Peak Alignment of Urine NMR Spectra Using Fuzzy Warping

Wen Wu,[†] Michal Daszykowski,[‡] Beata Walczak,*[,‡] Brian C. Sweatman,[§] Susan C. Connor,[§] John N. Haselden,[§] Daniel J. Crowther,[†] Rob W. Gill,[†] and Michael W. Lutz[†]

Bioinformatics Science & Technology, GlaxoSmithKline, Gunnels Wood Road, Stevenage, Hertfordshire SG1 2NY, U.K., Department of Chemometrics, Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland, and Safety Assessment, GlaxoSmithKline, Park Road, Ware, Hertfordshire SG12 0DP, U.K.

Received August 8, 2005

Proton nuclear magnetic resonance (^{1}H NMR) spectroscopic analysis of mixtures has been used extensively for a variety of applications ranging from the analysis of plant extracts, wine, and food to the evaluation of toxicity in animals. For example, NMR analysis of urine samples has been used extensively for biomarker discovery and, more simply, for the construction of classification models of toxicity, disease, and biochemical phenotype. However, NMR spectra of complex mixtures typically show unwanted local peak shifts caused by matrix and instrument variability, which must be compensated for prior to statistical analysis and interpretation of the data. One approach is to align the spectral peaks across the data set. An efficient and fast warping algorithm is required as the signals typically contain ca. 32 000−64 000 data points and there can be several thousand spectra in a data set. As demonstrated in our study, the iterative fuzzy warping algorithm fulfills these requirements and can be used on-line for an alignment of the NMR spectra. Correlation coefficients between the aligned and target spectra are used as the evaluation function for the algorithm, and its performance is compared with those of other published warping methods.

INTRODUCTION

Proton nuclear magnetic resonance (^{1}H NMR) spectroscopy has been used increasingly in recent years to obtain and explore signal profiles of a wide range of complex mixtures. Applications of NMR mixture analysis have included the profiling of plant extracts;[1] verification of the authenticity of various types of fruit juices, wines, and tobacco;[2−4] and the metabolic profiling of biological fluids from several vertebrate and invertebrate systems.[5,6] NMR has also been applied to batch analysis and impurity profiling of discreet chemical entities. For all of these applications, analysis of the resulting ^{1}H NMR profile is considerably helped by the use of multivariate statistical analysis (MVDA) techniques such as principal components analysis (PCA) and partial least squares (PLS).

Several features of NMR spectroscopy make it useful for mixture analysis, including the potential for revealing unexpected changes which otherwise would not be detected without extensive knowledge of the complex matrix. This advantage is facilitated by the ability of NMR to detect structural information for all hydrogen-containing compounds present above the NMR detection limit. The chemical shift position of each NMR signal is particularly useful as it is completely dependent on the electronic environment of each proton and, hence, directly reflects the chemical and conformational structure of the component it depicts. These large diagnostic influences on chemical shift are consistent for individual proton-containing moieties across different NMR instruments and diverse aqueous systems and, hence, aid in the component identification.

However, there are several other factors which also have a significant, but much smaller, effect on chemical shift but which may vary from sample to sample. These include temperature shifts arising from instrumental imperfections and concentration shifts arising from small differences in the amount of either the molecule itself or other components in the surrounding matrix. Differences in sample ionic strength can also have a large impact on peak position, especially in the case of charged molecules where the effective pK_a's of the acidic or basic groups alter with ion concentration. This is particularly relevant for the metabolic profiling of urine samples, where there may be considerable intersubject variation in both sample dilution and content.

MVDA techniques rely on a consistent representation of the data for each sample. Small non-group-dependent variations in chemical shift can therefore complicate the data analysis considerably, and it is usually desirable to minimize their impact. One approach to reducing peak shift problems is to bucket the data prior to analysis. This has been used extensively with some success in the metabolic profiling of urine by NMR but has several disadvantages. These include the loss of spectral resolution and complications with data interpretation as it can be difficult to conclude the exact change within a particular bucket to which the statistics refer. Another possible approach to minimizing the impact of chemical shift is to align the peaks prior to statistical analysis. Several peak alignment algorithms have been developed, some of which have been applied previously to NMR spectra.[7]

The main goal of our study was to develop a fast and effective peak alignment algorithm for NMR spectra of rat urine and to assess the usefulness of the alignment algorithm by comparing classification models of hepatotoxicity for aligned and bucketed data. To achieve this, we used urine

* Corresponding author fax: +48-32-25-99-978; e-mail: beata@us.edu.pl.
[†] Bioinformatics Science & Technology, GlaxoSmithKline.
[‡] The University of Silesia.
[§] Safety Assessment, GlaxoSmithKline.

10.1021/ci050316w CCC: $33.50 © 2006 American Chemical Society
Published on Web 01/24/2006

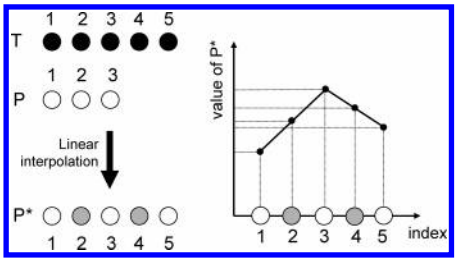

Figure 1. Principles of linear interpolation: the black circles are points of a target signal (T), and the white circles are the points of signal to be aligned (P). To stretch the section of signal P to the same length as that of the corresponding section in signal T, the linear interpolation is performed between points 1-2 and 2-3 of signal P and two new points (gray circles) in positions 2 and 4 in the interpolated signal P* are inserted. These two new points are average values of points 1-2 and 2-3 of signal P, respectively.

samples collected from Sprague−Dawley rats dosed with a model hepatotoxin, cycloheximide, or vehicle as control. The samples from this study were used purely to serve as a test data set on which to develop the peak alignment algorithm. The toxicological endpoints were therefore used to verify the model but not for detailed interpretation.

Effective and fast warping of signals is a nontrivial problem, especially for such signals, which contain thousands of data points. In our study, the fuzzy warping approach was applied. The main idea of this approach was described in ref 8 and can be summarized as an iterative procedure, alternating between fuzzy matching and signal transform, with the parameters weighted according to the correspondence of the extracted features. The correlation coefficients between the aligned test spectrum and the target spectrum were used as the evaluation function of successful alignment. A spectral comparison was also carried out to assess the performance of the alignment procedure in terms of any erroneous alignment, peak elimination, or change of peak shape. Aligned spectra were then used to develop a classification model using uninformative variable elimination−partial least squares (UVE−PLS) regression, a biased linear method endowed with feature selection. Models were compared with those obtained with bucketed data using an identical procedure.

THEORY

Fuzzy Warping (FW). For notation purposes, T and P denote the target spectrum and the spectrum to be aligned, respectively. The feature vector, **x**, is associated with each spectrum, and the elements of **x** represent the maxima of the identified peaks.

The goal of fuzzy warping is to establish correspondence between the most intense peaks of the spectra to be aligned. Once the peaks' correspondence is known, spectral fragments of the signal and the target signal are aligned using linear interpolation (see Figure 1).

The peak correspondence is estimated in an iterative manner, using a global transform with the parameters weighted according to the similarity of the peaks. Namely, features **x** are centered and scaled:

$$\mathbf{x}^c = \frac{\mathbf{x} - \bar{x}}{s} \tag{1}$$

using the following mean ($\bar{x}$) and scale (s^2) estimations:

$$\bar{x} = \frac{\sum_{i=1}^{N} m_i x_i}{\sum_{i=1}^{N} m_i} \tag{2}$$

$$s^2 = \sum_{i=1}^{N} m_i (x_i - \bar{x})^2 \tag{3}$$

where N denotes the number of elements in **x**, and **x** is the feature vector, which contains maxima of the spectrum T or P.

Weights, m, are calculated on the basis of matching the fuzzy membership Gaussian functions, N_1, centered at the positions of the peaks identified in the target spectrum, with the crisp membership functions, N_2, at the positions of peaks in a spectrum to be warped (see Figure 2).

Outputs of the ith Gaussian function for all the features of P form the ith row in matrix **G** ($N_1 \times N_2$). To calculate the one-to-one peak correspondence, matrix **G** has to be augmented with additional rows and columns to the dimensionality ($M \times M$), where $M = \max(N_1, N_2) + 1$. The last row and the last column of the new matrix, denoted as **W**, are added to deal with the peaks that have no correspondence. Initially, their elements are set to $1/M$. An iterative normalization of the columns and rows (known as the Sinkhorn normalization[9]) of **W** leads to a matrix, for which the sum of the elements in each row and column equals 1 (with an assumed degree of accuracy). After the Sinkhorn normalization, matrix **W** is reduced to matrix **W***, containing the first N_1 rows of **W** and the N_2 columns of **W** only. Then, the weight of the ith feature of T is defined as a sum of elements of the ith row of matrix **W***, whereas the weight of the jth feature of P is defined as the sum of elements of the jth column of **W***.

After transform with the actual transform parameters ($\bar{x}$ and s), the whole procedure is repeated, until convergence is obtained. However, at each iteration, the degree of fuzziness (i.e., the width of the Gaussian functions) is reduced, as

$$\sigma_{\text{iter}+1} = 0.6\sigma_{\text{iter}} \tag{4}$$

Finally, for each peak of T, the corresponding peak of P is found as having the highest value of the similarity measure; that is, for the ith peak of T, the maximal element of the ith row of matrix **W*** ($N_1 \times N_2$) is determined and its index is the index of the corresponding peak of P. To ensure proper assignment of the corresponding peaks, only those peaks which show correspondence higher than 0.99 are further taken into account. Once the peak correspondence is established, signal P is piecewise interpolated to the corresponding parts of the target T. In our study, warping was applied to the spectra in the range −0.2 to 10 ppm. After warping, all spectra were normalized to the same area, followed by exclusion of the water, urea, and chemical shift reference regions (δ_H = 4.9−4.7, 5.5−6.00, and −0.2 to 0.2, respectively).

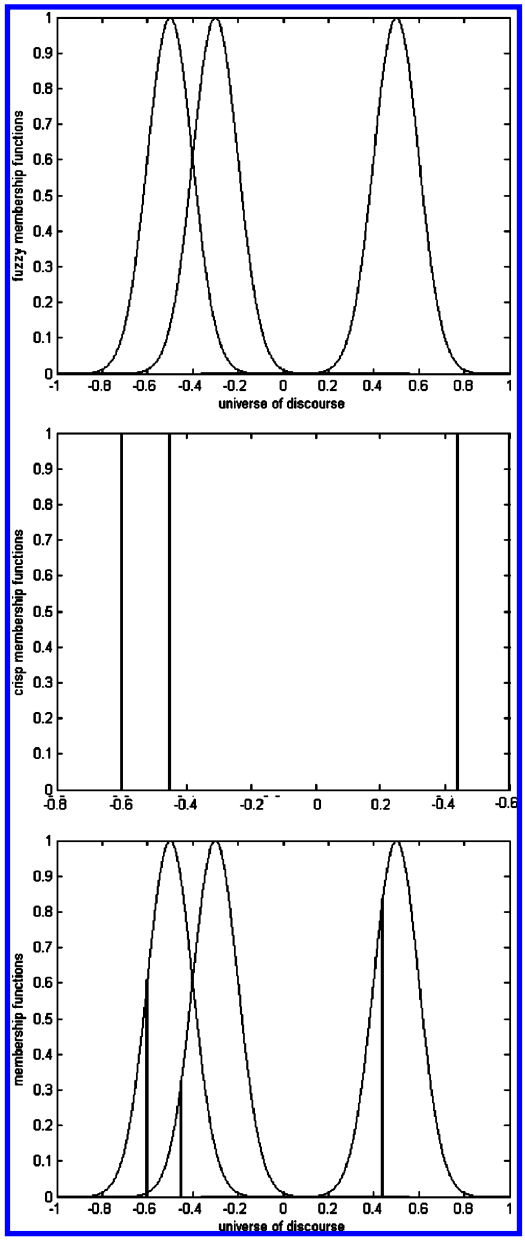

Figure 2. Graphical illustration of the main idea of the fuzzy matching approach: (a) fuzzy membership functions (Gaussian functions), centered at the positions of the identified peaks of the target signal; (b) crisp membership functions at the positions of the peaks identified in the warped signal; and (c) fuzzy matching of these two sets of membership functions.

The main steps of the FW algorithm are summarized as follows:

1. If the aligned signals have different lengths, interpolate signal P to the length of target T.

2. Find coordinates of the most intense peaks in P and T to set up feature vectors.

3. Insert Gaussian functions at the positions of the peaks in T.

4. Calculate the outputs of the Gaussian functions for the peaks of P to form matrix **G**.

5. Use matrix **G** to construct matrix **W** and, hence, account for peaks in P with no match in T.

6. Use the Sinkhorn iterative procedure to calculate the weights of the transform parameters.

7. Perform the transform of the feature vectors.

8. Check the convergence, and if it has not been achieved, decrease the degree of fuzziness and return to step 4.

9. Find the peaks with highest correspondence.

10. Perform piecewise interpolation of signal P to the corresponding parts of target T.

Depending on the problem at hand, different interpolation methods (such as linear, cubic, splines, etc.) can be used. Their choice can be based upon an improvement of the correlation coefficient, calculated for the aligned spectra.

Other Warping Algorithms. The FW algorithm described above was compared with other warping algorithms, including correlation optimized warping (COW),[10−11] dynamic time warping (DTW),[11−12] parametric time warping (PTW),[13] peak alignment by a genetic algorithm (PAGA),[7] and its two simplified versions, denoted as PALSI (peak alignment with linear shift and interpolation) and PALS (peak alignment with linear shift). In PAGA, the target and the unaligned signals are first divided into a number of segments. For each segment, the unaligned signal is then matched to its corresponding target by linear shifting and interpolation. Two integer numbers representing shifting, *s*, and interpolation, *i*, are optimized by a genetic algorithm (GA). GA is a powerful method applied to the optimization of multivariate functions; however, it needs a lot of calculation in order to obtain the optimal results. In PALSI, the two-step direct searching is used instead of GA. In the first step, *s* is optimized by scanning of an entire predefined range. In the second step (with *s* fixed), *i* is optimized by scanning of the predefined range. In PALS, only the one-step searching is used to optimize *s* and the interpolation step is deleted.

The collection of available warping algorithms is very large. Except compared to the approaches in this manuscript, there exist other ones such as the method of Brown and Stoyanova,[14] modified later by Witjes et al.[15]

Exploratory Analysis. Exploratory analysis and the modeling of the data were performed for the normalized spectra, followed by exclusion of the water, urea, and chemical shift reference regions.

Principal Component Analysis (PCA). PCA is a most popular data compression and visualization method.[16,17] It allows the representation of objects in the space of principal components (PCs), constructed as a linear combination of the original variables, and a maximizing description of data variance. The effectiveness of data compression is determined by the data rank (i.e., variables correlation). The projection of objects on the plane defined by the two main PCs reveals the main characteristics of the data structure. In our study, PCA was applied to the normalized spectra and the centered data set.

Hierarchical Clustering. An exploratory analysis of a studied data set often starts with hierarchical clustering of the data. Hierarchical clustering can be applied to reveal similarities (or dissimilarities) of objects in the multidimensional variables space. A detailed description of the hierarchical clustering methods can be found in refs 18−20.

Any hierarchical clustering method is characterized by the similarity measure and by the way the resulting subclusters are merged (linked). The final results of hierarchical clustering are presented in the form of a dendrogram. On the *x* axis of the dendrogram, the indices (or classes) of clustered objects are displayed, whereas the *y* axis represents the

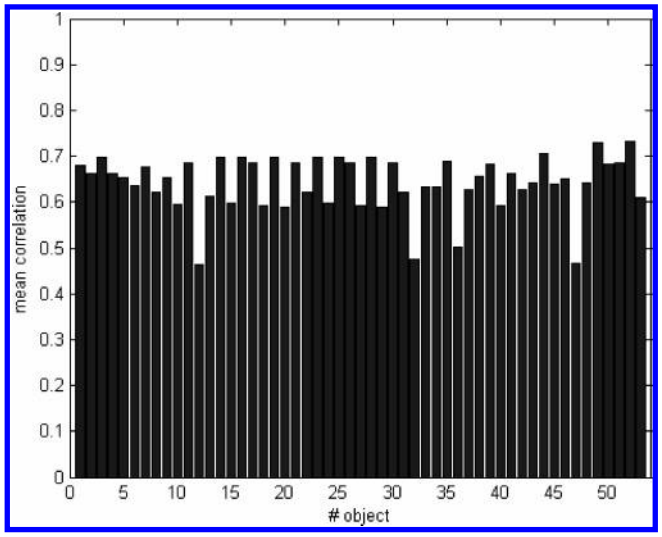

Figure 3. Mean values of correlation coefficients of individual spectra correlated with all spectra.

corresponding linkage distance or an adequate measure of similarity, between the two objects or clusters, which are merged.

In our study, the Euclidean distance was used as the similarity measure, and the Ward method was applied for subcluster linkage. The Ward linkage is based on the inner squared distance of clusters, so that at each stage these two subclusters are merged, for which the minimum increase in the total within-group error sums of squares is observed. Clustering was performed for the normalized spectra.

Classification. The warped signals are organized in the form of a data matrix **X** ($m \times n$), where m denotes the number of objects (samples) and n denotes the length of each spectrum (i.e., the number of variables). Binary vector **y** ($m \times 1$) describes sample classification.

Partial Least Squares (PLS). PLS is a bilinear regression method, well-suited for the modeling of multidimensional collinear data.[16,17,22−24] The PLS model can be presented in the following form:

$$\mathbf{X} = \mathbf{TP}^{\mathrm{T}} \tag{5}$$

$$\mathbf{y} = \mathbf{Tq} \tag{6}$$

where matrix **T** represents the PLS features of dimensionality $m \times f$, constructed to maximize the covariance between **X** and **y**; matrix **P** ($n \times f$) and vector **q** ($f \times 1$) represent x loadings and y loadings, respectively, and f denotes the number of factors.

There are many PLS algorithms[25−29] designed to deal with the data of different dimensionalities and using different types of kernels to speed up calculations. In our study, the SIMPLS algorithm by de Jong[29] was applied.

A PLS model can be constructed for either a continuous or a discrete y variable. For classification purposes, the variable y is a binary variable; all objects belonging to class 1 are assigned to the value of 1, whereas the objects from class 2 are assigned to the value of 0.

The complexity of the PLS model (i.e., the number of the PLS features, f) can be calculated, on the basis of the leave-k-out cross validation (CV) procedure (e.g., refs 16, 17, 22−24). In the case of the data set with a small number of objects, the leave-one-out CV is usually used. The main idea of this procedure is to leave out the consecutive objects, construct the PLS model for the remaining subset, and predict the y value for the object left out using 1, 2, ..., k factors. Once the y values are predicted for all the m objects, the root-mean-square error of cross validation

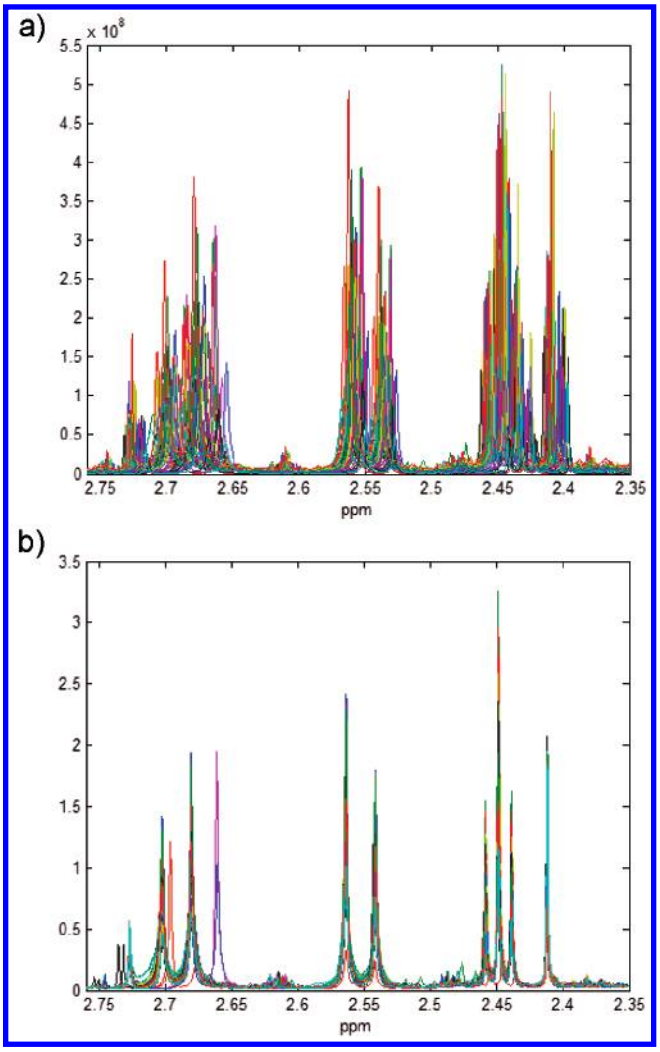

Figure 4. (a) Original ^{1}H NMR spectra in the selected region. (b) Normalized spectra warped to spectrum number 52 in the same region as in subplot a.

Table 1. Results of Different Warping Methods[a] Applied to 52 NMR Spectra, Using Object Number 52 as a Target

		mean value of correlation coefficient		
method	time (s)	original spectra	warped spectra	mean improvement (%)
COW	out of memory	0.55	NA	NA
DTW	out of memory	0.55	NA	NA
PTW	21	0.55	0.70	46
PLF	13	0.55	0.73	52
PAGA	3048	0.55	0.87	89
PALSI	474	0.55	0.87	89
PALS	167	0.55	0.87	88
FW	125	0.55	0.87	88

[a] COW, correlation optimized warping; DTW, dynamic time warping; PTW, parametric time warping; PLF, partial linear fit; PAGA, peak alignment by genetic algorithm; PALSI, peak alignment with linear shift and interpolation; and PALS, peak alignment with linear shift.

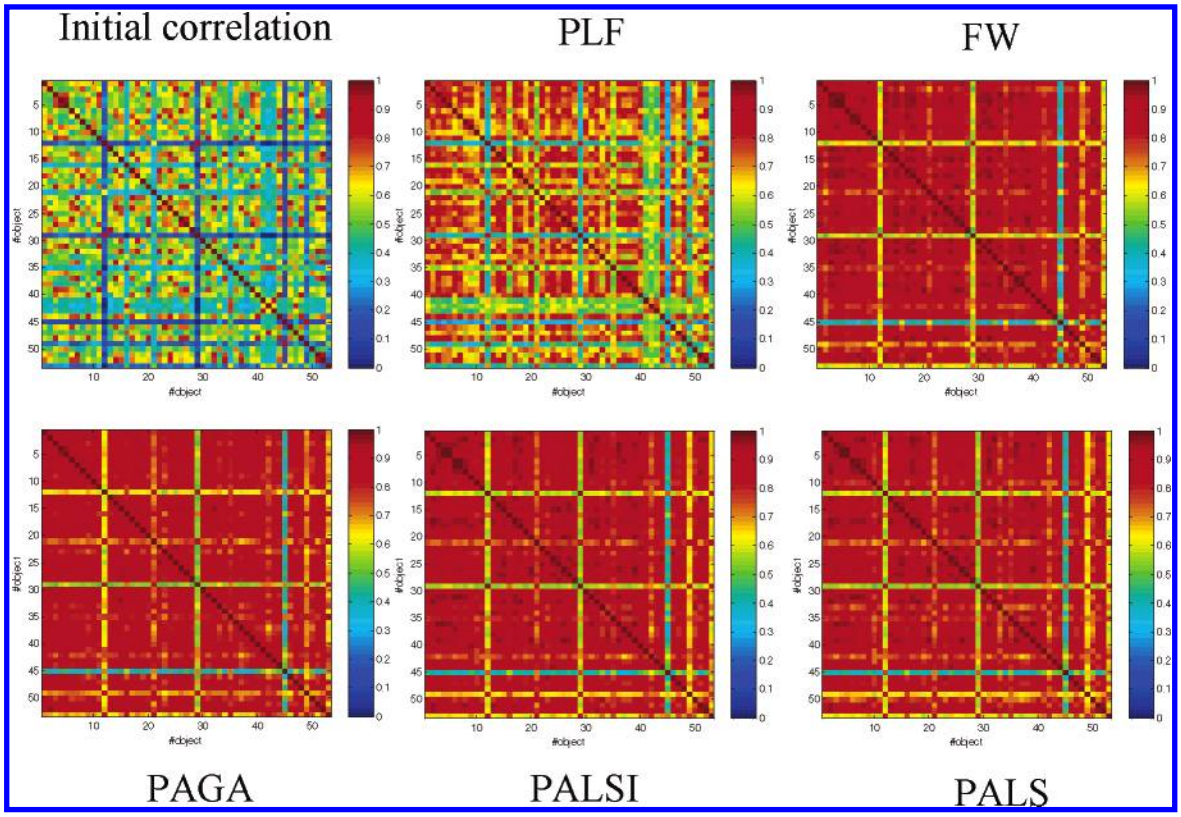

Figure 5. Images of the correlation matrices obtained from the different warping algorithms: partial linear fit (PLF), fuzzy warping (FW), peak alignment by genetic algorithm (PAGA), peak alignment with linear shift and interpolation (PALSI), and peak alignment with linear shift (PALS).

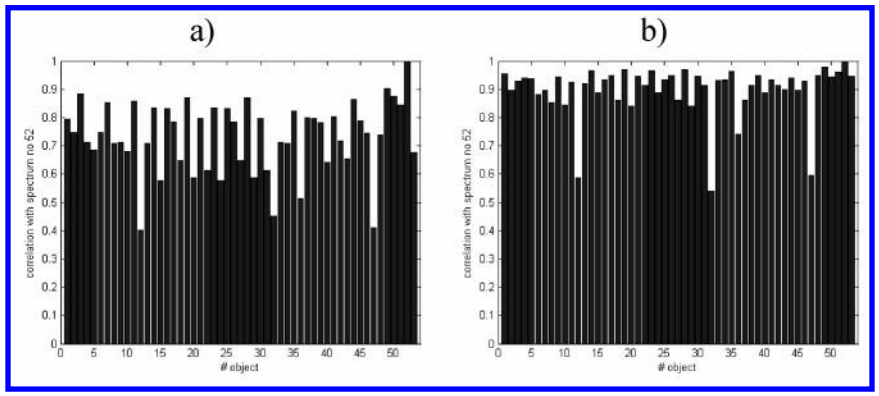

Figure 6. Correlation coefficients of spectrum number 52 with individual spectra: (a) before and (b) after warping to spectrum number 52.

(RMSCV) can be calculated as

$$\mathrm{RMSCV}(f) = \sqrt{\frac{[\hat{\mathbf{y}}(f) - \mathbf{y}]^2}{m}} \qquad \text{for } f = 1, 2, ..., k \qquad (7)$$

where $\hat{\mathbf{y}}(f)$ denotes the dependent variable predicted with f factors.

The model with the minimal value of RMSCV is selected as the model with an optimal predictive ability. The final PLS model can also be presented as

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e} \qquad (8)$$

where **b** denotes the vector ($m \times 1$) of the estimated regression coefficients and **e** ($m \times 1$) represents the vector of the y residuals.

Uninformative Variables Elimination−Partial Least Squares (UVE−PLS). The majority of calibration and classification models can be improved by eliminating the uninformative x variables, that is, the variables which are irrelevant for the y modeling. In the UVE−PLS approach (proposed by Centner et al.[30]), the identification of irrelevant variables is performed using the stability of the regression coefficients associated with these variables. Subsets of stable and unstable regression coefficients are determined, on the basis of the stability of the regression coefficients associated with the noise variables added to the original data matrix.

More precisely, the initial **X** ($m \times n$) matrix is augmented by the **R** ($m \times k$) matrix, which contains random normally distributed variables of very small magnitudes (their order of magnitude equals 10^{-10}). Let us denote the new matrix as **A** ($m \times n^*$); where $n^* = n + k$. On the basis of the leave-one-out cross-validation procedure, m vectors of the regression coefficients, **b**, of the m constructed PLS models, are calculated and collected in matrix **B** ($m \times n^*$). The stabilities

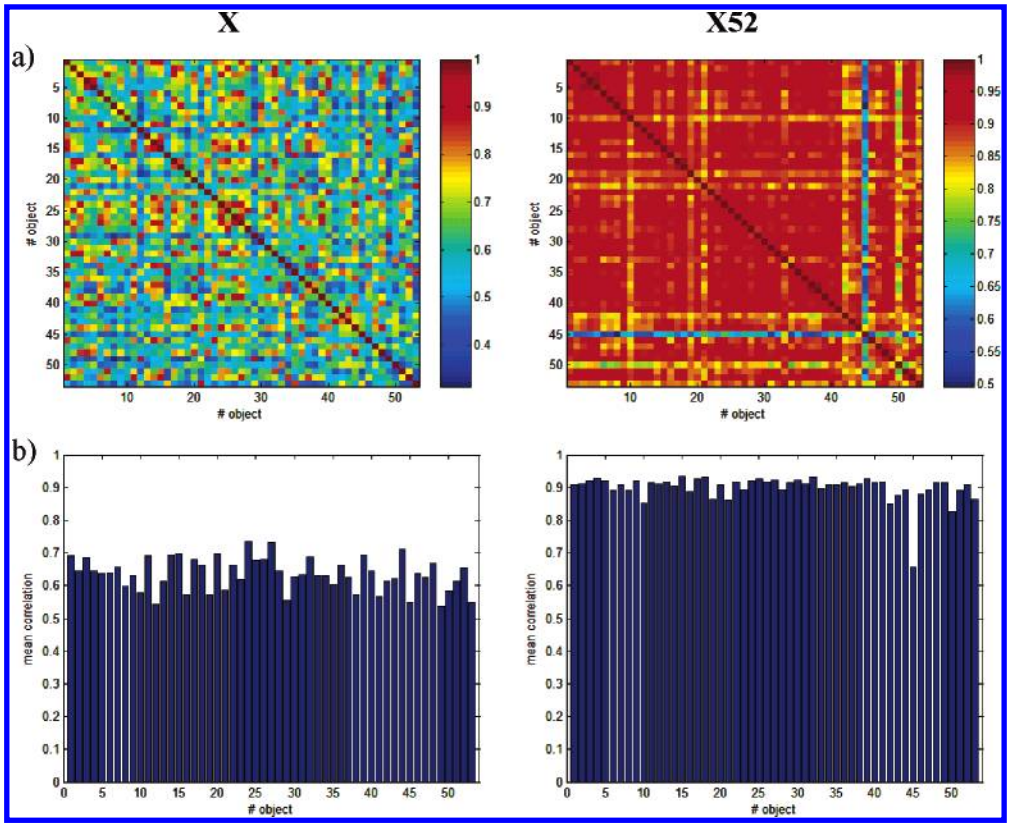

Figure 7. (a) Images of correlation matrices for the **X** and **X52** data sets and (b) mean values of the correlation coefficients of individual spectra before and after alignment.

of the n^* regression coefficients are defined as

$$\mathbf{s} = \frac{\bar{\mathbf{b}}}{\sigma} \tag{9}$$

where $\bar{\mathbf{b}}$ denotes the vector ($n^* \times 1$), the elements thereof being the mean values of the n^* columns of matrix **B**, and σ denotes the vector ($n^* \times 1$), the elements thereof being the standard deviations of the n^* column of matrix **B**.

The resulting vector **s** contains n^* elements, the first n elements corresponding to the experimental variables, whereas the remaining k elements correspond to the artificial random variables.

The cutoff value for the stability of the regression coefficients is defined as

$$\text{th} = \alpha \max |\mathbf{s}_i| \tag{10}$$

where $i = n + 1, ..., n + k$ and α denotes an additional parameter, which allows the cutoff modification. In our study, $\alpha = 0.99$.

All experimental variables associated with the regression coefficients, having their stabilities higher than the stability of the regression coefficients associated with random variables (cutoff value is defined by eq 10), are considered as informative for the y modeling, whereas the remaining ones are considered as uninformative and, therefore, are removed from the **X** data. The final PLS model is constructed for the reduced **X** matrix, which contains the relevant variables only.

EXPERIMENTAL SECTION

Materials. Reagents for the ^{1}H NMR analyses were purchased from Sigma-Aldrich Co. Ltd., Gillingham, Dorset, U.K., including monobasic sodium phosphate monohydrate ($NaH_2PO_4 \cdot H_2O$), dibasic sodium phosphate heptahydrate ($Na_2HPO_4 \cdot 7H_2O$), sodium 3-trimethyl-silyl-[2,2,3,3-$^2H_4$]-1-propionate (TSP), and sodium azide. Deuterated water (D_2O) was purchased from Fluorochem Ltd., Old Glossop, Derbyshire, U.K.

Urine Samples. Four groups of five male Sprague−Dawley rats (225−275 g body weight) were assigned to the treatment regimens that included once-per-day dosing with the vehicle (1% w/v methyl cellulose) or with 0.10, 0.25, or 0.50 mg/kg cycloheximide in 1% methyl cellulose. The animals were dosed orally by gavage each day for 3 days, and the urine samples were collected from each animal for a predose period on the day prior to dosing and then on days 1 and 3. The animals were maintained in a facility with a 12 h light/12 h dark cycle, and their urine was collected for a continuous period from 8 to 24 h following the dosing, mostly during the dark cycle. The animals were singly housed in metabolism cages during urine collection, and the urine was collected in plastic receptacles containing a 1 mL aliquot of 1% w/v sodium azide solution (aqueous) to act as a preservative. All urines were frozen immediately after collection and stored at −70 °C prior to analysis.

^{1}H NMR Analysis of Urine. Aliquots of urine (400 μL) were mixed with a phosphate buffer (200 μL, 0.2 M, pH 7.4) and frozen at −20 °C for at least 1 h. After thawing, the samples were centrifuged at 5000 rpm for 15 min in a refrigerated centrifuge (4 °C), to remove any resulting precipitate. A 500 μL aliquot of the supernatant was added to a 5 mm NMR tube (Wilmad 507PP) containing 50 μL of the TSP/D_2O/sodium azide solution (0.1% w/v TSP and 1% w/v sodium azide in D_2O). After capping, the NMR tubes

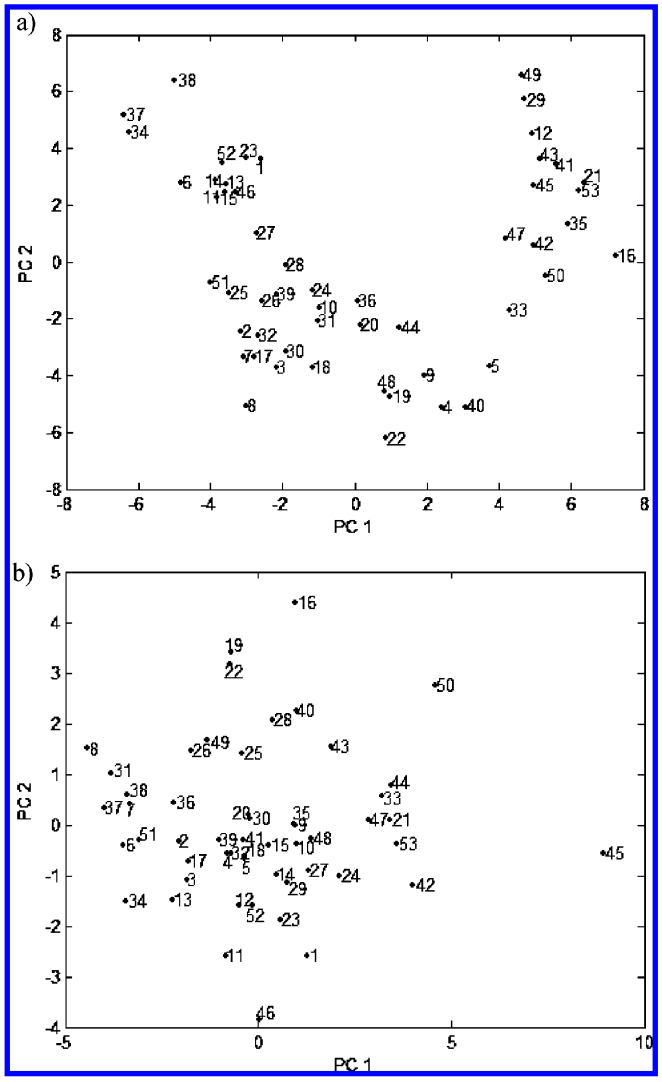

Figure 8. Projections of (a) original and (b) warped spectra on the planes defined by the first two PCs.

Table 2. Classification Results for the Partial Least Squares (PLS) and Uninformative Variable Elimination−Partial Least Squares (UVE−PLS) Models, Applied to the Original (**X**), Warped (**X52**), Bucketed (**Xb**), and Warped and Bucketed (**X52b**) Signals

data	model	complexity	RMSCV	CCR (%)	variables
X	PLS	5	0.4753	75.5	all
	UVE−PLS	11	0.1786	100	4286
X52	PLS	8	0.2667	86.8	all
	UVE−PLS	7	0.1566	100	3768
Xb	PLS	10	0.2667	92.2	all
	UVE−PLS	10	0.1753	100	34
X52b	PLS	10	0.2052	100	all
	UVE−PLS	9	0.1716	100	42

were inverted several times to ensure complete mixing of their contents.

TSP and D_2O provided a chemical shift reference (δ_H = 0.00 ppm) and the deuterium frequency lock signal, respectively, for the NMR spectrometer. Samples were analyzed at 700.01 MHz on a Bruker DRX-700 spectrometer at a probe temperature of 300 K, using a 5 mm TXI ATMA probe. Spectra were acquired and processed, using the XWIN-NMR version 3.5 software (Bruker Biospin GmbH, Rheinstetten, Germany). Acquisition of the ^{1}H NMR spectra was achieved, using a standard water presaturation pulse sequence, which incorporated the first increment of the nuclear Overhauser effect spectrometry (NOESY), RD−90°−t_1−90°−t_m−90°−collect FID. RD and t_m were relaxation delays of 2 s and 100 ms, respectively, during which time the water was selectively irradiated. The constant t_1 (5 μs) represented the delay for the first increment in a NOESY experiment. After eight dummy scans, a total of 64 scans were collected into 64 000 computer data points, with a spectral width of 14 005.63 Hz, an acquisition time of 2.34 s, and a total recycle delay of 4.34 s. An exponential line broadening function of 0.30 Hz was used to multiply the free-induction decay (FID) signals prior to Fourier transformation into 64 000 data points. After a manual adjustment of the phase and an automatic baseline correction, the real spectra were converted to JCAMP-DX fixed, uncompressed format prior to further data processing.

RESULTS AND DISCUSSION

Cycloheximide was used in this study at doses expected to produce liver pathology without compromising the survival of the animals over a 4-day period. Conventional toxicological endpoints confirmed a dose-dependent response to cycloheximide. There was a clear treatment-related decrease in serum total protein (albumin and globulin) at the high dose as well as a decrease in haematocrit and reticulocyte count. Overall, the general picture is of a dose-dependent inhibition of hepatic protein synthesis. Generalized protein synthesis inhibition has also had an effect on red blood cell formation, although over the time course of this study, this has not manifested itself in a decreased red blood cell count. Increased liver weights and concomitant pathology [cytoplasmic basophilic alteration (5/5, 4/5, and 0/5) and periportal vacuolation (3/5, 2/5, and 0/5) for the high, medium, and low doses, respectively] were consistent with the liver trying to increase protein synthesis as a rebound effect of cycloheximide.

The main goal of our study was to construct a classification model, which allows distinguishing of the control samples from the low- (class 1), medium-, and high-dosed samples (class 2), on the basis of the registered NMR spectra. Prior to statistical analysis, an initial visual inspection of the NMR data revealed that several control samples and dosed samples contained regions of the spectra that had an unusually high degree of shift variation. The cause of this phenomenon was uncertain but may have arisen from temperature instability during NMR acquisition. To determine any compound-related effects, it was necessary to reduce the impact of these shifts. This data set was, therefore, chosen as an ideal candidate for the evaluation of a peak alignment algorithm.

The warping procedure requires selection of a target spectrum for an alignment of all the remaining signals. In our study, a few different targets were tested. Their choice was based on the mean value of the correlation coefficient of the individual spectra with all the remaining NMR signals (see Figure 3).

To better visualize the differences in the peak shifts, Figure 4 shows the NMR spectra plotted in a narrow spectral region from 2.35 to 2.75 ppm. Spectrum number 52 showed the highest mean correlation with all other spectra belonging to both classes. This spectrum had the highest correlation with all the spectra from class 2 also. Among the considered and tested targets, spectrum number 3 was also included as it

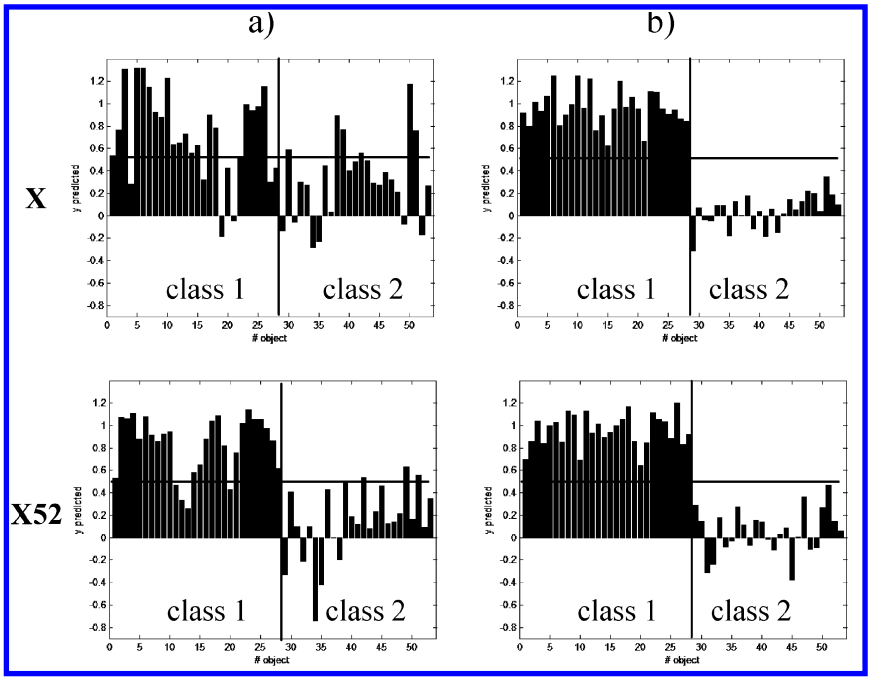

Figure 9. The *y* values predicted with the PLS (a) and UVE−PLS (b) models for the original and warped spectra; the first 28 objects belong to class 1, and all the remaining ones belong to class 2. In an ideal case, *y* predicted for all the objects from class 1 ought to equal 1, whereas *y* predicted for class 2 ought to equal 0.

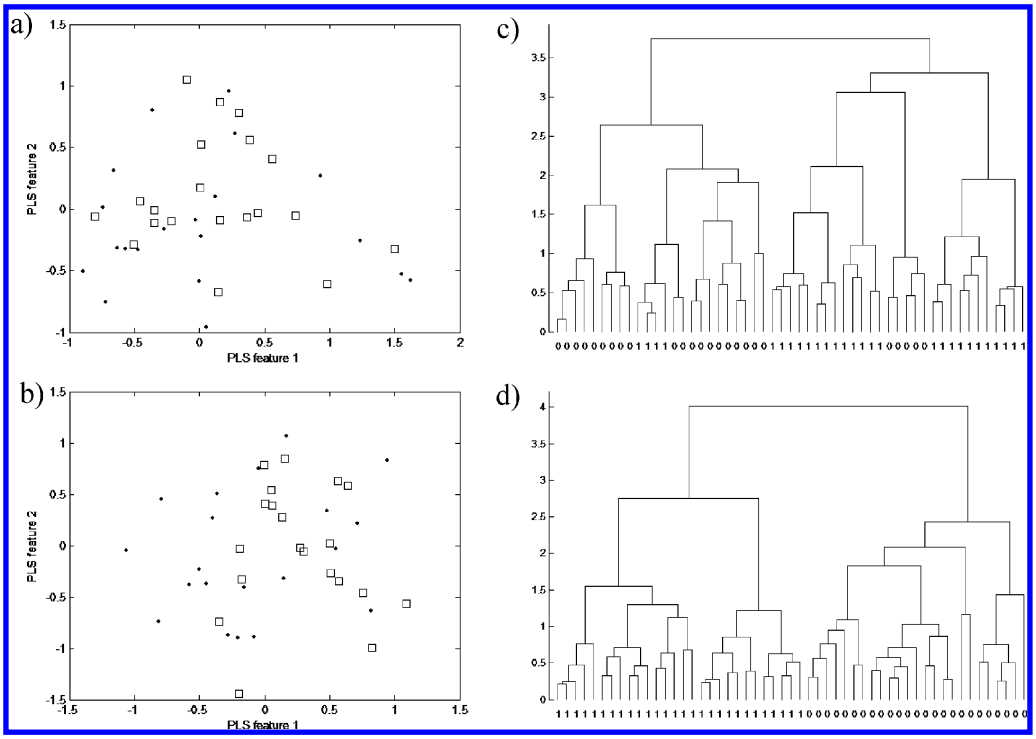

Figure 10. Projections of spectra on the plane defined by the first two latent variables of the UVE−PLS models constructed for (a) original data, **X**, and (b) warped signals, **X52**; dendrograms of 53 urine samples (constructed on the basis of the Euclidean distance and the Ward's method of objects grouping) in the UVE−PLS space for the original spectra (c) and warped spectra (d).

had the highest correlation with all the spectra belonging to class 1. However, as the target choice did not influence the final results, further results and discussion were limited to spectrum number 52 as the target spectrum (see Figure 4).

The time required to align 53 spectra to the selected target was ca. 125 s; that is, it appeared to be the fastest warping procedure among those compared and also showed a good performance (see Table 1 and Figure 5).

The results of FW can be improved by optimization of the number of the most intense peaks used for signal alignment. In the present study, the following numbers of the most intense peaks were tested: 10, 20, 30, ..., 100. For each signal, the optimal number of the most intense peaks was selected, on the basis of the value of the correlation coefficients of this aligned signal and the target spectrum. After the optimization process, the mean value of the correlation coefficients of all the signals with the target

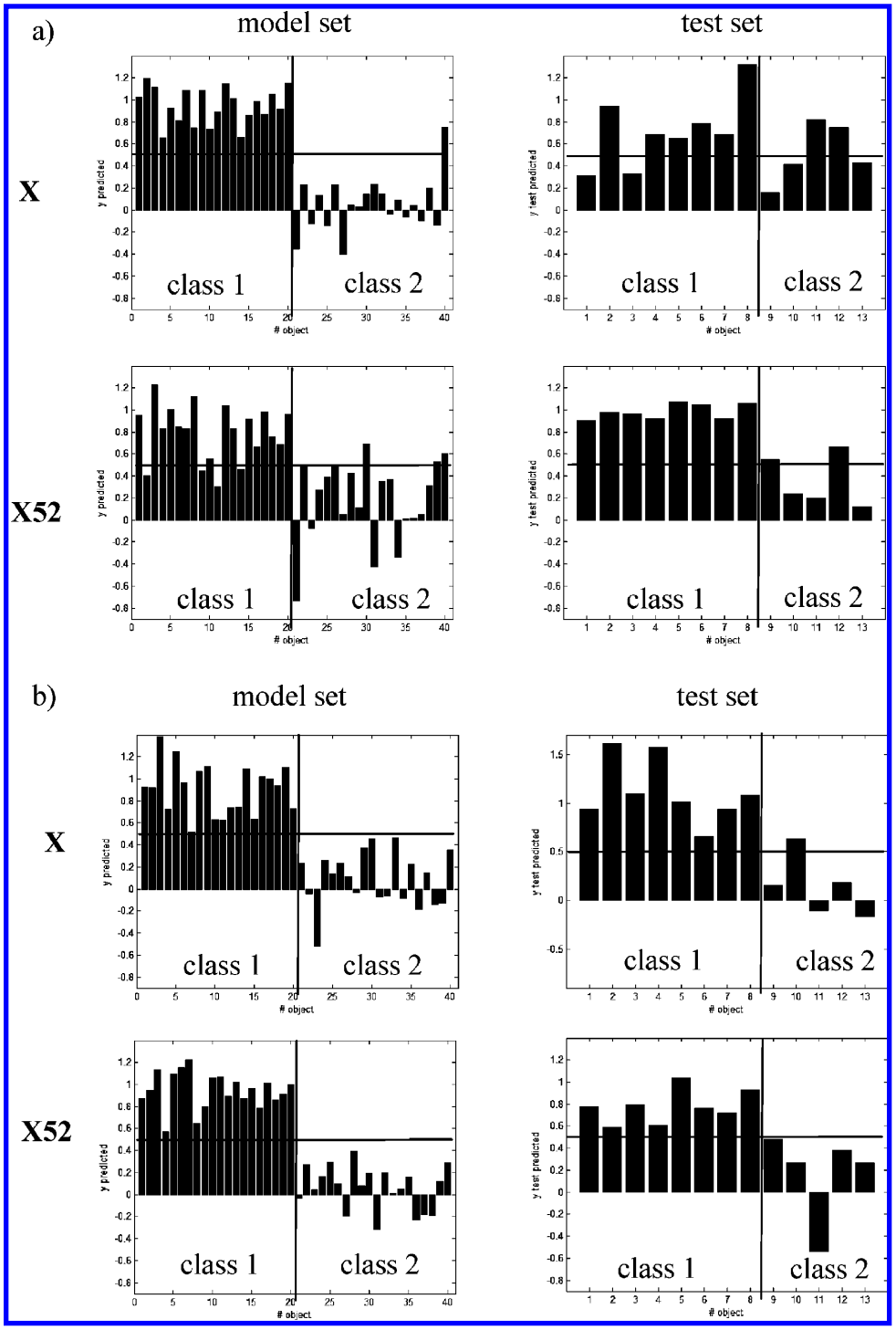

Figure 11. The *y* values predicted with the PLS (a) and UVE−PLS (b) models for the original, **X**, and warped spectra, **X52**. Model set: the first 20 objects belong to class 1, and the 20 remaining ones belong to class 2. Test set: the first eight objects belong to class 1, and the remaining five objects belong to class 2. In an ideal case, *y* predicted for all the objects from class 1 ought to equal 1, whereas *y* predicted for class 2 ought to equal 0.

spectrum increased to 0.90 (see Figure 6).

Prior to alignment, four spectra, numbers 12, 32, 36, and 47, showed a relatively low correlation with other signals (Figure 5a). After warping to spectrum number 52, their correlations with the other spectra increased, but they still remained relatively low. These four spectra did not correlate well with the target spectrum either (see Figure 6b).

Although warping procedures were applied to the spectra in the range from −0.2 to 10 ppm, the data exploration and classification was performed for normalized spectra, followed by the exclusion of the water, urea, and TSP regions, and organized in the form of matrix **X**. A set of spectra warped to spectrum number 52 is denoted as **X52**. After elimination of these regions, the correlation pattern changed significantly (see Figure 7). The mean value of the correlation coefficient increases from 0.64 for **X** to 0.90 for **X52** (see Figure 7b).

The effectiveness of warping can also be expressed as a decrease of the data variance from 96.8 for **X** to 34.0 for **X52**. The relative similarity among the original spectra and the warped ones can be traced on the principal com-

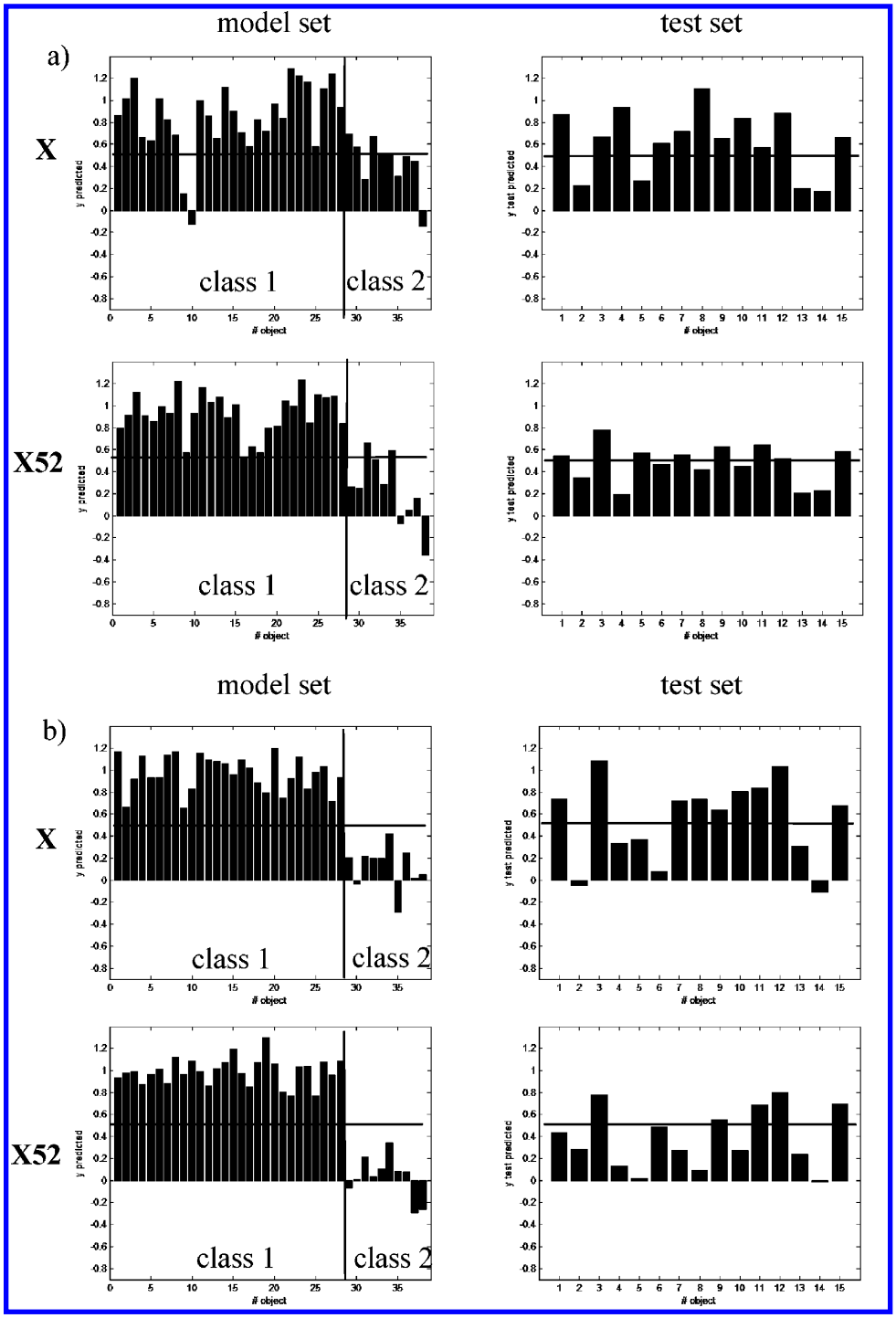

Figure 12. The *y* values predicted with the PLS (a) and UVE−PLS (b) models for the original, **X**, and warped spectra, **X52**. Model set: the first 28 objects belong to class 1, and the 10 remaining ones belong to class 2. Test set: all objects belong to class 2. In an ideal case, *y* predicted for all the objects from class 1 ought to equal 1, whereas *y* predicted for class 2 ought to equal 0.

ponents projections (see Figure 8). The first two principal components extracted for the unaligned signals describe 37.5% of the data variance, whereas those of the warped signals describe 42.2% of the data variance. In both cases, objects from the two classes overlap completely, but a high clustering tendency observed for the unaligned spectra on the plane defined by the first two PCs disappears after spectra warping.

The PLS discrimination models, constructed for the data sets studied (i.e., the unaligned and aligned spectra; **X** and **X52**), do not lead to satisfactory results (see Table 2). One of the possible reasons for unsatisfactory modeling is a high number of uninformative variables contained in the ^{1}H NMR spectra of the urine samples. High within-group variance but small covariance with the dependent variable *y* can be caused by natural diversity of the biological material or by an insufficient alignment of the processed spectra. Uninformative variables can be, however, eliminated by the UVE−PLS approach. The UVE−PLS models, constructed for the **X** and **X52** data sets, lead to the correct classification of all samples (see Figure 9). The observed lower complexity and the lower value of RMSCV for the **X52** set indicate

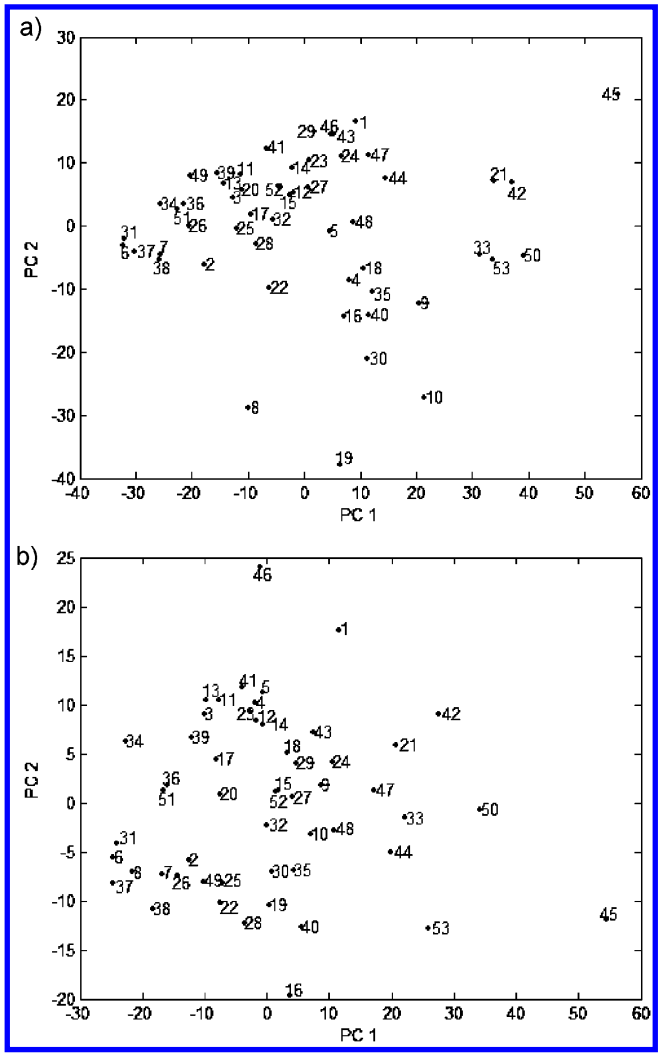

Figure 13. Projections of samples on the plane defined by the first two principal components for (a) bucketed data, **Xb**, and (b) bucketed warped signals, **X52b**.

that the model for this data set has a better discrimination power.

Projections of the unaligned and aligned signals on the planes, defined by the first two UVE−PLS features, are presented in Figure 10a and b. The remaining features (factors) were included in the models for a further reduction of the within-class variances. Dendrograms constructed for the spectra studied in the UVE−PLS model spaces, that is, based on the sample scores (using the Euclidean distance as a similarity measure and Ward's method of the subclusters grouping), are presented in Figure 10c and d. For the warped data set, two classes of objects are well separated in the two main branches of the dendrogram, whereas for the unaligned data set, subclusters of objects belonging to class 1 are mixed with subclusters of objects belonging to class 2.

As the studied data sets contain 53 objects only, the predictive power of the constructed models was evaluated on the basis of the leave-one-out cross validation and expressed as RMSCV. For better evaluation of the models' predictive ability, an analogous modeling was performed for the model set containing 40 objects (20 samples from class 1 and 20 samples from class 2) and the resulting models were evaluated for the independent test sets. The model set was selected using the Kennard and Stone algorithm,[31] to ensure data representativity, and all of the modeling steps, such as feature selection, were performed for the model set only. The independent test set contained eight samples from class 1 and five samples from class 2. Also in this case, PLS models were unsatisfactory, and only 69.23% of the objects from the test set were classified correctly for unaligned data and 84.6% for aligned data. The predictive ability of the UVE−PLS models greatly improved, but for the unaligned data, one object from the test set was still wrongly classified, whereas for the warped data, 100% of the objects from the test set were correctly classified (see Figure 11).

From a practical point of view, it was interesting to test the predictive power of the classification model, which was constructed for the model set containing only control and predose samples (class 1) and the samples with medium and high doses (0.25 and 0.50 mg/kg cycloheximide), collected on day 3 (class 2), while the remaining samples were kept as an independent test set. Unfortunately, the constructed UVE−PLS models for both the unaligned and aligned data sets did not properly predict whether the test samples belonged to class 2 (see Figure 12). This was probably associated with the fact that the model set was not representative, that is, not all data variations were included in the model.

All of the above considerations were limited to the original and warped signals (**X** and **X52**, respectively), but they can, however, be extended to the bucketed data (**Xb** and **X52b**) as well. The bucketing of the spectra was done by dividing the spectra into 0.02 ppm regions from 0 to 10 ppm using the Bruker Amix software.

A comparison of the score plots of the original and the warped data with the score plots of the bucketed original and the bucketed warped data (see Figures 8 and 13) reveals changes due to bucketing.

These changes are very significant for **Xb**, when compared with **X**, and less significant for **X52b**, when compared with **X52**. There are also some differences between the score plots of **Xb** and **X52b**. The first two principal components describe 68.6% and 69.9% of the data variance for **Xb** and **X52b**, respectively.

Differences between **Xb** and **X52b** can also be visualized in the plots representing the mean spectrum of **Xb** versus the mean spectrum of **X52b** and representing the standard deviation of **Xb** versus the standard deviation of **X52b** (See Figure 14).

Significant differences between the mean spectra of **Xb** and **X52b** are observed for features 152, 167, 149, 168, and 153, whereas the highest differences of the standard deviations of **Xb** and **X52b** are observed for features 152, 167, 168, 170, 171, 153, and 149. Only, for feature 152, the standard deviation of **X52b** is higher than the standard deviation of **Xb**, whereas for the remaining features—as it could easily be expected—bucketing causes the opposite effect.

For the bucketed warped data set, the PLS discrimination model alone leads to a 100% correct classification (see Table 2 and Figure 15), whereas the corresponding model, constructed for the bucketed original data, gives a 92.2% correct classification. Both UVE−PLS models of the bucketed data

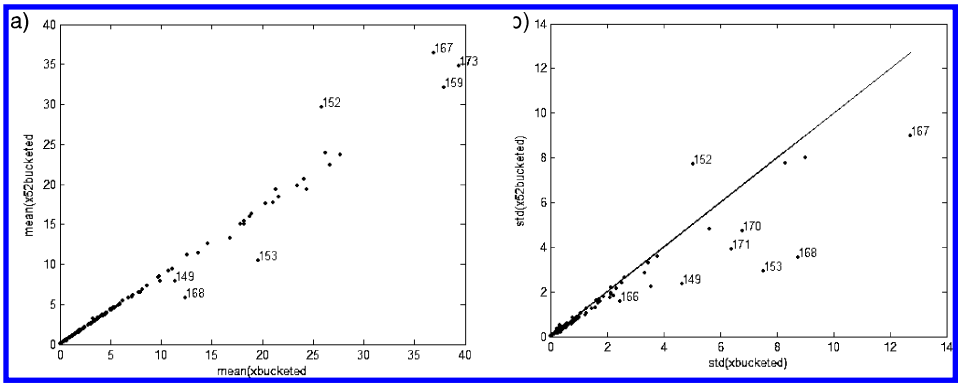

Figure 14. (a) Mean spectrum of the bucketed data, **Xb**, versus mean spectrum of the warped and bucketed data, **X52b**. (b) Standard deviation of the bucketed data versus standard deviation of the warped bucketed data.

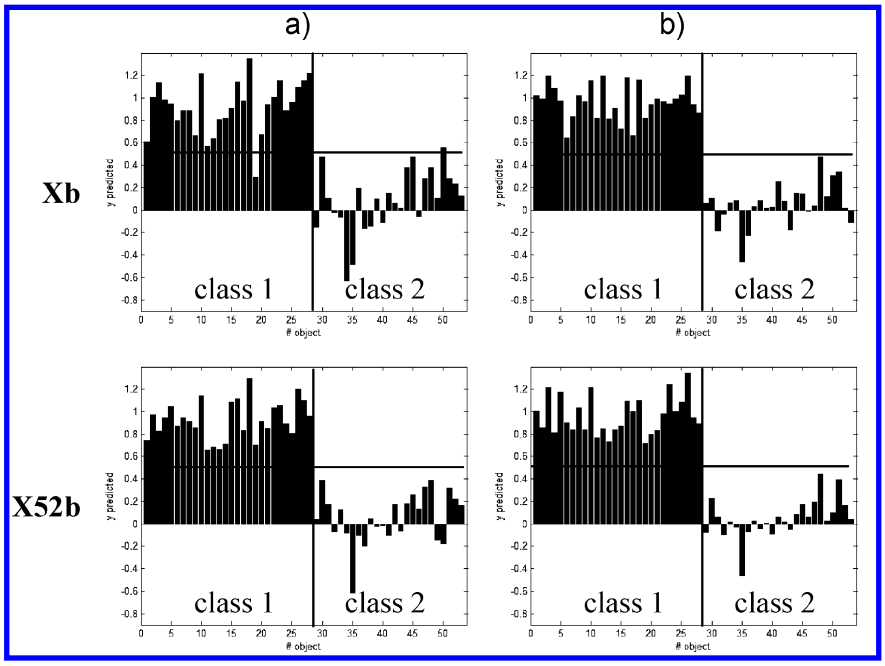

Figure 15. The *y* values predicted with the PLS (a) and UVE−PLS (b) models for the bucketed, **Xb**, and bucketed warped spectra, **X52b**; the first 28 objects belong to class 1, and all the remaining ones belong to class 2. In an ideal case, *y* predicted for all the objects from class 1 ought to equal 1, whereas *y* predicted for class 2 ought to equal 0.

sets, that is, the bucketed original and the warped data, are satisfactory.

In summary, only the PLS model, leading to a 100% correct classification, can be constructed for the bucketed warped data. UVE−PLS models for the original, warped, and bucketed data sets are all satisfactory, but the smallest complexity and the smallest RMSCV is achieved for **X52** (i.e., for the warped spectra; see Table 2).

CONCLUSIONS

The fuzzy warping algorithm, applied to the alignment of the two NMR spectra (containing ca. 30 000 data points each) requires less than 3 s, and it can therefore be used on-line. The warping procedure leads to a significant increase of correlation among the spectra and to a significant reduction of data variance. Spectra warping, followed by bucketing, allows a successful PLS classification of the urine samples of the control and drug-treated animals.

For the warped only or bucketed only spectra, the construction of a proper classification model is possible after elimination of the noninformative variables. The best discrimination model (with the lowest complexity and the smallest RMSCV) is, however, the UVE−PLS model, constructed for the warped spectra.

REFERENCES AND NOTES

(1) Ott, K. H.; Arabinar, N.; Singh, B.; Stockton, G. W. Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts. *Phytochemistry* **2003**, *62*, 971−85.

(2) Le Gall, G.; Puaud, M.; Colquhoun, I. J. Discrimination between orange juice and pulp wash by ^{1}H nuclear magnetic resonance spectroscopy: Identification of marker compounds. *J. Agric. Food Chem.* **2001**, *49*, 580−588.

(3) Brescia, M. A.; Kosir, I. J.; Caldarola, V.; Kidric, J.; Sacco, A. Chemometric Classification of Apulian and Slovenian Wines using 1H NMR and ICP-OES together with HPICE Data. *J. Agric. Food Chem.* **2003**, *51*, 21−26.

(4) Choi, H.-K.; Choi, Y. H.; Verberne, M.; Lefeber, A. W. M.; Erkelens, C.; Verpoorte, R. Metabolic fingerprinting of wild type and transgenic tobacco plants by 1H NMR and multivariate analysis technique. *Phytochemistry* **2004**, *65*, 857−864.

(5) Nicholson, J. K.; Lindon, J.; Holmes, E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **1999**, *29*, 1181−1198.

(6) Nicholson, J. K.; Connelly, J.; Lindon, J. C.; Holmes, E. Metabonomics: a platform for studying drug toxicity and gene function. *Nat. Rev. Drug Discovery* **2002**, *1*, 153−161.

(7) Forshed, J.; Schuppe-Koistinen, I.; Jacobsson, S. P. Peak alignment of NMR signals by means of a genetic algorithm. *Anal. Chim. Acta* **2003**, *487*, 189−199.
(8) Walczak, B.; Wu, W. Fuzzy warping of chromatograms. *Chemom. Intell. Lab. Syst.* **2005**, *77*, 173−180.
(9) Sinkhorn, R. A. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Stat.* **1996**, *35*, 876−879.
(10) Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *J. Chromatogr., A* **1998**, *805*, 17−35.
(11) Pravdova, V.; Walczak, B.; Massart, D. L. A comparison of two algorithms for warping of analytical signals. *Anal. Chim. Acta* **2002**, *456*, 77−92.
(12) Kassidas, A.; MacGregor, J. F.; Taylor, P. A. Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J.* **1998**, *44*, 864−875.
(13) Eilers, P. Parametric time warping. *Anal. Chem.* **2004**, *76*, 404−411.
(14) Brown, T. R.; Stoyanova, R. NMR spectral quantitation by principal-component analysis. II. Determination of frequency and phase shifts. *J. Magn. Reson., Ser. B* **1996**, *112*, 32−43.
(15) Wijes, H.; Melssen, W. J.; in't Zandt, H. J. A.; van der Graaf, M.; Heerschap, A.; Buydens, L. M. C. Authomatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in larger spectral data sets. *J. Magn. Reson.* **2000**, *144*, 35−44.
(16) Massart, D. L.; Vandeginste, B. G. M.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part A*; Elsevier: Amsterdam, 1997.
(17) Vandeginste, B. G. M.; Massart, D. L.; Buydens, L. M. C.; de Jong, S.; Lewi, P. J.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B*; Elsevier: Amsterdam, 1997.
(18) Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Data by the Use of Cluster Analysis*; John Wiley & Sons: New York, 1983.
(19) Vogt, W.; Nagel, D.; Sator, H. *Cluster Analysis in Clinical Chemistry; A Model*; Wiley: New York, 1987.
(20) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data; An Introduction to Cluster Analysis*; Wiley: New York, 1990.
(21) Romesburg, H. C. *Cluster Analysis for Researchers*; Lifetime Learning Publications: Belmont, CA, 1984.
(22) Martens, H.; Naes, T. *Mutivariate Calibration*; John Wiley & Sons: Chichester, U.K., 1989.
(23) Martens, H.; Martens, M. *Multivariate Analysis of Quality*; John Wiley & Sons: Chichester, U.K., 2001.
(24) Naes, T.; Isaksson, T.; Fearn, T.; Davies, T. *Multivariate Calibration and Classification*; NIR Publications: Chichester, U.K., 2002.
(25) Dayal, B. S.; MacGregor, J. F. Improved PLS algorithms. *J. Chemom.* **1997**, *11*, 73−85.
(26) Rannar, S.; Lindgren, F.; Geladi, P.; Wold, S. A PLS kernel algorithm for data sets with many variables and fewer objects: part 1: Theory and algorithm. *J. Chemom.* **1994**, *8*, 111−125.
(27) Lindgren, F.; Geladi, P.; Wold, S. The kernel algorithm for PLS. *J. Chemom.* **1993**, *7*, 45−59.
(28) Wu, W.; Manne, R. Fast regression methods in a Lanczos (or PLS-1) basis. Theory and applications. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 145−161.
(29) de Jong, S. SIMPLS: an alternative approach to partial least-squares regression. *Chemom. Intell. Lab. Syst.* **1993**, *18*, 251−263.
(30) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B.; Sterna, C. Elimination of uninformative variables for multivariate calibration. *Anal. Chem.* **1996**, *68*, 3851−3858.
(31) Kennard, R. W.; Stone, L. A. Computer aided design of experiments. *Technometrics* **1969**, *11*, 137−148.

CI050316W