Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Magnetic Resonance 170 (2004) 329–335

JMR Journal of Magnetic Resonance

www.elsevier.com/locate/jmr

Automatic alignment of individual peaks in large high-resolution spectral data sets

Radka Stoyanova[a,*], Andrew W. Nicholls[b], Jeremy K. Nicholson[c], John C. Lindon[c], Truman R. Brown[d]

[a] *Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA*
[b] *Metabometrix Ltd., RSM, Prince Consort Road, London, SW7 2BP, UK*
[c] *Biomedical Sciences Division, Biological Chemistry, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London, SW7 2AZ, UK*
[d] *Hatch Center for MR Research, Columbia University, 710 W. 168th St., New York, New York 10032, USA*

Received 30 December 2003; revised 19 July 2004
Available online 19 August 2004

Abstract

Pattern recognition techniques are effective tools for reducing the information contained in large spectral data sets to a much smaller number of significant features which can then be used to make interpretations about the chemical or biochemical system under study. Often the effectiveness of such approaches is impeded by experimental and instrument induced variations in the position, phase, and line width of the spectral peaks. Although characterizing the cause and magnitude of these fluctuations could be important in its own right (pH-induced NMR chemical shift changes, for example) in general they obscure the process of pattern discovery. One major area of application is the use of large databases of ^{1}H NMR spectra of biofluids such as urine for investigating perturbations in metabolic profiles caused by drugs or disease, a process now termed metabonomics. Frequency shifts of individual peaks are the dominant source of such unwanted variations in this type of data. In this paper, an automatic procedure for aligning the individual peaks in the data set is described and evaluated. The proposed method will be vital for the efficient and automatic analysis of large metabonomic data sets and should also be applicable to other types of data.

Keywords: Pattern recognition; Principal component analysis; Spectroscopy; Spectral correction; Metabonomics

1. Introduction

Given the vast amount and complexity of biochemical and spectroscopic data that can be obtained from metabonomics studies [1,2], it is necessary to invoke computer-based pattern recognition (PR) methods to ensure optimum retrieval of information. These techniques are effective tools for reducing the complexity of information contained in multiple data sets, such as from nuclear magnetic resonance (NMR) spectra, to a smaller number of features, which can be used to identify meaningful biochemical effects [3]. These spectral features can then be directly related to biological endpoints, such as drug toxicity or disease diagnosis [4,5].

Before such PR analyses can be carried out, effects of unwanted experimental variations need to be removed from the data. Principal component analysis (PCA) has proven to be an essential step for such preprocessing and general exploration of these complex data sets. In some cases PCA is sufficient to supply the sought information by itself [6]. PCA is a well-known statistical technique for the analysis of large, multivariate data sets, which extracts the basic features (patterns, factors) from the data [7]. The experimental and instrument-induced variations consist of fluctuations in the peak position,

[*] Corresponding author.
E-mail address: Radka.Stoyanova@fccc.edu (R. Stoyanova).

1090-7807/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2004.07.009

phase, and line width [8]. Although characterizing these fluctuations can be sometimes important (pH or drug binding-induced chemical shifts, for example) in general they obscure the process of pattern discovery, since often the biologically significant changes are related to the changes in the amplitude of the different signals. Thus, before applying PR techniques, the unwanted variations in frequency, line width, and phase need to be removed. Aggregating the spectral intensities within fixed frequency intervals ("binning") is a simple but effective way to overcome the effect of peak line shape fluctuations on the pattern discovery process. This technique has produced successful classification in numerous studies [4,6]. It is also clear, however, that with this method subtle changes in some peaks will be masked by larger changes in neighboring ones. PCA-based approaches for estimating phase variations [9]; phase and frequency [10] and for simultaneous correction of frequency, line width, and phase variations [11,12] have been proposed. The technique has to be applied to a single resonance peak common to all spectra in the data set, whose behavior can be assumed to be related only to experimental/instrumental artifacts. Correction factors are determined for this reference peak and then applied to the entire spectrum.

For high-resolution ^{1}H NMR data, however, the peak variability is much higher and correcting the data using a single reference peak is not sufficient. There are variable frequency shifts in the spectra, primarily due to pH and temperature variations. Recently, a procedure for further refinement of the data, using a genetic algorithm has been proposed [13]. The approach aligns the peaks in automatically selected segments in each spectrum to the corresponding peaks in a preselected reference spectrum. The main drawback to this approach is the use of the reference spectrum because, for example, in many metabonomic studies the spectra from perturbed systems can show complete loss of some peaks, or appearance of new peaks.

In this paper, a completely different approach for simultaneous alignment of the individual peaks in the entire data set is presented. The technique is based on detecting the peak-regions, where shifts occur and aligning the peaks in these spectral regions by the highest frequency point. The detection process consists of sliding first derivatives of a variety of simulated peak-shapes along the second PC of the spectral data set and calculating the correlation of these shapes with the corresponding frequency points of the underlying PC. High correlations (typically >0.80) at given points are indicative of the presence of peak shifts and at these positions, the corresponding peak-regions are selected for alignment. By adjusting the local frequency shifts in these series of spectral regions, the second and higher order PCs then become mainly related to the desired amplitude changes, providing data sets amenable to PR analysis.

2. Theory and methods

PCA has been used extensively to analyze large multidimensional data sets [7]. It identifies fundamental structures in a data matrix, called principal components (PCs), through an orthogonal decomposition of the data, using the PCs $(\vec{\mathbf{P}}_1, \vec{\mathbf{P}}_2, \ldots)$ as a basis set. The general theory behind the application of PCA to spectral data is presented in detail elsewhere [14] so in this paper, PCA is only described in relation to the explanation of the proposed peak alignment procedure.

When applied to spectral data matrix **D**, containing n spectra with m points each,the PCs are spectral shapes, providing the representation:

$$\underset{(n\times m)}{\mathbf{D}} = \underset{(n\times 1)}{S_1}\ \underset{(1\times m)}{\vec{\mathbf{P}}_1} + \underset{(n\times 1)}{S_2}\ \underset{(1\times m)}{\vec{\mathbf{P}}_2} + \underset{(n\times 1)}{S_3}\ \underset{(1\times m)}{\vec{\mathbf{P}}_3} + \cdots + \underset{(n\times 1)}{S_m}\ \underset{(1\times m)}{\vec{\mathbf{P}}_m}. \tag{1}$$

S_j are the projections of the data along the PCs, called scores. The PCs are orthonormal, i.e., each PC is orthogonal to the rest of the PCs and the length of each PC vector is 1. They are calculated as the eigenvectors of the data covariance matrix and in this particular implementation the covariance matrix is estimated around the origin, rather than the mean. The PCs are ordered by the decreasing amount of variation in the data they explain. For spectral data, the sources of coherent variations are typically small relative to the large number of original variables and a significant data-reduction can be achieved by representing the data, using just a few of the PCs in Eq. (1). The remaining PCs are noise related and can largely be ignored without loss of information.

If **D** is a spectral data set in which the only coherent variation is the amplitude and frequency of a single peak-shape **f**, then each individual spectrum in **D** (ignoring the noise) can be represented as:

$$\vec{\mathbf{s}}_i = A_i\,\vec{\mathbf{f}}(\omega_j - \omega_i) \quad i = 1,\ldots,n;\ j = 1,\ldots,m, \tag{2}$$

where for the ith spectrum A_i is the peak amplitude and $\omega_i = \omega_0 + \delta\omega_i$ is the frequency offset determined as $\delta\omega_i$ shift from some average frequency position ω_0. If it is assumed that **D** contains at least two distinct spectral shapes, i.e. at least two $\vec{\mathbf{s}}_i$ having different frequency shifts, based on the assumptions implicit in Eq. (2), it follows that the remaining peak parameters, such as line width and phase are identical for the spectra in **D**. Both **A** and $\boldsymbol{\delta\omega}$ are $(n \times 1)$ matrices, containing the amplitudes A_i and frequency variations $\delta\omega_i$. A Taylor series expansion of the signal **f** around ω_0 yields:

$$\mathbf{D} = \mathbf{A}\left(\vec{\mathbf{f}} + \left.\frac{\partial \vec{\mathbf{f}}}{\partial\omega}\right|_{\omega_0} \delta\omega + \frac{1}{2}\left.\frac{\partial^2 \vec{\mathbf{f}}}{\partial\omega^2}\right|_{\omega_0} \boldsymbol{\delta\omega}^2 + \cdots\right). \tag{3}$$

Assuming that the frequency variations are small, i.e., $\delta\omega \gg \delta\omega^2$, the second and higher order members in the Taylor series can be ignored. PCA in this case yields two PCs, that describe $\mathbf{D}$. It can be shown that for spectral data, defined as in Eq. (2) the shape of $\mathbf{P}_2$ is similar to the shape of the derivative $\left.\frac{\partial \vec{\mathbf{f}}}{\partial \omega}\right|_{\omega_0}$. Indeed, because the dominant variation in $\mathbf{D}$ is along the direction of $\vec{\mathbf{f}}$, then the first PC is also along this direction (by definition). Because $\mathbf{f}$ is a peak-shape, implying symmetry around ω_0, then $\left.\frac{\partial \vec{\mathbf{f}}}{\partial \omega}\right|_{\omega_0}$ is orthogonal to $\vec{\mathbf{f}}$. Since the signals in $\mathbf{D}$ are presented in 2D space, both in the Taylor series expansion, as well as in the PCA representation, it follows that $\mathbf{P}_2$, orthogonal by definition to $\mathbf{P}_1$, is collinear to $\left.\frac{\partial \vec{\mathbf{f}}}{\partial \omega}\right|_{\omega_0}$, i.e., $\vec{\mathbf{P}}_2$ has a derivative shape.

Therefore, if frequency shifts exist in multi-spectral data sets and if they are the dominant source of variation, then the second PC of this data set will show first derivative line shapes. Spectral regions, containing the derivatives can then be detected by their high correlations with simulated first derivative shapes, generated with different line widths. The goal is to identify spectral shapes in $\mathbf{P}_2$ that match the simulated derivatives, and this will indicate presence of frequency shifts in the corresponding spectral regions.

A spectral region of interest (SROI) is defined as a spectral sub-region of q points (the SROI may contain several peaks or the entire spectral width) and PCA is applied to this region. P_{2j} is defined as the component of $\mathbf{P}_2$ at frequency j (in points), j = $1,\ldots,q$. $f(k-\omega_0,\tau)$ which describes a simulated peak-shape with center frequency ω_0 and line width parameter τ (both in points). Without loss of generality, $f(k-\omega_0,\tau)$ can be assumed to be of a Lorentzian functional form. Only the spectral width of f containing points in a symmetrical interval of a few (ℓ) line widths around ω_0 will be considered, i.e., $k = \omega_0 - \ell\tau,\ldots,\omega_0 + \ell\tau$, where ℓ is a small positive integer. $\frac{\partial f}{\partial k}$ is defined as the derivative of f, calculated in the same interval. Typically, k is much smaller than the entire spectral width q of the SROI. The process, which follows, can be described as 'sliding' the derivative $\frac{\partial f}{\partial k}$ along $\mathbf{P}_2$ and calculating the correlation coefficients between $\frac{\partial f}{\partial k}$ and the underlying part of $\mathbf{P}_2$ at each frequency point. Formally, correlation coefficients are calculated as follows:

$$\mathbf{R}^p_\tau = \mathrm{abs}\left(\mathrm{CORREL}\left(\frac{\partial f}{\partial k}, P_{2_p}\right)\right),$$
$$p = r,\ldots,r+2\ell\tau,\ r = 1,\ldots,q-2\ell\tau. \tag{4}$$

Thus, for a given line width τ, a total of $q - 2\ell\tau$ correlation coefficients are estimated. Furthermore, τ can be varied incrementally by $\delta\tau$ over a range $\tau_{\min}$ until $\tau_{\max}$, that are considered to be the best guesses for minimal and maximum line width, present in the data. Finally, for magnetic resonance spectra, besides a single Lorentzian line, $f(k-\omega_0,\tau)$ may have the functional from of 1:1 Lorentzian doublets and 1:2:1 Lorentzian triplets or other higher multiplets in less common cases. The spectral region with the highest correlation coefficient is isolated and the peaks in this region are aligned appropriately and then, the corrected spectral region is amended in the data set.

The process continues until all spectral regions with significant correlations between the simulated first derivatives and $\mathbf{P}_2$ are corrected. The entire procedure was implemented in the IDL programming language (RSI, Boulder, CO). The program can be obtained by request from the authors. At present the input and output files are written in binary format as IEEE floats. PCA is applied to the real part of the data set signals in the frequency domain. For a typical data set such as NMR spectra as used in the exemplification here, this implies that a time-to-frequency domain processing has been performed. Typically these steps consist of data filtering, zero-filling, Fourier transformation and phasing. Also the spectra are usually co-aligned, using a resonance from an internal reference compound.

It can be assumed that amplitude and frequency shifts are the major sources of variations in high-resolution ^{1}H NMR spectral data, resulting in second and higher order PCs, containing in parts some derivative shapes. Three functional shapes of first derivatives of a single Lorentzian line, 1:1 doublets and 1:2:1 triplets were simulated since these comprise the major classes of multiplets seen in ^{1}H NMR spectra. This is not a limitation, however, since more complex multiplets can easily be simulated and even complex second order multiplets can be regarded as a superposition of singlets. Table 1 summarizes the specific parameters for simulation of these shapes, as well as the initial spectral widths considered for each one of them. The correlation coefficients are estimated using Eq. (4) and among all the correlations, the corresponding spectral regions with highest correlation $[k_1,k_2]$ are marked for further processing.

The analyzed data set is split in two: one, containing the SROI with the points between $[k_1,k_2]$ zeroed and a complementary one to which the correction procedure is applied. The spectral data in $[k_1,k_2]$ is aligned by shifting the frequencies of the points with maximum amplitude in each spectrum to a certain average frequency, which typically is the middle of the spectral region. Now, the aligned data is added to the spectral subset with zeroed intensities between $[k_1,k_2]$.

The entire procedure of calculation of $\mathbf{R}$'s is repeated, skipping the region of $[k_1,k_2]$ in $\mathbf{P}_2$. If there are correlations higher than a certain value (our experience shows that 0.80 is sufficient for indicating significant correlation), then the procedure proceeds with extraction of the corresponding spectral region and alignment. The

Table 1
Summary of parameters for simulating the derivative shapes and typical line widths limits and increments for each shape

Derivative type	ℓ[a]	ω_0	$\tau_{\min}$ (in Hz)	$\tau_{\max}$ (in Hz)	$\delta\tau$ (in Hz)
Single Lorentzian line	6	3τ	1	10	1
1:1 Lorentzian doublet	9	3τ, 6τ	0.5	10	0.5
1:2:1 Lorentzian triplet	12	3τ, 6τ, 9τ	0.1	5	0.1

[a] Number of line widths in the total number of points.

process continues until all spectral regions with significant correlations are corrected.

3. Results

The above procedure is illustrated on a data set comprising 57,600 MHz ^{1}H NMR spectra of rat urine, being a sub-set of the data acquired from a study on the hydrazine toxicity as previously published [6]. In that study, hydrazine (at three doses, 75, 90, and 120 mg/kg) was administered orally to Han Wistar rats and urine samples were collected prior to the administration and at various time periods after dosing. In addition, urine from normal control rats was collected at the same time-intervals [6].

The FIDs were line-broadened by 0.3 Hz, transformed to the frequency domain, and phased. The resonance of 3-trimethylsilyl-(2,2,3,3-$^2H_4$)-1-propionic acid sodium salt (TSP) was used as a chemical shift reference at 0 ppm. A typical spectrum from one of the control rats is shown in Fig. 1. This global alignment, based on TSP, however, is not sufficient for the complete alignment of the rest of the peaks. As an example, two spectral regions from 9 selected spectra are presented in Fig. 2: the spectral region between 2.35 and

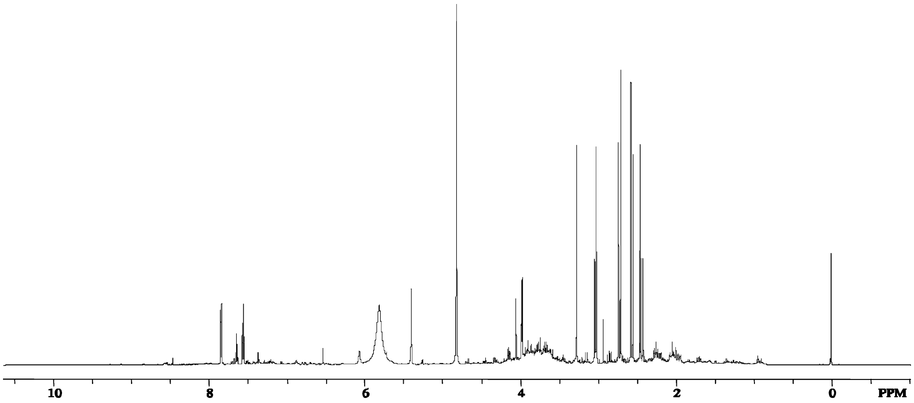

Fig. 1. Typical 600 MHz ^{1}H NMR spectrum from whole rat urine. The resonance at 0 ppm is the reference signal of TSP.

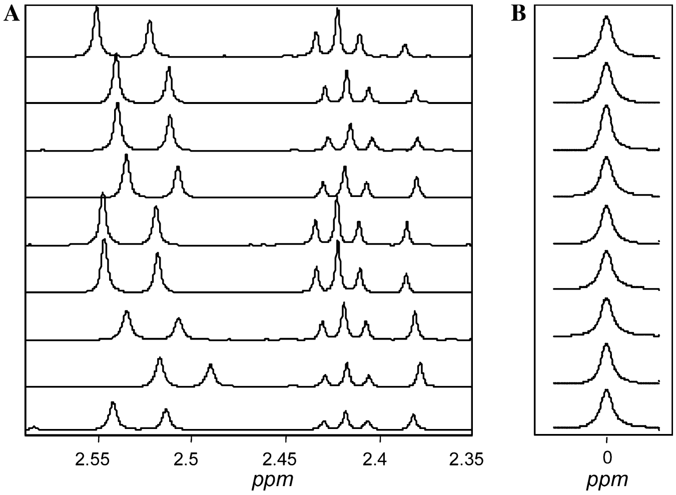

Fig. 2. (A) The spectral region of interest (SROI) in 9 spectra from the hydrazine data set. (B) The spectral region of TSP in the same spectra.

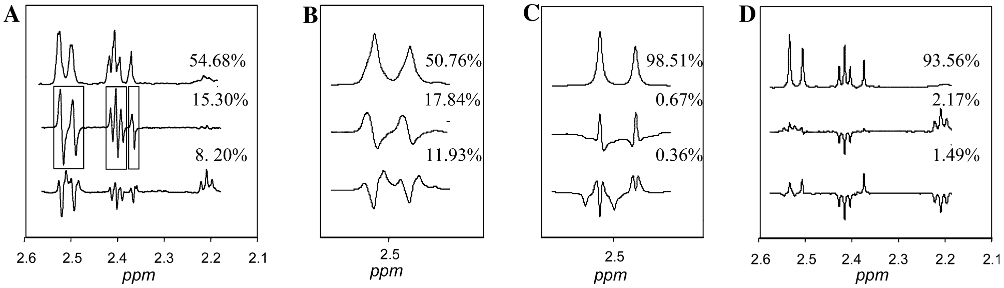

Fig. 3. The first three PCs and their corresponding normalized eigenvalues of (A) SROI, (B) before, and (C) after application of the procedure for individual peak alignment to the doublet of the AB citrate resonance. (D) SROI after application of the procedure for the individual peak alignment. The boxes around the derivative shapes in the second PC in (A) indicate the three spectral regions in which corrections were preformed, namely the doublet of the AB citrate resonance, the triplet of 2-oxoglutarate and the singlet peak of succinate.

2.60 ppm (Fig. 2A), which will be referred to later as the SROI, and the TSP peak-region [−0.03 to 0.03 ppm] (Fig. 2B). The SROI contains the following peaks: one doublet from the AB type spectrum of citrate, a triplet from 2-oxoglutarate and the singlet from succinate. Whilst the shape of the TSP peak across these spectra is almost identical (Fig. 2B), large frequency shifts in both directions can be seen for the peaks in Fig. 2A.

PCA of the SROI also reveals substantial frequency shifts in this region. The first three PCs from the SROI are presented in Fig. 3A. The second PC is composed entirely from derivative shapes and contributes ~15% to the total variance.

The procedure for the individual peak alignment was applied to this region. The correlation (Eq. (4)) between the region of the AB doublet of citrate and a doublet with line width of 6 Hz had the highest correlation coefficient (0.96) and this region was selected for alignment. PCA of the peak-region (Fig. 3B) shows that 17% of the total variance is related to frequency shifts. It should be noted that in 'real' data one of the peaks of the citrate "doublet" is consistently higher then the other and so aligning the data by the higher peak is successful. This is demonstrated in Fig. 3C, where PCA of the adjusted for frequency shifts region shows that the variance, associated with the second PC is only 0.67%.The total amount of the adjusted shift was 90 points (~10 Hz). The procedure for detecting remaining frequency shifts was applied again, skipping the region of the AB citrate in $\vec{\mathbf{P}}_2$. The correlation between the region of succinate and a single Lorentzian derivative (line width = 4 Hz) had the highest correlation (0.87). The data were aligned in this region and lastly, the region of 2-oxoglutarate was strongly correlated with a triplet derivative (line width = 2 Hz) and the shifts in this region were corrected. The total amount of adjusted shifts for these metabolites were 27 and 17 points, respectively.

The first three PCs of the SROI after the described corrections are presented in Fig. 3D. The first PC now explains 94% instead of 55% of the total variation in the data set, indicating that about 40% of the variance was related to the shifts of the three metabolites aligned. The shape of the second PC dramatically changed—instead of entirely comprised of derivative shapes it is now related to variations in peak magnitudes. It contains 'negative' peaks from citrate, 2-oxoglutarate and succinate and a positive triplet for 2-aminoadipic acid, indicating that in the data set there are processes, related to simultaneous decrease of the first three metabolites and increase of the latter [6]. As a consequence, the scores of the second PCs are now related primarily to the dose- and time-related spectral changes, rather than artifactual frequency shifts as in the case prior to application of the alignment procedure.

4. Discussion

The proposed approach substantially improves the data prior to application of PR techniques by removing experimental variations that often can mask subtle, but biochemically important, spectral changes. It is particularly useful for high-resolution data because frequency shifts are the dominant source of 'unwanted' variations in these spectra. On the one hand, these data are extremely sensitive to small variation in pH, ion concentrations, and/or temperature variations and these variations affect the various NMR peaks differently. On the other hand, variations, induced by other spectral imperfections, such as line widths, phase distortions, and baseline are minimal.

The detection of the spectral regions containing such artifactual shifts is insensitive of the precise functional line shape and width of the peaks—even in cases of a less-than-perfect match between the simulated line shapes and the regions of derivatives in the second PC, the evaluated correlations are significant. Although the peak-shapes, found in high-resolution ^{1}H NMR data are very close to Lorentzian lines, as used in this paper, the sensitivity of the detection allows flexibility in how well the synthetic shapes model the true data. It is only sufficient that the line widths of the simulated line shapes approximate the real peak line widths.

The procedure is relatively fast (processing time is less than 10 min on Digital Personal Workstation 600 a.u. with 576 MB of RAM, running Digital Unix 4.0 D), with most time being taken by estimation of the coefficients **R** (Eq. (4)). Although calculating these coefficients at each correction step is redundant it proved to be a robust way for estimation. An alternative approach would be to store all correlations, together with the corresponding spectral region and once the correlations are sorted, to execute the corrections in descending order. Because of the high-sensitivity of the detection, as discussed above, numerous line widths and accompanying spectral regions will show significant correlation with the same section of the second PC. It proved impossible after correcting the data in a given region to effectively eliminate all other correlations, related to the same data. Thus, eliminating this region for further consideration and calculating new correlations, although computationally more intensive, proved to be more accurate way to perform this part of the procedure.

Extracting the peak-regions for correction and reinserting the corrected data can introduce slight offsets when connecting the edges of the peak-region with the remaining of the data. However, these offsets are typically small compared to the peak-amplitudes and the added variation introduced into the data are negligible. In addition, it should be noted that the correction procedure entirely preserves the area of the peaks, given that the peaks are well separated and there is no loss of the peak-area due to truncation of the peak's wings.

The procedure also assumes that the peaks (including doublets and triplets) are well separated, that the shifts among the peaks in a given region is a continuum of small shifts, and that the baseline variation in the spectral region are minimal. The procedure proposed here has the advantage that it detects exclusively the peak-regions where shifts occur and where the alignment procedure will be successful while introducing minimal error. Other methods rely on comparison with a reference spectrum [13] and are thus dependent on the degree of appropriateness of this spectrum. Variations in frequency for strongly overlapping peaks do not result in derivative shapes in the second PC and thus they are naturally excluded from this analysis. Precise modeling of the existing in the data multiplets (and subsequently their derivatives) will improve the performance of the procedure, both in terms of detection and correction of the frequency shifts. The use of 1:1 doublets and 1:2:1 triplets was adequate for the type of NMR data presented in this paper; alternatively, complex spin systems may be modeled using available Gamma simulation toolkits [15]. Of course, if the single peaks in multiplets are well separated the procedure will be successful of correcting the existing frequency shifts. In summary, the developed procedure increases the capability of PR to detect subtle, but biochemically important, spectral changes, which are often masked by experimental fluctuations. This is of particular importance as the loss of discrimination caused by spectral artifacts affects in the first place the detection of changes in the levels low and intermediate molecular weight compounds, the key components of the biochemical networks. Second, the PC scores of the corrected data now become coherent function of the experimental variables, such as dose or time. Third, the identified patterns are a collection of real peaks, rather than aggregates, which allows their interpretation in biochemical terms. Lastly, it is possible to combine and reanalyze data sets acquired under a variety of conditions or even from different techniques. The development and application of these procedures facilitate the process of pattern identification and will be used to investigate metabonomic changes due to disease processes. The approach described here will be of special applicability when analyzing large data bases of metabonomic data such as that generated by the COMET consortium investigating metabonomic approaches to drug toxicity testing [16].

Acknowledgments

The work described in this report was supported by funds provided through NIH Grant P01-CA41078 to T.R. Brown. R.S. was in part supported by a Guzik Foundation Award (P.I.: C. Patriotis).

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 29 (1999) 1181–1189.

[2] J.K. Nicholson, J. Connelly, J.C. Lindon, E. Holmes, Metabonomics: a platform for studying drug toxicity and gene function, Nat. Rev. Drug Discov. 1 (2002) 153–161.

[3] G. Hagberg, From magnetic resonance spectroscopy to classification of tumors. A review of pattern recognition methods, NMR Biomed. 11 (1998) 148–156.

[4] E. Holmes, A.W. Nicholls, J.C. Lindon, S. Ramos, M. Spraul, P. Neidig, S.C. Connor, J. Connelly, S.J.P. Damment, J. Haselden, J.K. Nicholson, Development of a model for classification of toxin-induced lesions using 1H NMR spectroscopy of urine combined with pattern recognition, NMR Biomed. 11 (1998) 235–244.

[5] M.C. Preul, Z. Caramanos, D.L. Collins, J.G. Villemure, R. Leblanc, A. Olivier, R. Pokrupa, D.L. Arnold, Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy, Nat. Med. 2 (1996) 323–325.

[6] A.W. Nicholls, E. Holmes, J.C. Lindon, J.P. Shockcor, R.D. Farrant, J.N. Haselden, S.J.P. Damment, C.J. Waterfield, J.K. Nicholson, Metabonomic investigations into hydrazine toxicity in the rat, Chem. Res. Toxicol. 14 (2001) 975–987.

[7] T.W. Anderson, An Introduction to Multivariate Statistical Analysis, Wiley, New York, 1971.

[8] A.C. Kuesel, R. Stoyanova, N.R. Aiken, C.W. Li, B.S. Szwergold, C. Shaller, T.R. Brown, Quantitation of resonances in biological 31P NMR spectra via principal component analysis: potential and limitations, NMR Biomed. 9 (1996) 93–104.

[9] M.A. Elliott, G.A. Walter, A. Swift, K. Vandenborne, J.C. Schotland, J.S. Leigh, Spectral quantitation by principal component analysis using complex singular value decomposition, Magn. Reson. Med. 41 (1999) 450–455.

[10] T.R. Brown, R. Stoyanova, NMR spectral quantitation by principal-component analysis. II. Determination of frequency and phase shifts, J. Magn. Reson. B 112 (1996) 32–43.

[11] H. Wites, W.J. Melssen, H.J.A. 'tZandt, M. van der Graaf, A. Heerchap, L.M.C. Buydens, Automatic correction for phase shifts, frequency shifts, and lineshape distortions across of series of single resonance lines in large spectral data sets, J. Magn. Reson. Ser. A 144 (2000) 35–44.

[12] R. Stoyanova, T.R. Brown, NMR spectral quantitation by principal component analysis. III. A generalized procedure for determination of lineshape variations, J. Magn. Reson. 154 (2002) 163–175.

[13] J. Forshed, I. Schuppe-Koisyinen, S.P. Jacobson, Peak alignment of NMR signals by means of a genetic algorithm, Anal. Chim. Acta 487 (2003) 189–199.

[14] R. Stoyanova, T.R. Brown, NMR spectral quantitation by principal component analysis, NMR Biomed. 14 (2001) 271–277.

[15] S.A. Smith, T.O. Levante, B.H. Meier, R.R. Ernst, Computer simulations in magnetic resonance. An object oriented programming approach, J. Magn. Reson. Ser. 106 (1994) 75–105.

[16] J.C. Lindon, J.K. Nicholson, E. Holmes, H. Antti, M.E. Bollard, H. Keun, O. Beckonert, T.M. Ebbels, M.D. Reily, D. Robertson, Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project, Toxicol. Appl. Pharmacol. 187 (2003) 137–146.