretation if the concentration variations of these different metabolites are anticorrelated.

The O-PLS model loading is interpreted using the method that combines the back-transformed loading and the variable weight. To get a clearer visualization, the part of the loading between δ 0.5 and 3.5 is enlarged in Figure 6. Different biomarkers can be identified easily on this plot, and for some of them, the variation between the control and the treated group is summarized in Table 4. The combination of the back-transformed loading with the variable weight allows the characterization of the variation of a metabolite with two parameters, its concentration variation and its discrimination weight (square of the correlation coefficient). The interpretation of the biological consequences of HgCl₂-induced

Table 4. Predicted Example of Metabolite Variation According to the HgCl₂ Treatment

metabolite	mean concn variation ^a	discrimination weight
creatinine	--	0.78
adipate	-	0.89
citrate	--	0.43
lactate	++	0.57
taurine	++	0.47
β -hydroxybutyrate	+	0.22

^a Key: ++, major increase; --, major decrease; +, minor increase; -, minor decrease.

toxicity is beyond the scope of the current paper. Therefore, only six of the many perturbed metabolites have been selected for the purpose of illustration as depicted in Table 4.

It should also be noted that, without including the peak corresponding to creatinine, the other metabolites with the highest concentration are not the ones that provide the more important weight in the discrimination. Citrate has a low discrimination weight and is affected by a peak position variation not related to the difference between the control and the treated group.

CONCLUSION

In this paper, we have demonstrated that metabolomic studies can be carried out without any reduction of the true ¹H NMR spectral resolution. From simulation studies, variation of peak position can reduce slightly the prediction ability of the model but at the same time this peak position variation can be modeled by the pattern recognition method and therefore may provide useful information about physicochemical variations in the biofluid matrix, thereby offering an improvement over realignment methods this result is not only important for NMR-based studies but also for any other analytical techniques (i.e., LC-MS) where peak position variation is often considered as problematic. The complexity of the loading interpretation due to the high number of variables can be simplified using O-PLS, which reduces the number of loadings to be interpreted and allows their representation and interpretation in the same format as a spectrum. Finally, combining back-transformed loading of an autoscaled model for each variable with its corresponding weight in the same plot is very useful for accurate interpretation of chemometric models and allows more sensitivity in the detection of metabolic perturbation biomarkers.

ACKNOWLEDGMENT

O.C., M.E.D., A.C., R.H.B., J.C.L., J.K.N. and E.H. are members of Biomedical Sciences, Faculty of Medicine, Imperial College London, The authors acknowledge gratefully the financial support from the Wellcome Trust for the Biological Atlas of Insulin Resistance (BAIR) Consortium and helpful discussion with members of the BAIR project team (www.bair.org.uk).

Received for review August 12, 2004. Accepted October 27, 2004.

AC048803I