Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemometrics and Intelligent Laboratory Systems 73 (2004) 139–149
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Statistical experimental design and partial least squares regression analysis of biofluid metabonomic NMR and clinical chemistry data for screening of adverse drug effects

Henrik Antti[a,*], Timothy M.D. Ebbels[b], Hector C. Keun[b], Mary E. Bollard[b], Olaf Beckonert[b], John C. Lindon[b], Jeremy K. Nicholson[b], Elaine Holmes[b]

[a] *Metabometrix Ltd, RSM. Prince Consort Road, London SW7 2BP, UK*
[b] *Biological Chemistry, Biomedical Sciences Division, Faculty of Medicine, Imperial College of Science Technology and Medicine, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, UK*

Received 24 July 2003; accepted 26 November 2003
Available online 11 March 2004

Abstract

Metabonomic analysis is increasingly recognised as a powerful approach for delineating the integrated metabolic changes in biofluids and tissues due to toxicity, disease processes or genetic modification in whole animal systems. When dealing with complex biological data sets, as generated within metabonomics, as well as related fields such as genomics and proteomics, reliability and significance of identified biomarkers associated with specific states related to toxicity or disease are crucial in order to gain detailed and relevant interpretations of the metabolic fluxes in the studied systems. Since various physiological factors, such as diet, state of health, age, diurnal cycles, stress, genetic drift, and strain differences, affect the metabolic composition of biological matrices, it is of great importance to create statistically reliable decision tools for distinguishing between physiological and pathological responses in animal models. In the screening for new biomarkers or patterns of pathological dysfunction, methods providing statistically valid measures of effect-related changes will become increasingly important as the data within areas such as genomics, proteomics and metabonomics continues to grow in size and complexity. ^{1}H NMR spectroscopy and mass spectrometry are the principal analytical platforms used to derive the data and, because extensively large data sets are required, as much consideration has to be given to optimum design of experiments (DoE) as for subsequent data analysis. Thus, statistical experimental design combined with partial least squares (PLS) regression is proposed as an efficient approach for undertaking metabonomic studies and for analysis of the results. The method was applied to data from a liver toxicology study in the rat using hydrazine as a model toxin. 1D projections of 2D J-resolved (J-RES) ^{1}H NMR spectra and the corresponding clinical chemistry parameters of blood serum samples from control and dosed rats (30 and 90 mg/kg) collected at 48 and 168 h post dose were analysed. Confidence intervals for the PLS regression coefficients were used to create a statistical means for screening of biomarkers in the two combined data blocks (NMR and clinical chemistry data). PLS analysis was also used to reveal the correlation pattern between the two blocks of data as well as the within the two blocks according to dose, time and the interaction dose × time.

Keywords: NMR; Clinical chemistry; PLS analysis; Design of experiments; Metabonomics

1. Introduction

Within the area of metabonomics, defined as the multiparametric metabolic response of living systems to pathophysiological stimuli or genetic modification, biofluids and tissues are analysed by the sequential combination of high field NMR spectroscopy and multivariate chemometric techniques. The metabonomic approach has proved an efficient technique in characterising and predicting the nature and target organ of toxicity for a large number of different xenobiotics [1–6].

High-resolution ^{1}H NMR spectroscopy is an efficient and nondestructive tool for generating data on a multitude of metabolites in biofluids or tissues [7]. The acquired spectral profile of a biofluid reflects the metabolic status of the organism, which alters in response to stressors in order to maintain a homeostatic balance [8].

Information recovery, in terms of relationships between the NMR spectral profiles and their biochemical interpretation, can be maximised by applying multivariate statistical

* Corresponding author. Tel.: +44-20-7594-3541; fax: +44-20-7594-6818.
E-mail address: h.antti@metabometrix.com.

0169-7439/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2003.11.013

tools to the analysis of these complex, information-rich, NMR data. As previously shown, ^{1}H NMR spectroscopy of complex biological samples, coupled with multivariate statistical analysis methods provides an alternative in vivo approach to the investigation of drug induced toxicity, altered gene function and also for disease diagnosis. Applications of this metabonomic technology include the identification of biomarkers of toxicity and disease [9,10], monitoring of time related metabolic perturbations in biofluids and tissues following toxic insult [11–13] and metabolic characterisation of physiological variance in humans under mild physiological stress [14].

One important aspect of the metabonomic approach is reliability and significance of the identified biomarkers associated with specific metabolic states. Since various physiological factors, such as diet, state of health, age, diurnal cycles, stress, genetic drift and strain differences, affect the metabolic composition of biological samples, it is of great importance to create statistically reliable decision tools for distinguishing between physiological and pathophysiological responses in animal models. In metabonomic data, a drug-induced biomarker, is by definition linked to a change with dose or with both dose and time. In the screening for new biomarkers or patterns of toxic response, methods providing statistically significant measures of effect-related changes will become more important as genomic, proteomic and metabonomic data sets grow in size and complexity. Such methods will also be a prerequisite in the work towards combining blocks of data generated with different analytical techniques or within different areas of science, e.g. combining genomic, proteomic and metabonomic matrices, in order to extract relevant patterns, enhance interpretation and hence facilitate the understanding of processes and mechanisms induced by toxicity and disease.

Statistical experimental design or design of experiments (DoE) [15,16] is a powerful tool for defining the effect of one or more variables on a set of measured responses by using multiple linear regression (MLR) or generalised regression methods such as partial least squares (PLS) [17,18]. Confidence intervals for the calculated regression coefficients create a means for understanding the significance of the variables and the interactions between variables on the measured responses. DoE also provides a strict mathematical framework for changing all pertinent experimental variables simultaneously and independently of each other, and achieve this in the smallest possible number of experimental runs. The strength of DoE, compared to univariate approaches considering only one variable at the time, is its ability to detect and estimate nonadditive variable interactions as well as providing a higher precision to estimates of the variables' effects on the measured responses.

Here we have evaluated DoE combined with PLS regression as a multivariate metabonomic screening tool for large biological data sets with the aim of biomarker identification, toxic response detection and interpretation of variable correlations within and between blocks of complex multiparametric data based on statistical significance. Hydrazine is a well-documented hepatotoxin that induces steatosis and the effects of hydrazine administration to animal models on both conventional clinical parameters and the ^{1}H NMR biofluid profiles have been reported [19].

The type of NMR experiment used influences the visibility of different components of the biofluid profile. Plasma spectra contain a mixture of high molecular components such as lipoproteins which generate broad resonances on which are superimposed sharper resonances from the low molecular weight components such as organic and amino acids. Use of a spin echo experiment such as the Carr-Purcell-Meiboom-Gill (CPMG) or J-resolved results in the suppression of the broader elements and therefore enhances visualization of the low molecular weight metabolites. Conversely diffusion edited pulse sequences can be used to enhance the small molecule profiles [20]. Here we have chosen to use J-resolved projections to characterise the effect of hydrazine on the low molecular weight profile since hydrazine is known to induce modifications in several low molecular weight species present in the plasma profile.

The application to the combined blood serum 2D J-resolved (J-RES) NMR data and clinical chemistry parameters for a hydrazine dose study in the rat exemplified the versatility of the proposed DoE–PLS methodology in toxicology screening and metabolic profiling.

All data presented here were acquired within the COnsortium for MEtabonomic Toxicology (COMET). COMET is an academic project involving five major pharmaceutical companies (BMS, Eli Lilly, Hoffman La Roche, Novo Nordisk, Pfizer) and Imperial College. Its ultimate goal is to build expert systems capable of predicting the toxicity of candidate drug compounds. This is being achieved through the construction of a database of ~ 100,000 ^{1}H NMR spectra of biofluids from toxicological studies. Multivariate statistical data mining techniques are being used to build mathematical models of the NMR data for classifying new samples according to their most likely site or mechanism of toxicity, and also to discover new biomarkers of these effects. Results of studies on individual toxins will be published elsewhere.

2. Methods

2.1. Animal studies

Male 8–10 week Sprague–Dawley (SD) rats were randomly assigned to dose groups (1—control (saline), 2—low dose (hydrazine, 30 mg/kg) and 3—high dose (hydrazine, 90 mg/kg)). Serum samples were collected at 48 and 168 h post treatment.

2.2. Acquisition of 2D J-resolved (J-RES) ^{1}H NMR serum spectra

All ^{1}H NMR 2D J-resolved (J-RES) serum spectra [21] were measured at 600.13 MHz ^{1}H NMR frequency and 300

K on a Bruker DRX-600 using the BEST™ flow-injection system (Bruker Efficient Sample Transfer, Bruker Biospin, Rheinstetten, Germany) for sample delivery. The spectra were acquired according to a standard procedure [22] in order to focus on the small molecules present in the serum. One-dimensional sum projections on to the chemical shift axis were calculated from the 2D-JRES spectra and used for subsequent analysis. The assignment of ^{1}H NMR urine spectra and the ^{1}H NMR 2D J-resolved (J-RES) serum spectra were made with reference to published literature data [22,23].

2.3. Clinical chemistry measurements

The serum samples were also characterised by measurement of the following clinical chemistry parameters: blood urea nitrogen, alanine aminotransferase (ALT), aspartate aminotransferase (AST), alkaline phosphatase (ALP), gamma glutamyl transferase (GGT), sodium, potassium, calcium, phosphate, albumin, total protein and total bilirubin.

2.4. Data reduction and pattern recognition

Each NMR spectrum was reduced to 245 integrated regions of equal width (0.04 ppm) corresponding to the region δ 0.2–10.0 using AMIX (version 2.5.9, Bruker) and the data collated into a single data table. The region (δ 4.50–5.98) was deleted to remove any spurious effects of variability in the suppression of the water resonance and any consequent chemical exchange effects on the urea signal. Finally, all spectra were normalised to a constant integrated intensity.

Multivariate analysis was performed using the MODDE software (version 5.0, Umetrics, Umeå, Sweden). Prior to data analysis, the NMR and clinical chemistry data were mean centered followed by scaling to unit variance, in which the variable mean was subtracted from each variable (column of the data) and then each variable was divided by its standard deviation.

2.5. DoE–PLS method

Design of experiments (DoE) was applied separately to the two study data sets in order to create the possibility of investigating and statistically validating the effect of dose and time on the metabonomic NMR patterns and the changes in serum clinical chemistry parameters.

The idea of using DoE was to systematically vary the two parameters (variables), dose and time, known to affect the outcome of the metabolic evolution pattern, independently of each other. The matrix of designed experiments (**X**) was then correlated to the corresponding combined matrix of NMR spectra and clinical chemical parameters (**Y**) for the hydrazine serum study by using PLS. For the calculated PLS model, a set of weights (***w***, ***c***) were calculated component wise for the two data matrices, **X** and **Y**. The **X**-weights (***w***) described the importance of the *X*-variables (dose, time) and the interaction between the two (dose × time) for correlating to **Y** and similarly the **Y**-weights (***c***) described the *Y*-variables (NMR spectral regions and clinical chemistry parameters) which were important in maximising the covariation between **X** and **Y** [16,17]. Consequently, the effect of each variable (dose and time) on the metabolic changes described by NMR data and the clinical chemistry data was investigated at all levels of the other variable included in the design. This allowed calculation of PLS regression coefficients (*b*) for the effects of each variable (dose and time) as well as the interaction effect (dose × time) between the included variables. A significant interaction effect between two variables implies that the effect of one variable is dependent on the setting of the other variable, which is defined as a nonadditive relationship between the variables. The statistical significance of the metabolic changes according to dose and time were expressed as 95% confidence intervals for the PLS regression coefficients *(b)* calculated from the residual standard deviation in **Y** based on the *t*-distribution. The calculated confidence intervals were then used as selective criteria in screening for significant changes in the NMR data and the clinical chemistry parameters. One advantage of the approach was the possibility of characterising each NMR spectral region and each clinical chemistry parameter as nonsignificant or significant according to dose, time, dose × time or combinations thereof and thereby improving the understanding of the complex time-related metabolic variable patterns.

The design for the hydrazine serum study was set to vary in time between the two time points 48 and 168 h post dose and in hydrazine dose concentration between the three concentrations 0, 30 and 90 mg/kg. This gave a total of six experimental settings and using four replicates provided a data set (**X**) of 24 observations for PLS model calculations, where **X** was regressed against the multiblock **Y**-matrix (the combined matrix of the reduced J-resolved NMR data and the corresponding clinical chemistry parameters measured on the blood serum samples) (Fig. 1).

2.6. Correlation of designed data (X) and two combined blocks of data (Y) (metabonomic ^{1}H J-resolved NMR serum data and clinical chemical data) using PLS

The combined **Y**-matrix for the hydrazine serum data consisting of the J-resolved NMR data and the corresponding clinical chemical parameters allowed intra- and inter-block correlation and covariation studies based on the calculated PLS weights (***w***, ***c***) and regression coefficients (*b*).

By interpretation of the PLS weights plot ($\boldsymbol{w} \times \boldsymbol{c}_1 / \boldsymbol{w} \times \boldsymbol{c}_2$), the correlation and covariation between the variables from the different blocks according to variations in dose and time were revealed. Hence, variables in the two blocks responding in the same way to toxic insult would cluster together in the plot. The extracted PLS regression coefficients (*b*) with corresponding 95% confidence intervals were used for interpretation of the dose–time response as well as for judging the

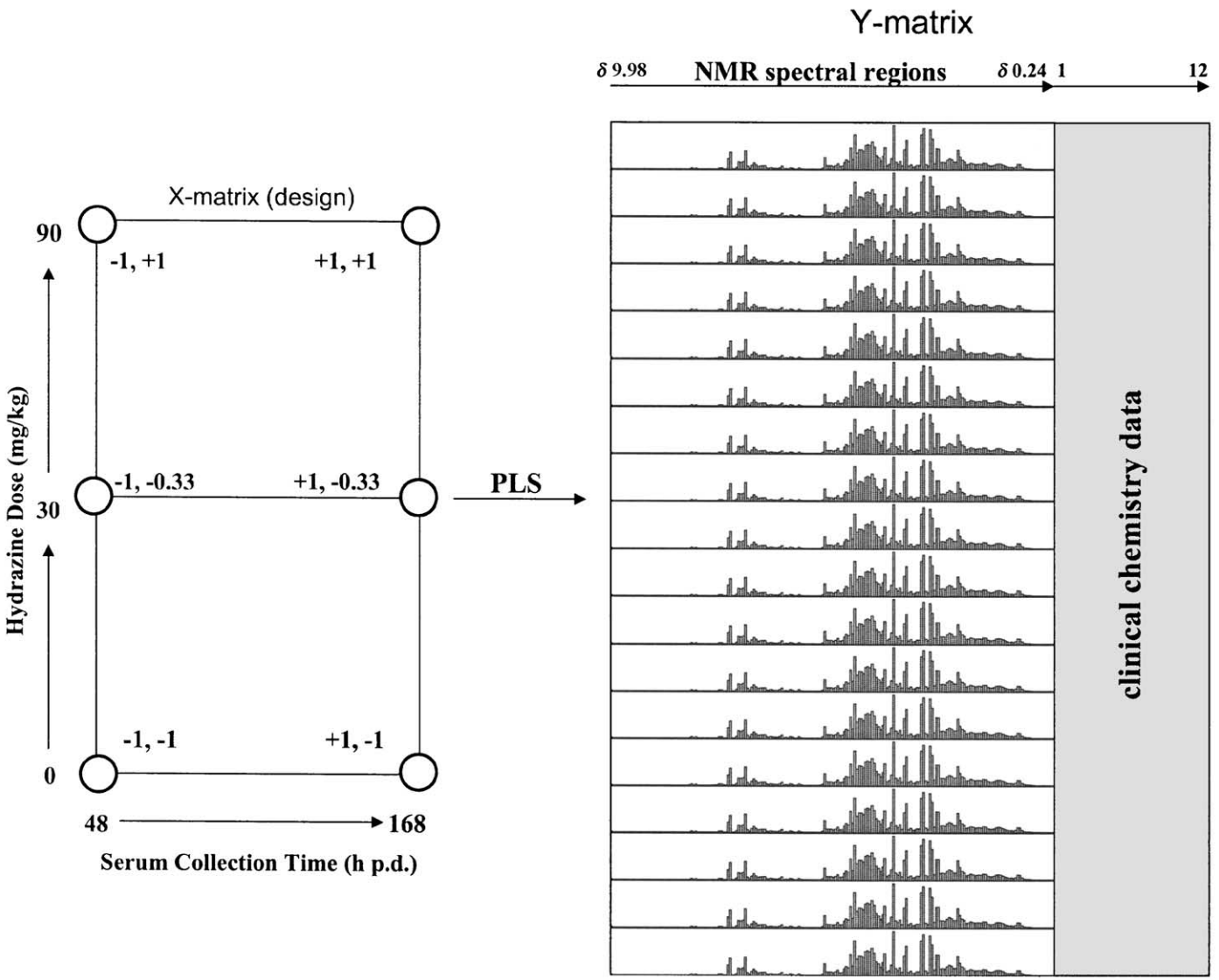

Fig. 1. Scheme explaining the experimental design and modelling strategy for the hydrazine serum study.

significance of each single variable according to dose, time and the interaction dose × time.

2.7. Algorithms

2.7.1. PLS algorithm

The PLS NIPALS algorithm [24–27] used to extract the PLS components can be described as follows,

$$\boldsymbol{w} = \boldsymbol{X}'\,\boldsymbol{u}/(\boldsymbol{u}'\,\boldsymbol{u}) \tag{1}$$

$$\boldsymbol{w} = \boldsymbol{w}/\|\boldsymbol{w}\| \tag{2}$$

$$\boldsymbol{t} = \boldsymbol{X}\boldsymbol{w} \tag{3}$$

$$\boldsymbol{c} = \boldsymbol{Y}'\,\boldsymbol{t}/(\boldsymbol{t}'\,\boldsymbol{t}) \tag{4}$$

$$\boldsymbol{u} = \boldsymbol{Y}\boldsymbol{c}/(\boldsymbol{c}'\,\boldsymbol{c}) \tag{5}$$

$$\boldsymbol{p} = \boldsymbol{X}'\,\boldsymbol{t}/(\boldsymbol{t}'\,\boldsymbol{t}) \tag{6}$$

$$\boldsymbol{E}(\boldsymbol{X} - \text{residual}) = \boldsymbol{X} - \boldsymbol{t}\boldsymbol{p}'$$
$$\boldsymbol{F}(\boldsymbol{Y} - \text{residual}) = \boldsymbol{Y} - \boldsymbol{t}\boldsymbol{c}' \tag{7}$$

where **X** is the predictor matrix (design matrix), **Y** is the response matrix, $\boldsymbol{w}$ is the PLS-weight vector for **X**, $\boldsymbol{u}$ is the PLS-score vector for **Y**, $\boldsymbol{t}$ is the PLS-score vector for **X**, $\boldsymbol{c}$ is the PLS-weight vector for **Y** also used to calculate the **Y**-residual matrix, **F**, and $\boldsymbol{p}$ is the PLS-loading vector for **X** used in the calculation of the **X**-residual matrix, **E**. A matrix or vector followed by $'$ indicates a transposed matrix or vector.

The PLS regression coefficients (b) can then be calculated:

$$b = \boldsymbol{w}(\boldsymbol{p}'\,\boldsymbol{w})^{-1}\boldsymbol{c}' \tag{8}$$

2.7.2. Calculation of confidence intervals for the PLS regression coefficients

Confidence intervals (>95%) were calculated for the PLS regression coefficients according to the following formula [16],

$$\sqrt{(\boldsymbol{X}'\,\boldsymbol{X})^{-1}} \times \text{RSD} \times \boldsymbol{t}(\alpha/2, \text{DF}_{\text{resid}}) \tag{9}$$

where RSD is the standard deviation of the **Y**-residual, $\boldsymbol{t}$ is the tabulated t-value for a given significance level (α) and the degrees of freedom for the **Y**-residual (DF_{resid}). $^{-1}$ refers to a matrix inversion.

3. Results

3.1. Correlation of designed data (X) and two combined blocks of data (Y) (metabonomic ^{1}H J-resolved NMR serum data and clinical chemical data) using PLS

The weights plot ($\boldsymbol{w} \times \boldsymbol{c}_1/\boldsymbol{w} \times \boldsymbol{c}_2$) for the calculated PLS model (Fig. 2) showed that the number of significant

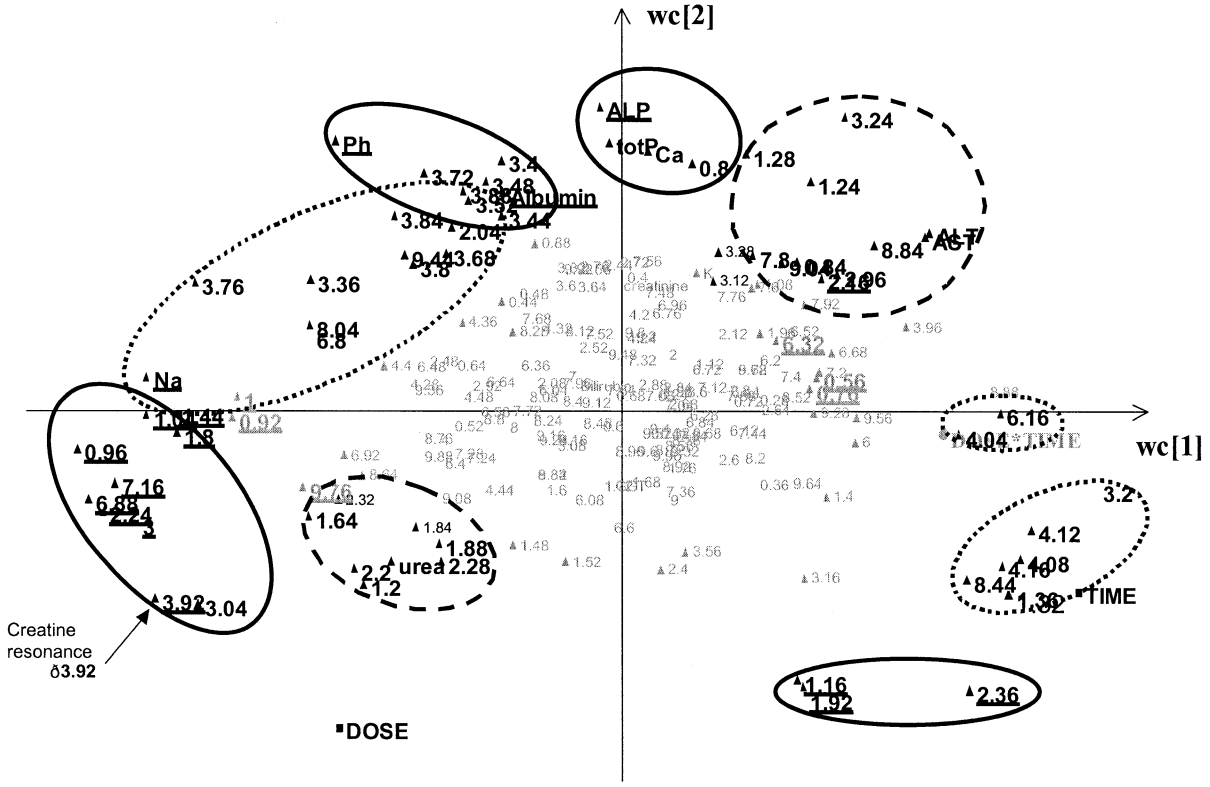

Fig. 2. PLS weights plot ($\boldsymbol{w} \times \boldsymbol{c}_1/\boldsymbol{w} \times \boldsymbol{c}_2$) revealing the model correlation structure for the hydrazine blood serum study. Variables divided into the following classes: nonsignificant (grey), significant according to: dose (dashed line), time (dotted line), dose and time (solid line), dose × time (underlined).

variables was much smaller compared to the original number of acquired variables. Based on the PLS weights plot, interpretations were made of the correlation structure in the data.

To exemplify the interpretations that could be made in terms of correlation and significance from the PLS weights, a few variables were chosen for more detailed analysis. Looking at the spectral region at δ 3.92 (creatine), located in the lower left quadrant of the PLS weights (R2Y = 0.83, Q2 = 0.6) high positive correlation was seen with the other region associated with creatine at δ 3.04 and to the spectral regions circled with a solid line, implying that all these spectral regions have a similar response to hydrazine treatment (dose) with time. Positive correlation with the hydrazine dose can also be seen suggesting that the intensities of these spectral regions increase when hydrazine is dosed. To investigate the correlations connected to δ 3.92, a line was fitted which connected δ 3.92 with the origin of the plot. By projecting all other variables perpendicular to this line, the correlation to δ 3.92 could be decided based on the distance

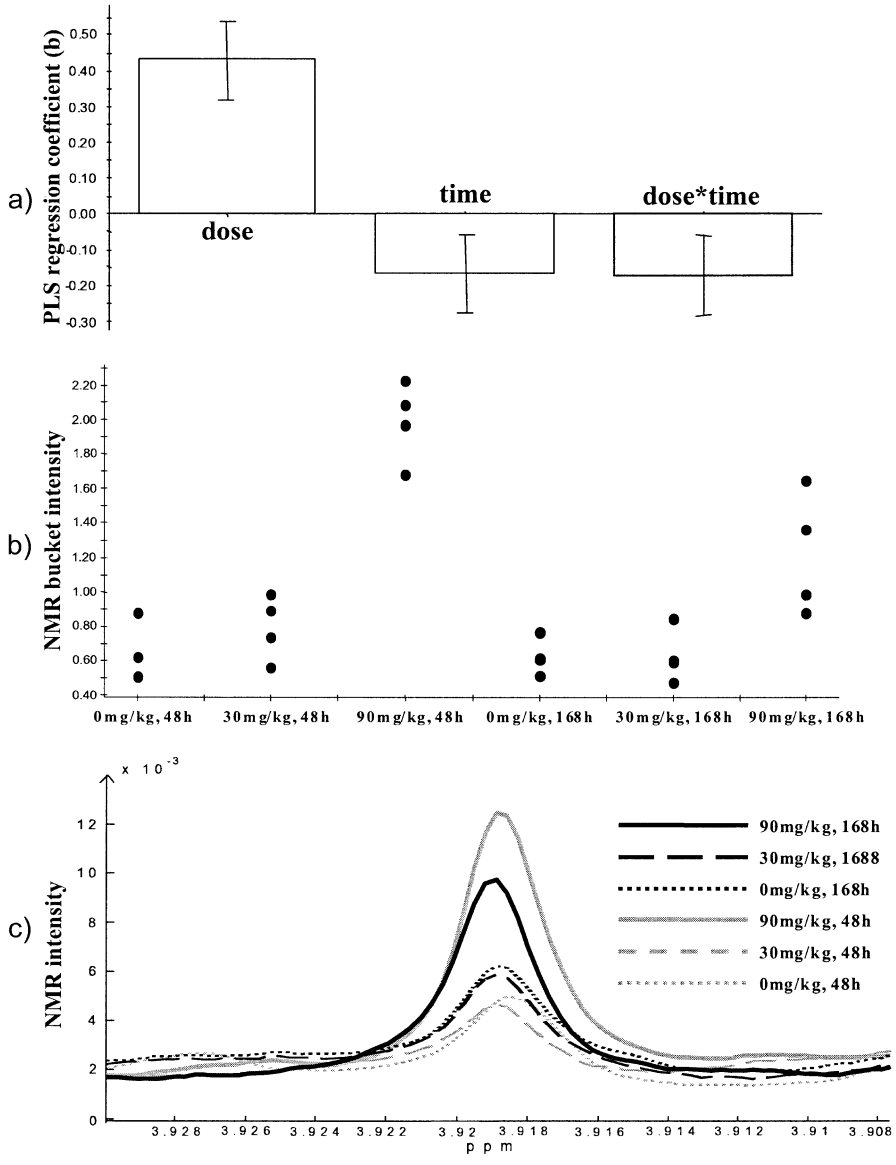

Fig. 3. (a) PLS regression coefficients with 95% confidence intervals for the spectral region (δ 3.92) associated with a creatine resonance. Coefficient for: dose, time and dose × time. (b) Replicate plot for the spectral region (δ 3.92) showing peak intensity versus dose and time for the included samples. (c) Normalised ^{1}H NMR spectra (δ 3.932–3.908) for six representative serum samples, one from each experimental setting in the design, 0 mg/kg, 48 h (grey dotted line), 30 mg/kg 48 h (grey dashed line) and 90 mg/kg 48 h (grey solid line), 0 mg/kg 168 h (black dotted line), 30 mg/kg 168 h (black dashed line) and 90 mg/kg 168 h (black solid line).

from the projections to the origin (leverage). Projection of the *X*-variables dose and time on to the imaginary line showed that δ 3.92 was positively correlated with hydrazine dose but negatively correlated with time. The PLS regression coefficients (*b*) for δ 3.92 (Fig. 3a) verified the positive correlation with dose and the negative correlation with time and also provided evidence for their significance (at >95% confidence level). In addition, the interaction dose × time was significantly negatively correlated to δ 3.92. An explanation to this could be found by looking in the replicate plot (Fig. 3b), where the effect of the high dose at both time points (48 and 168 h) was evident compared with the replicate variation but where it was also evident that this effect of hydrazine dosing decreased between 48 and 168

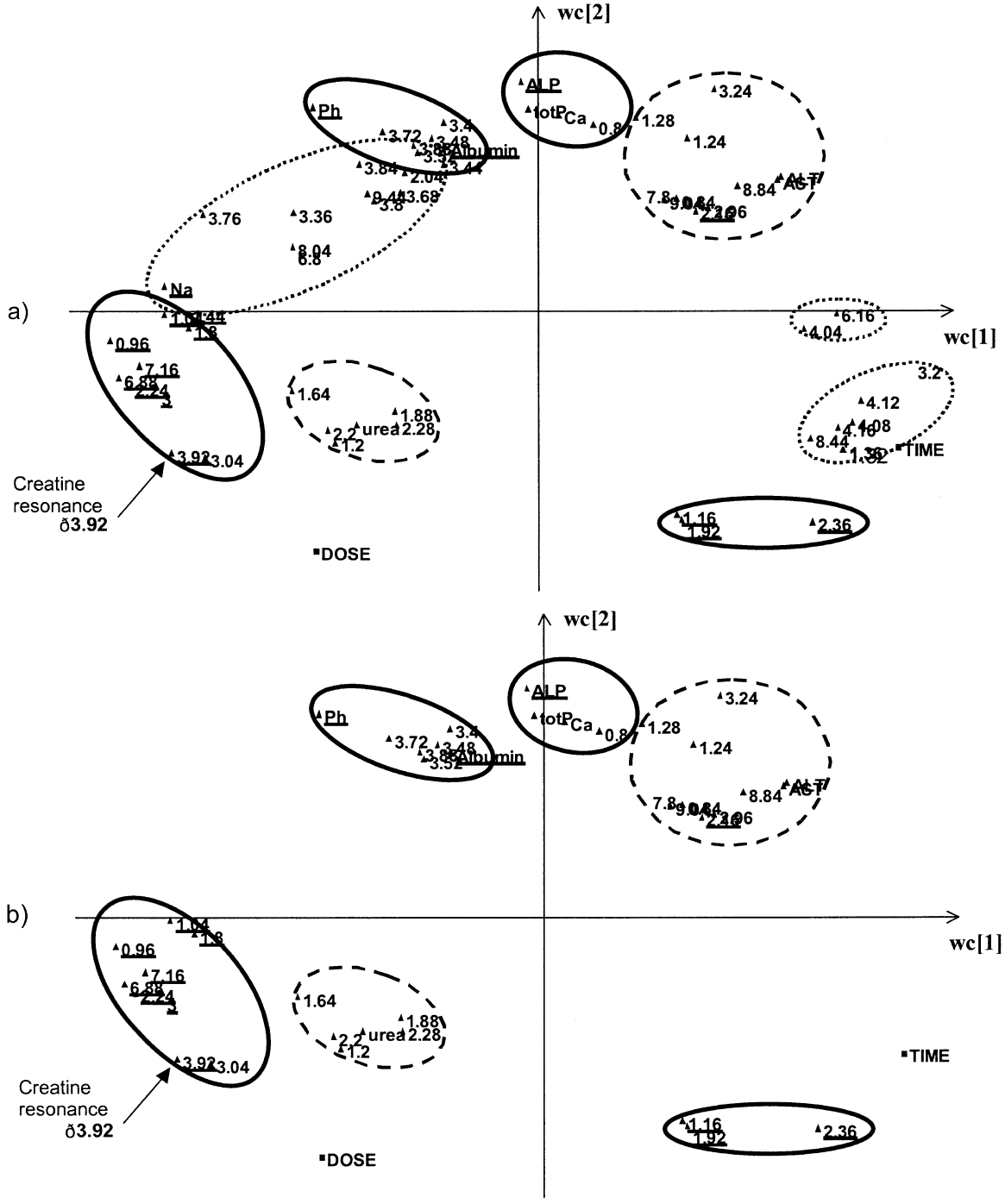

Fig. 4. (a) PLS weights plot ($\boldsymbol{w} \times \boldsymbol{c}_1/\boldsymbol{w} \times \boldsymbol{c}_2$) after the first screening step (significance screening) showing only the significant variables according to dose (dashed line), time (dotted line), dose and time (solid line). (b) PLS weights plot ($\boldsymbol{w} \times \boldsymbol{c}_1/\boldsymbol{w} \times \boldsymbol{c}_2$) after the second screening step (biomarker screening) showing only the potential markers for hydrazine toxicity.

h post dose. The original spectral data for the resonance at δ 3.92 for six representative spectra, one for each experimental setting in the design, verified the model results (Fig. 3c). In the plot, the effect of the high hydrazine dose (90 mg/kg) can be seen as a large intensity increase at 48 h post dose. This significant change according to time can also be verified by the fact that the increase in peak intensity is smaller at 168 h compared to 48 h post hydrazine administration. In addition, δ 3.92 was negatively correlated to the clinical chemistry parameters ALT and AST in the PLS weights plot (Fig. 2), implying that when δ 3.92 (creatine) increases in intensity, levels of ALT and AST decrease in the

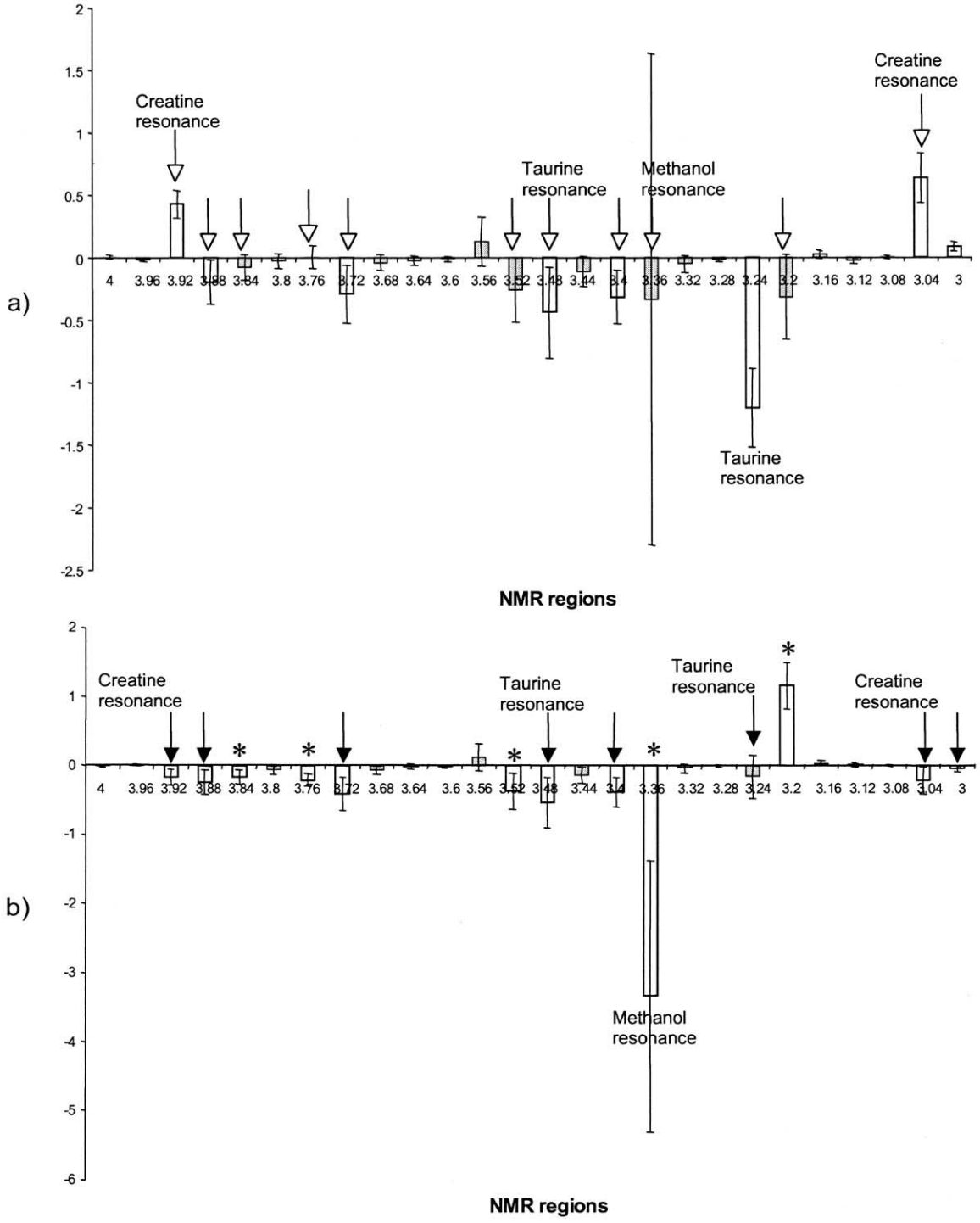

Fig. 5. PLS regression coefficients (δ3.0–4.0) for the hydrazine blood serum study. Significant variables (white bars), nonsignificant variables (grey bars). (a) Regression coefficients related to hydrazine dose. Variables significant according to time are denoted by open arrows. (b) Regression coefficients related to serum collection time. Variables significant according to dose are denoted by filled arrows and variables significant according to time only are denoted by *.

serum samples, also implying that ALT and AST levels are depleted due to hydrazine dosing.

Thus in this way, interpretations could be made for all variables (NMR regions and clinical chemistry parameters) in the model and significant changes due to dose, time and dose × time could be detected and verified statistically.

3.2. Two step variable screening procedure for hydrazine toxicity

A two step screening strategy was applied to the combined blocks of NMR and clinical chemistry variables for the hydrazine data set. Removal of the nonsignificant variables (significance screening) leaves only the variables significantly changing with dose or time or combinations thereof (Fig. 4a). In the second screening step (biomarker screening), all NMR and clinical variables significant with time only were removed leaving a relatively small number of variables significantly affected by hydrazine dose (Fig. 4b). The remaining NMR spectral regions and clinical chemistry parameters thus become the variables to focus on in terms of serum biomarkers for hydrazine toxicity, mechanistic understanding of the toxic insult and the correlation between NMR spectral data and serum clinical chemistry parameters.

3.3. Detailed interpretation of PLS regression coefficients for a selected NMR spectral area

To further emphasize the benefits of this proposed method, the PLS regression coefficients (b) for the variables in a selected spectral range were investigated and interpreted in more detail for the hydrazine blood serum study. The PLS regression coefficients (b) for the spectral range δ 3.0–4.0 were plotted together with the corresponding 95% confidence interval for each coefficient (Fig. 5a and b). The range was selected to include the two spectral regions associated with the metabolite creatine (δ 3.92 and δ 3.04), which were discussed earlier. Viewing the PLS regression coefficients according to hydrazine dose (Fig. 5a), it was evident that eight of the displayed spectral regions were significant. Among those, the two regions attributed to creatine (δ 3.92 and δ 3.04) showed a significant increase with hydrazine dose. The segment at δ 3.24 associated with the metabolite taurine also showed a high degree of significance related to an intensity decrease with dosing of hydrazine. This could also be detected in the PLS weights plot for the hydrazine serum study (Fig. 5a), where this region showed a high negative correlation with hydrazine dose. The spectral segment at δ 3.36 associated with a methanol contaminant proved to be insignificant with dose implying that the changes in that spectral region were completely unrelated to hydrazine administration. From consideration of the regression coefficients (b) for time (Fig. 5b), it was evident that 11 of the displayed spectral regions turned out to be significant with time. Notably, the region 3.36 (methanol) showed a high degree of significance related to a decrease with time, highlighting the fact that by using DoE the effects of dose and time can be efficiently separated. The filled arrows in the plot indicate the spectral regions that were considered significant according to dose and the stars indicate the spectral regions only significant according to time. From the coefficient plot for time (Fig. 5b), it was possible to distinguish between potential biomarkers for hydrazine dose, spectral changes unrelated to hydrazine dose, which could not be considered as biomarkers of hydrazine toxicity but as potential carriers of valuable information related to non-dose specific events, and nonsignificant spectral areas in all respects. Within the group considered as potential biomarkers, a division could still be made into regions or metabolites significant according to various combinations including dose (i.e. dose, dose and time, dose and dose × time or dose, time and dose × time). This subtler classification could hence provide a further more detailed understanding of the metabolic evolution due to toxic insult.

4. Discussion

By applying DoE to the variables corresponding to hydrazine dose concentration and serum collection time and regressing the design matrix (**X**) against the corresponding reduced J-resolved NMR spectra and Clinical Chemistry data (**Y**) using PLS, systematic intensity changes in NMR spectral regions and clinical chemistry parameters could be classified as significant or not according to dose, time, the interaction between the two (dose × time) or combinations thereof on a 95% significance level.

Combination of the J-resolved NMR spectra and the serum clinical chemistry parameters in the same model enabled the interpretation of inter-block variable correlations and covariations based on statistical significance providing greater information recovery as well as a more detailed biochemical explanation to changes occurring due to toxic insult as well as non-dose related events.

The fact that the significant changes caused by dose and time can be interpreted independently means that the suggested approach is able to separate changes occurring due to toxin treatment from events taking place with time, which are unrelated to dose. Examples of such changes could be diet, state of health, diurnal cycles, genetic drift, stress and strain differences as well as bacterial and other types of contamination, drifts in instrumentation and experimental conditions. This was exemplified by the methanol contamination of the serum NMR spectra in the hydrazine dosing study. The DoE–PLS approach managed to easily separate the effects of dose and time on the spectral region associated with methanol. In this way, it was seen that methanol was only changing significantly with time and not with dose and hence that spectral region could be discarded as a potential biomarker. In complex biological data sets, this kind of analysis is crucial in order not to draw false conclusions

about dose/disease related effects and thereby detect 'false' markers for toxicity or disease.

By applying DoE–PLS, the interpretation and understanding of changes in complex data matrices will be facilitated. The variables were classified as being significant or not and the significant variables could then be further characterised according to what they showed in terms of significant changes. Hence, biomarker detection and identification could be carried out in a more efficient and reliable manner, since overall nonsignificantly changing variables could be ignored together with variables significant only with time, leaving only the significant dose-related variables, associated to certain metabolites, as a result of the screening.

The approach presented here is not limited to the types of data used here. Instead there is a great potential for adding other types of data in order to find the statistical significance of these descriptors according to the variables varying in the design (in this case dose and time) and also, importantly, to find the correlation and covariation between variables within and between blocks of descriptors. A challenging task would be the application of this approach to the analysis of gene expression data in order to screen for genes or combination of genes significantly affected by toxic insult or disease as well as finding correlations and covariations between metabonomic descriptors and gene expression patterns.

5. Conclusions

The combination of design of experiments (DoE), multivariate projections (PLS) and statistical significance testing forms an efficient means of screening for biomarkers and detecting toxicity-related patterns in metabonomic NMR data. The method also proves a reliable approach for the analysis of combined blocks of multiparametric data suggesting its value in the aim of combining data sets from different analytical techniques, e.g. LC, NMR, MS, or combining data matrices generated within different fields of science, e.g. genomics, proteomics, metabonomics, in order to facilitate the understanding the processes taking place in biological systems.

By applying DoE–PLS in the variables dose and time significant changes according to toxic insult could be separated from purely time-related changes occurring from impurities, inherent physiological variation, drifts in instrumentation or experimental conditions. Separation of the effects into groups according to significance facilitated information recovery leading to better understanding of the occurring metabolic processes.

Correlation pattern interpretation in multiblock response data was facilitated by PLS projections, yielding clustering of variables according to significance and type of effect.

The two-step screening procedure vastly reduced the number of variables considered as markers for hydrazine toxicity facilitating the interpretation of the acquired complex multiparametric data structures.

The proven ability of the method to screen large sets of complex multiparametric data and assign the individual variables to a certain kind of functionality suggests that it could be of great value in studies of other complex data sets generated within various fields of science, e.g. genomics and proteomics and in screening for toxicity and disease.

Acknowledgements

The authors acknowledge the members of the Consortium for Metabonomic Toxicology (COMET), Eli Lilly, Bristol-Meyers-Squibb, Roche, Novo Nordisk, and Pfizer and Pharmacia, for financial support.

References

[1] J.K. Nicholson, J.C. Lindon, E. Holmes, "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data, Xenobiotica 11 (1999) 1181–1189.
[2] J.K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Metabonomics: a generic platform for the study of drug toxicity and gene function, Nat. Rev., Drug Discov. 1 (2002) 153–161.
[3] J.C. Lindon, J.K. Nicholson, E. Holmes, J.R. Everett, Metabonomics: metabolic processes studied by NMR spectroscopy of biofluids, Prog. Nucl. Magn. Reson. Spectrosc. 12 (2000) 289–320.
[4] K.P.R. Gartland, C. Beddell, J.C. Lindon, J.K. Nicholson, The application of pattern recognition methods to the analysis and classification of toxicological data derived from NMR spectroscopy of urine, Mol. Pharmacol. 39 (1991) 629–642.
[5] E. Holmes, J.K. Nicholson, G. Tranter, Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks, Chem. Res. Toxicol. 48 (2001) 182–191.
[6] B.M. Beckwith-Hall, J.K. Nicholson, A.W. Nicholls, P.J.D. Foxall, J.C. Lindon, S.C. Connor, M. Abdi, J. Connelly, E. Holmes, Nuclear magnetic resonance spectroscopic and principal components analysis investigations into biochemical effects of three model hepatotoxins, Chem. Res. Toxicol. 11 (1998) 260–272.
[7] J.K. Nicholson, I.D. Wilson, High resolution proton NMR spectroscopy of biological fluids, Prog. Nucl. Magn. Reson. Spectrosc. 21 (1998) 444–501.
[8] J.K. Nicholson, M.J. Buckingham, P.J. Sadler, High resolution proton NMR studies of vertebrate blood and plasma, J. Biochem. 211 (1983) 606–615.
[9] M.L. Anthony, V.S. Rose, J.K. Nicholson, J.C. Lindon, Classification of toxin-induced changes in 1H NMR spectra of urine using an artificial neural network, J. Pharm. Biomed. 13 (1995) 205–211.
[10] E. Holmes, A.W. Nichols, J.C. Lindon, S. Ramos, M. Spraul, P. Neidig, S.C. Connor, J. Connely, S.J.P. Damment, J.N. Haselden, J.K. Nicholson, Development of a model for classification of toxin-induced lesions using ^{1}H NMR spectroscopy of urine combined with pattern recognition NMR, Biomed 11 (1998) 235–244.
[11] E. Holmes, F.W. Bonner, B.C. Sweatman, J.C. Lindon, C.R. Beddell, E. Rahr, J.K. Nicholson, Nuclear magnetic resonance spectroscopy and pattern recognition analysis of the biochemical process associated with the progression of and recovery from nephrotoxic lesions in the rat induced by mercury (II) chloride and 2-bromoethanamine, Mol. Pharmacol. 42 (1992) 922–930.
[12] H. Antti, M.E. Bollard, T. Ebbels, H. Keun, J.C. Lindon, J.K. Holmes, E. Holmes, E. Batch, Statistical Processing of 1H NMR-derived urinary spectral data, J. Chemometrics. 16 (2002), 461–468.

[13] J. Azmi, J.L. Griffin, H. Antti, R.F. Shore, E. Johansson, J.K. Nicholson, E. Holmes, Metabolic trajectory characterisation of xenobiotic-induced hepatotoxic lesions using statistical batch processing of NMR data, Analyst 127 (2002) 271–276.
[14] E. Holmes, P.J.D. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr, J.C. Lindon, M. Spraul, P. Neidig, Automatic data reduction and pattern recognition methods for analysis of 1H nuclear magnetic resonance spectra of human urine from normal and pathological states, Anal. Biochem. 220 (1994) 284–296.
[15] G.E.P. Box, W.G. Hunter, J.S. Hunter, Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, Wiley, New York, 1978.
[16] L. Eriksson, E. Johansson, N. Kettaneh-Wold, C. Wikstrom, S. Wold, Design of Experiments: Principles and Applications, 1999, Umetrics AB, ISBN 91-973730-0-1.
[17] S. Wold, L. Eriksson, M. Sjostrom, Partial least squares projections to latent structures (PLS), Chemistry Encyclopedia of Computational Chemistry, Elsevier, Amsterdam, 1998.
[18] L. Eriksson, E. Johansson, N. Kettaneh-Wold, S. Wold, Multi- and Megavariate Data Analysis, Principles and Applications, 1999–2001, Umetrics AB, ISBN 91-973730-1-X.
[19] A.W. Nicholls, J. Haselden, J.C. Lindon, E. Holmes, J.K. Nicholson, Nuclear magnetic resonance spectroscopic investigations into hydrazine toxicity, Chem. Res. Toxicol. 27 (1999) 260–272.
[20] B.M. Beckwith-Hall, N.A. Thompson, J.K. Nicholson, J.C. Lindon, E. Holmes, A metabonomic investigation of hepatotoxicity using diffusion-edited 1H NMR spectroscopy of blood serum, Analyst 128 (2003) 814–818.
[21] T.W.M. Fan, Prog. Nucl. Magn. Reson. Spectrosc. 28 (1996) 161–219.
[22] J.K. Nicholson, P.D. Foxall, M. Spraul, R.D. Farrant, J.C. Lindon, 750 MHz ^{1}H and ^{1}H-^{13}C NMR spectroscopy of human blood plasma, Anal. Chem. 67 (1995) 793–811.
[23] J.C. Lindon, J.K. Nicholson, J.R. Everett, NMR spectroscopy of biofluids, Annu. Rep. NMR Spectrosc. 38 (1999) 1–87.
[24] H. Wold, Nonlinear estimation by iterative least squares procedures, in: F. David (Ed.), Research Papers in Statistics, Wiley, New York, 1966, pp. 441–444.
[25] H. Wold, Estimation of principal components and related models by iterative least squares, in: P.R. Krishnaiah (Ed.), Multivariate Analysis, Academic Press, New York, 1966, pp. 391–420.
[26] H. Martens, T. Naes, Multivariate Calibration, Wiley, Chichester, 1989.
[27] A. Hoskuldsson, Prediction Methods in Science an Technology, Basic Theory vol. 1,Thor Publishing, Warsaw, Poland, 2001.