Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Pharmaceutical and Biomedical Analysis 38 (2005) 824–832
JOURNAL OF PHARMACEUTICAL AND BIOMEDICAL ANALYSIS
www.elsevier.com/locate/jpba

A comparison of methods for alignment of NMR peaks in the context of cluster analysis

Jenny Forshed [a], Ralf J.O. Torgrip [a,b], K. Magnus Åberg [a], Bo Karlberg [a], Johan Lindberg [b], Sven P. Jacobsson [a,c,*]

[a] *Department of Analytical Chemistry, Stockholm University, SE-10691, Stockholm, Sweden*
[b] *Safety Assessment, AstraZeneca R&D Södertälje, SE-15185, Södertälje, Sweden*
[c] *Analytical Development, Pharmaceutical and Analytical R&D, AstraZeneca R&D, SE-15185, Södertälje, Sweden*

Received 24 September 2004; received in revised form 29 October 2004; accepted 30 January 2005
Available online 2 April 2005

Abstract

This paper compares the performance of two recently developed algorithms and methods for peak alignment of first-order NMR data of complex biological samples. The NMR spectra of such samples exhibit variations in peak position and peak shape due to variations in the sample matrix and to instrumental instabilities. The first method comprises an alignment of spectral segments with linear interpolation and shift correction to accommodate correspondence between a target and a test spectrum by a beam search or genetic algorithm. The second method is based on peak picking and needle vector representation of the NMR data with subsequent breadth-first search to establish shift corrections between the target and the test spectrum.

The two proposed peak alignment methods and their respective merits are discussed for a real metabonomics application. Both alignment methods have been shown to enhance the interpretability of the resulting multivariate models, thereby increasing the prospect of detecting and following the onset of subtle biological changes reflected in the NMR data.

Keywords: Peak shift; Peak alignment; NMR; Metabolomics; Metabonomics; Multivariate analysis; Bucketing

1. Introduction

NMR has many attractive features, making it a useful tool both for quantitative and qualitative analysis. NMR and pattern recognition techniques are indispensable combination tools frequently employed in systems biology and in the pharmaceutical industry. To exploit the full information content of NMR data acquired on the complex systems found in these research areas, various multivariate data analysis methods based on variance mapping have been implemented [1–6]. The common denominator for all of these methods is the necessity of proper data pre-processing prior to data analysis. Several approaches are available that deal with un-desired spectral artefacts, e.g. multiplicative scatter correction [7] for baseline correction, orthogonal scatter correction for removal of unwanted variances [4,8] and corrections for relative intensity variations such as normalization, mean-centring and autoscaling. However, the variation in the abscissa has been dealt with to a lesser extent.

If NMR peaks are un-aligned, any variance-mapping model generated will exhibit spurious artefacts, making it more complex, more difficult to interpret and possibly misleading [9]. Consequently, the usefulness of the results

Abbreviations: FID, Free Induction Decay; FWHM, Full Width at Half Maximum; GA, Genetic Algorithm; MCS, Measure of Class Separation; NR, Needle Representation; PARS, Peak alignment by Reduced Set mapping; PC, Principal Component; PCA, Principal Component Analysis; PLS-DA, Partial Least Squares-Discriminant Analysis; SWA, Segment-Wise Alignment

* Corresponding author at: Department of Analytical Chemistry, Stockholm University, SE-10691, Stockholm, Sweden. Tel.: +46 8 553 289 68; fax: +46 8 553 277 30.
E-mail address: sven.jacobsson@astrazeneca.com (S.P. Jacobsson).

0731-7085/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpba.2005.01.042

achieved will decrease. Examples of problems and artefacts originating from misaligned NMR signals have previously been reported [5,6,10–14]. From a multivariate point of view, it is interesting to note that any NMR peak originating from identical analytes should (between samples) be a rank-one peak: i.e. symmetric (Lorenzian) with identical *x* position and FWHM (full width at half maximum). If this is not the case, any deviations from this ideal will be interpreted by the model as mappable variance and will thus contribute to the model, even though the significance of the variance is less pronounced.

To overcome the peak shift problem in NMR, different approaches have previously been suggested, of which the predominant method is integration of predetermined spectral segments, i.e. bucketing, typically 0.04–0.07 ppm wide [1–4]. The bucketing approach deals with the peak shift problem but ruins the resolution of the acquired data and will certainly confound variance contributions from small peaks with variance contributions from large peaks in the same bucket. This makes multivariate detection of small variances (peaks) virtually impossible if they are in the same bucket as larger variances. Furthermore, one main advantage of the peak alignment methods over bucketing is interpretability. The loadings from the multivariate analysis will reveal the peaks responsible for the clustering, as is clearly shown in [6].

More refined alignment methods (than bucketing) for NMR data have been reported, one example being partial linear fit (PLF), a method outlined by Vogels et al. [15]. PLF automatically picks out segments in an NMR spectrum of size *d* depending on the peak frequency and shifts them *s* points left and right. Each possible and relevant combination of *d* and *s* is tried until the sum of squared differences between the spectrum and the target is minimized. Other approaches, such as the one outlined by Brown and Stoyanova [10], perform automatic removal of frequency shifts in NMR spectra by using PCA to determine the misalignment in a single peak across a series of spectra. This method has been extended and applied to in vivo NMR spectra by Witjes et al. [16], and has been recently further developed and applied to high-resolution spectral data [14].

It is hard to ascertain the quality of the results when alignment of first-order real data is performed, the reason being that no other data (as in second-order data) exist to guide or validate the alignment. A target, e.g. one spectrum, must be chosen for the alignment and the elucidation can only be made by comparisons with other existing techniques such as bucketing.

In this paper two different approaches to peak alignment, published in 2003 [6,17], are compared with respect to the qualitative class analysis of metabonomics samples. The impact of the alignment methods is critically examined when the aligned data are modeled by multivariate methods such as principal component analysis (PCA) and partial least squares discriminant analysis (PLS-DA) [18].

2. Theory and methods

2.1. NMR data

The metabonomics NMR data fully described in reference [6], were utilized for evaluation of the proposed method performances of the peak alignment methods. The data consist of urine spectra from two groups each of 6 rats: one control group orally dosed with water and one group dosed with citalopram (positive control for phospholipidosis). The rats were dosed once a day for 14 days and urine samples were collected on days–5 (pre-dose, two occasions), 1, 3, 7, 10 and 14. The samples were buffered (0.2 M phosphate solution) and centrifuged (13,000 rpm for 10 min) prior to ^{1}H-NMR analysis on a Bruker DRX600 instrument working at 600.13 MHz. A suppression of the water signal was achieved using pre-saturation during relaxation delay and mixing time with a shaped pulse for selective saturation. One of the 84 resulting ^{1}H-NMR spectra is shown in Fig. 1.

Comparing two urine spectra from the same class in the data set will reveal many differences. Various background matrices reflecting the varying urine concentration levels in the samples will influence different peaks in various ways. This may result in some peaks differing from one spectrum to another, although they are chemically and analytically equal. The magnetic field homogeneity was automatically adjusted for each sample in this study, although due to the variations in the background matrix between samples this could not be done with equal quality. Poor magnetic field homogeneity may give broad or asymmetric peaks. Other instrumental parameters, such as temperature, may also influence the peaks. Furthermore, different patterns of peaks will appear due to individual variations within a group of rats. The pH of the sample is a major source of variation in peak positions [12]. Even if the samples are buffered, small pH differences will be detected [1–3,19].

Samples from the same day should cluster together. The day clusters should also move away from the control, indicat-

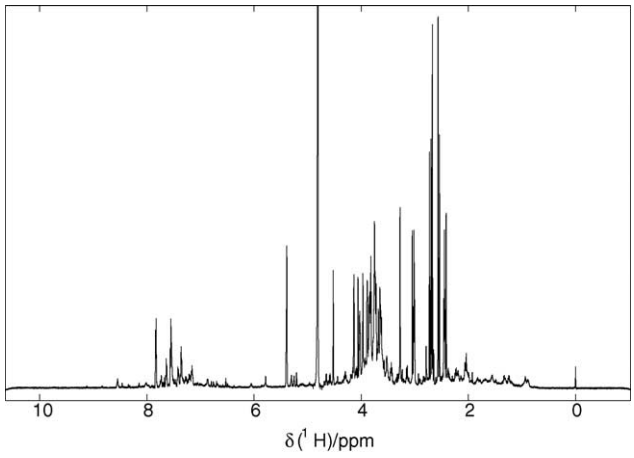

Fig. 1. Original ^{1}H-NMR spectrum of urine sample from rat dosed with citalopram, day 7 of study.

ing the onset of the toxic event and the progress of the lesion. Days 7–14 will be regarded a separate group since it is most likely that the responses have reached a steady state after 6 days. The observed differences between the groups of spectra will probably be traceable back to the metabolites stemming from the xenobiotic. Further analysis and identification of the detected peaks has not been pursued in this paper.

The acquired FID was zero-filled to yield 65,536 data points per spectrum and a reduction, down-sampling, of data to some extent will not result in loss of information, although the traditional bucketing at 0.04–0.07 ppm (typically resulting in ∼250 data points) gives a rough reduction. The two peak alignment methods presented here reduce the spectral data in two conceptually distinct ways. The segment-wise peak alignment method reduces spectra after the peak alignment with a bucketing approach and the appropriate size of the buckets are tried out and evaluated. PARS is initiated with and highly dependent on a peak-detection algorithm resulting in a sparse or needle representation of spectra, which reduces data to an arbitrary choice of resolution.

2.2. Segment-wise peak alignment

The segment-wise peak alignment (SWA) can be described as a segmented non-peak picking approach, with shift correction and linear interpolation to accommodate correspondence between target and test spectrum. The target spectrum is chosen as the spectrum supposed to reflect all possible peaks, i.e. a spectrum from a dosed rat, in this case the spectrum shown in Fig. 1. Each spectrum will comprise unique peaks or artefacts, found only in one or a few spectra; however, if the segments in the alignment are chosen so as to be wide enough, a pattern in the segment will be recognized between the target and the sample and one occasional peak will thus not influence the evaluation measure. The size of this possible peak as well as the evaluation function during the alignment will matter. Also, the maximum possible range of shifting and interpolation are enclosed to avoid misalignments in such cases.

The previously described peak alignment by genetic algorithms [6] involves dividing spectra into several segments, a minimum size of the segments being predetermined. Every segment is shifted sideways and stretched or shrunk by linear interpolation to fit a corresponding section of the target spectrum. To avoid damaging peaks during the segmentation, points with minimum intensity of both the spectrum for alignment and the target spectrum are automatically determined to be positioned in a low-intensity area of the data (within a predetermined data window) to avoid affecting the subsequent data analysis. This means that every segment will have a unique size depending on the spectral topography and will also have a unique number of added or removed points governed by the alignment algorithm. One example of SWA of two segments is shown in Fig. 2.

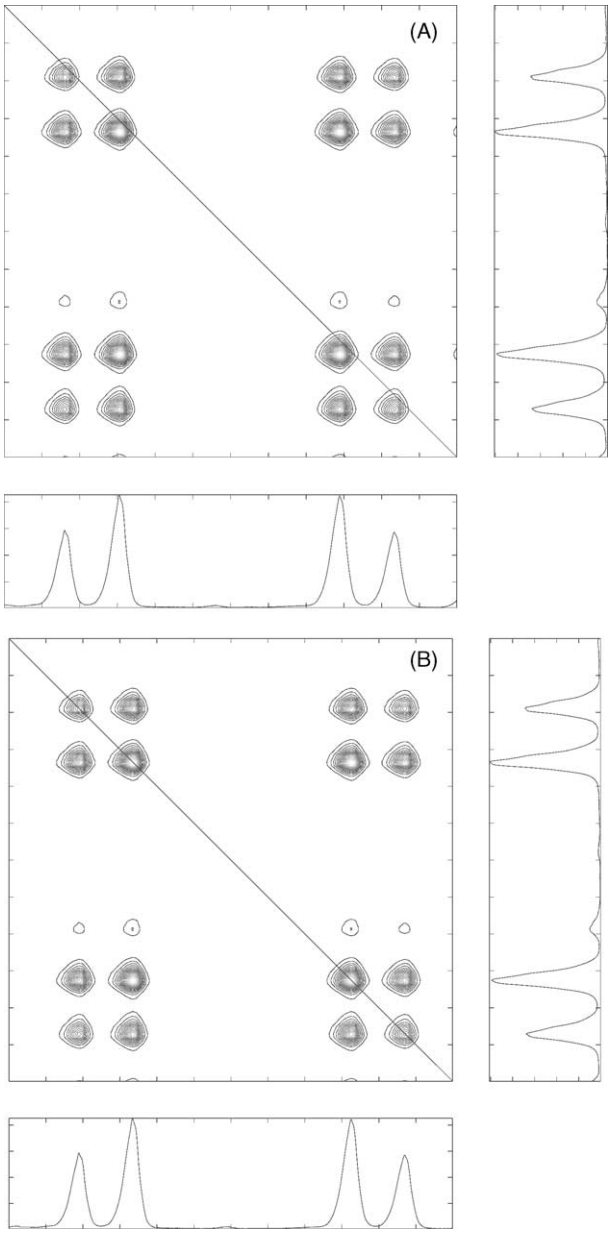

Fig. 2. Contour map of the cross-correlation of the citrate-peak area (2.84–2.32 ppm) in metabonomics NMR spectra, which represents two segments in the segment-wise peak alignment algorithm, before (A) and after (B) alignment. The diagonal line marks where the contours of the peaks from the two spectral segments should meet when they are aligned, as seen in (B).

The optimal correction of segments is carried out using genetic algorithms. An extension of this segmented approach, based on the beam search algorithm, is also presented here [20].

2.2.1. Segment-wise peak alignment by a genetic algorithm

Genetic algorithms (GAs) constitute a family of optimization methods introduced by John Holland [21]. They can be viewed as an evolutionary optimization process in which a population of candidate solutions to a problem

evolves over a sequence of generations. During each generation the fit of the solutions is evaluated. A better solution will have a higher probability of surviving and breeding as the GA proceeds. This means that not all possible solutions are evaluated, the best solutions in a set of candidate solutions being favored while the GA proceeds. Comprehensive treatments of GAs can be found in references [21–23].

The application of GAs to the alignment problem is done by allowing the parameters of the shifting, stretching and shrinking of a sample segment to evolve in an evolutionary manner to give the best fit with the target. The fit is evaluated as the correlation coefficient between the target and the sample. When all segments in the test spectrum are aligned, the segments are reassembled to form a reconstructed, aligned spectrum [6].

2.2.2. Segment-wise peak alignment by a beam search algorithm

To speed up the evolution of the alignment algorithm described above, Lee and Woodruff [20] proposed a beam search algorithm. This turns out to perform at least as well as the proposed genetic algorithm, though about seven times faster. Beam search is a heuristic search technique where a number of nearly optimal alternatives ("the beam") are examined [24]. Initially, a set of likely solutions is created on the edge of a pre-determined search radius and subsequently evaluated. The algorithm then assigns one or more new candidate solutions by selecting the k best steps from the current trial solution(s) where k is the parametric input beam width. For each loop of the algorithm, the radius is reduced until a stop criterion is met. The best solution of the end population is then finally reported.

2.2.3. Further development of the SWA method

Interpolation of spectral segments may introduce a change in peak area. To ascertain its influence on the data analysis, this part of the algorithm was excluded. Excluding the interpolation will give a lower correlation coefficient between spectra but preserve the peak area information. Furthermore, the results of class separation depending on the type of search algorithm (beam search or genetic algorithm) were studied.

Since rat urine NMR data is relatively noisy, it was assumed that appropriate bucketing after peak alignment would simplify the data but not decrease the latent information content. Bucketing in this case will, as in previous studies [1–4], reduce differences in data due to instrumental and background matrix differences, but can now be performed with much narrower buckets since the peak shifts do not have to be accounted for. The appropriate sizes of the buckets were tried out and compared to classical bucketing with 0.04 ppm per bucket. The algorithms for the work with the SWA method were implemented in Matlab [25] and the code is available upon request from the authors.

2.3. PARS–peak alignment by reduced set mapping

The PARS method exploits possibilities of sparse or "needle" representation (NR) and uses a fast tree-search algorithm with early pruning of alignment solutions, i.e. the spectral representation of the peaks is transformed into a sparse vector of zeros with peak maximum intensity found at the x-axis locations (ppm or frequency) corresponding to the peak maxima (see Fig. 3).

The NR is well suited for analysis by, for instance, breadth-first search (BFS) algorithms, making the method fast. The method relies on using one spectrum or some other representative shape within the data as the target and then evaluating combinations of sets of restricted, possible corrections of the unaligned data (test sample) to minimize a global combination search. The NR-mapped representation of the data is subjected to a BFS search to yield a correction scheme and the corrections found are made to the NR of the data. Furthermore, the algorithm used in this paper uses the intensity information as guide for the assessment of the best solution in the search space. The NR further opens the way for other interesting possibilities such as using recursively updated target vectors, i.e. the method assesses the currently aligned test spectrum for un-matched (un-aligned) peaks not present in the target. The new peaks are then subsequently inserted in the target to yield an updated target feature. In this way the target reflects information about all the spectra that have previously been aligned. The "recursive target update" feature used here is one of the strengths of the NR representation and circumvents the problem of target spectrum selection. The NR representation is also interesting from a data analysis point of view since the variable domain of resulting model (loadings) is annotated by the *true ppm shift* of the NMR peaks, thereby simplifying the interpretation of the model.

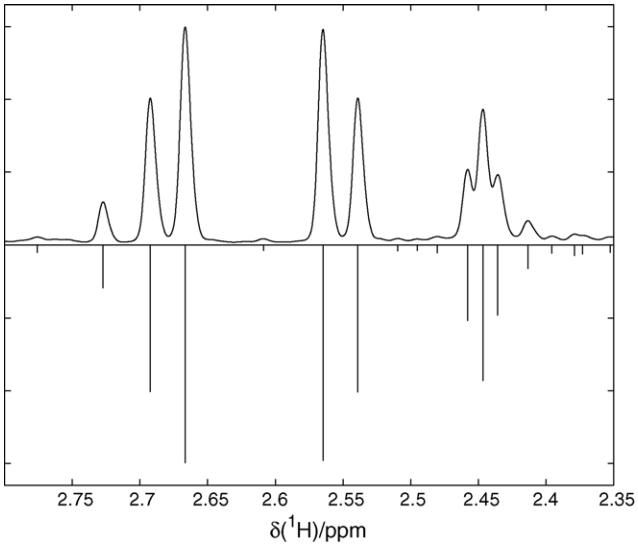

Fig. 3. A segment from the original NMR spectrum (positive) with the needle representation (negative for visualization) used in the peak alignment by reduced set mapping, PARS.

The main reason for using this alignment method is that the spectra can be aligned at an arbitrary resolution. It should be noted that the method presented here differs slightly from the original one reported [17]. In this work we used the NR both for shifting the data *and* also for the subsequent data analysis. The NR representation has pronounced advantages since any differing FWHMs or shim problems (between-sample) will effectively be avoided.

The NMR data were transformed and aligned using the NR at a resolution of 0.004 ppm (2500 data points/NMR spectrum [0–10 ppm]), which is slightly larger than the practical working resolution of the NMR spectrometer in question. PARS was invoked using intensity information. The mismatch penalty was invoked using a Harrington-type of desirability function [26] where the match error was expressed as:

$$S_d = \left(\frac{|p_x - p_t|}{2w}\right)^{P_{\text{loc}}} \quad (1)$$

$$S_I = M_w \left(\frac{|I_x - I_t|}{\max([I_x, I_t])}\right)^{P_{\text{rel}}} \left(\frac{\max([I_x, I_t])}{I_{\max}}\right)^{P_{\text{abs}}} \quad (2)$$

$$S_{\text{tot}} = S_d + S_I \quad (3)$$

where S_d (Eq. (1)) is the penalty for the (mis)location, p_x and p_t are the locations of the sample and target peak, respectively, w the search window and P_{loc} the exponent (weight) for the location penalty. Likewise, S_t (Eq. (2)) is the penalty for the intensity, I_x and I_t are the intensities of the test and target peak, respectively, P_{rel} and P_{abs} are the weights for the relative and absolute intensity part of the penalty, $I_{\max}$ the maximum intensity for the set in question (target and test) and, finally, M_w the weight balance for the intensity/location penalty. The penalty for each possible match is then calculated by Eq. (3).

The data were aligned using $P_{\text{loc}} = 1$, $P_{\text{rel}} = 1$, $P_{\text{abs}} = 0$ and $M_w = 1$. These parameters gave the most straightforward and simple implementation and were not further optimized. The maximum allowed alignment distance was $w = 0.1$ ppm. The alignment resulted in an 84×2500 matrix to be modeled by the multivariate analysis methods.

The peak alignment of the data was performed using the PARS algorithm with an in-house written code for Matlab [25].

2.4. Target spectrum

In any peak alignment method a target must be chosen, and various alternatives are possible. When the same target is used for all the spectra in a data set, all the data may be illustratively represented by score plots from multivariate analysis of the whole data set that show differences between groups of spectra, assuming that peaks appearing from the same substance appear at the same position in the *x*-direction of the spectra after the alignment. An unknown sample would be aligned to the previously chosen target and projected into the multivariate model space (loadings) for visual interpretation of the resulting scores. This is the case in the SWA method [6], where one target spectrum is assumed to reflect all peaks of interest and is therefore chosen as a target for all the spectra in the study, as well as in PARS [17], where one spectrum is chosen as the start for the alignment according to the "recursive target update". When interpreting the data with nominal classes, we could also use this class information. All the spectra in one class may be aligned to one chosen target in its own class, e.g. as in Andersson et al. [27]. To classify an unknown sample with this method, this sample must be aligned to each class-target spectrum separately and the probability of belonging to that class must be calculated after alignment, e.g. by SIMCA [28]. The interpretation with all classes in one multivariate model will not be valid in this case.

2.5. Comparison of SWA and PARS

Since the two methods for alignment described above differ conceptually, it is difficult to arrive at a fair comparison between them. Here, the classification is evaluated after the peak alignment has been performed. This is done by computing the distance between classes of spectra using the measure of class separation (described below) and evaluating the interpretability in a multivariate space. PCA was used as the method of compression/projection for generating the reduced data space for interpretation, and PLS-DA [18,29] was applied as a supervised classification method where the distance between the groups was measured. The results from the two different types of peak alignment methods described above are also compared to using the full raw spectra and the classical approach of bucketing.

2.5.1. The measure of class separation

The quality of the alignment was evaluated by projecting the entire data set into the multivariate models PCA and PLS-DA. From the resulting score vectors, two vectors showing a good separation between classes were chosen and in this space each class was approximated by a bivariate normal probability distribution. The boundary between a pair of classes is defined as the hypersurface (curve in 2D) on which the probability densities of the two classes are equal. The measure of class separation is calculated as the minimum Mahalanobis distance between the class boundary and the class centre. Since there are two such distances for each pair of classes, the shortest distance of them all was chosen as the measure of class separation. For simplicity, the abbreviation MCS is used below [30]. All calculations were performed in MATLAB [25].

2.5.2. Data pre-treatment

From the NMR study, all the samples from control rats and the samples from day–5 (pre-dose) were selected as a single group, the control group. Samples from the citalopram group (day 1, day 3 and days 7–14) represent three dosed groups. The NMR spectra were area-normalized to equal area (water peak excluded) and centred to zero mean prior to multivariate

modeling. The resulting spectra from PARS were autoscaled (to unit variance) within variables. One outlier in the spectral raw data due to instrumental error was detected and excluded. Another outlier was detected by visual inspection of the first two PCA scores after PARS had been applied; it was excluded on the basis that it differed from the group with approximately 10 standard deviations.

3. Results and discussion

3.1. Segment-wise peak alignment

Fig. 4 shows the results of the PLS-DA classifications. Each plot illustrates how different segment sizes for the bucketing after alignment compared to the classical bucketing of raw spectra influences the distances between the different dosed groups and control. The peak alignment result from the beam search and the genetic algorithm are here compared, together with the influence of linear interpolation. For day 1 with 512 or more than 8192 buckets, excluding the interpolation in the peak alignment clearly shows the best separation. This is probably due to loss of area information in the interpolation step. The rather "noisy" results from day 1 may be due to the fact that the group is heterogeneous because of the rats having different response rates. For days 3–14, any alignment method results in better group discrimination, compared to the raw data, if the bucketing exceeds 512 buckets. It is notable that no significant difference at any rate in class discrimination is detected when the data are reduced down to 8192 buckets, which, according to this analysis, may be considered to be the "real" resolution. It is also noteworthy that all the alignments were repeated with exactly the same results.

From this investigation one case was chosen for further analysis by PCA and PLS-DA. Since the interpolation step in the alignment did not improve the class separation in all cases, and the beam search algorithm works about seven times faster with results equal to the genetic algorithm; the beam search peak alignment with only shift correction was henceforth used. These calculations will take about 2 s for a metabonomics NMR spectrum of 65,536 data-points, partitioned in about 100 segments, on a PC with a processor of 2.4 MHz and 512 MB RAM. According to these results, the number of buckets should be 512 from days 1–3 but 286 according to the best results from days 7–14. To reduce the risk of loosing information, 512 buckets were used in the following PCA and PLS-DA analysis.

It should be noted that other search algorithms (than beam search) might be better suited to finding the optimum peak alignment in this one-dimensional search (sideways movement).

3.2. Comparing SWA and PARS

The main difference between the two methods for peak alignment is the peak picking. In the PARS method the peaks are defined, whereas in the SWA method the morphological information in the spectrum is preserved. One risk with the SWA is that the peaks are not defined and parts of the peaks may be damaged if a segment is moved too far. On the other hand, there is no need to define the appearance of a peak,

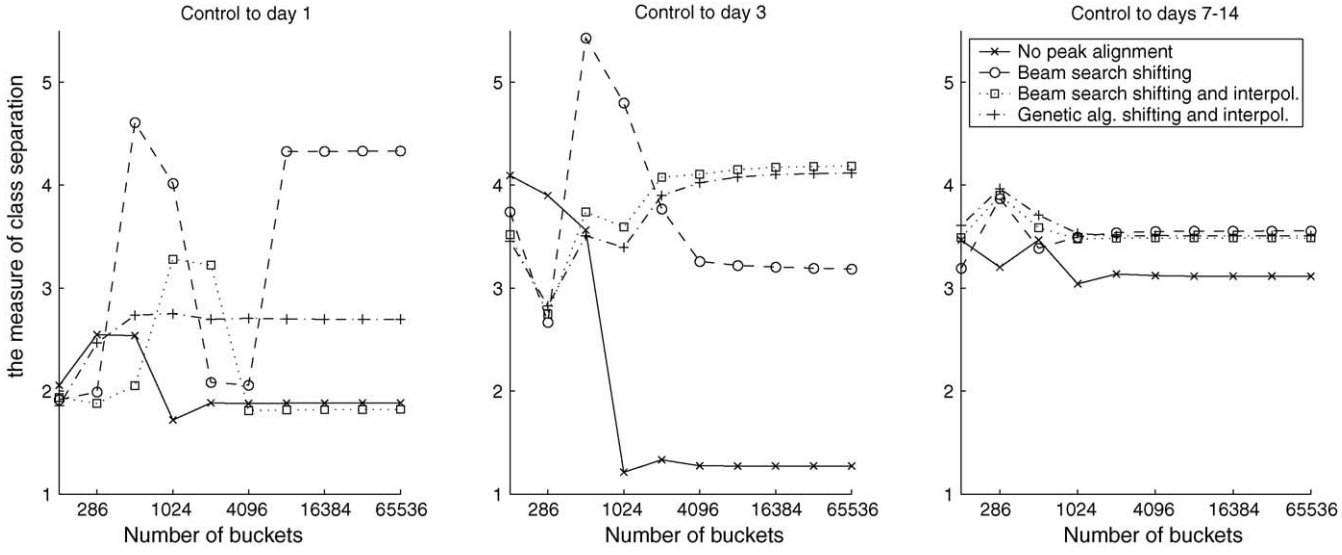

Fig. 4. The measure of class separation denotes the distance between groups after PLS-DA (2 PLS components). The *x*-axis denotes the number of buckets for each analysis, PLS-DA having been evaluated at 128; 286 (0.04 ppm/bucket); 512; 1024; 2048; 4096; 8192; 16,384; 32,768 buckets and full spectra (denoted as 65,536 buckets). The control samples in the study are all the samples from control group (days −5 to 14) and samples from day −5 from the dosed group. The dosed samples are represented by samples from day 1, day 3 and days 7–14, respectively in the dosed group. The four different graphs in the three plots represent: no peak alignment (×), beam search peak alignment including sideways shifting of segments only (◯), beam search peak alignment with sideways shifting and interpolation (□), and peak alignment by genetic algorithm with sideways shifting and interpolation (+).

but only to determine the minimum segment widths and the maximum shifting allowed. The possible danger with peak picking is that some peaks of importance may be missed or that artefacts are detected as peaks.

One major advantage of PARS is the possibility to use autoscaling, which scales every variable (true peak) to unit variance. This will magnify the variances of small peaks, which can be of major interest, and assist in the finding of these peaks in the multivariate analysis. Autoscaling on bucketed or full spectra, peak-aligned or not, will decrease the distances between groups in the analysis since noisy buckets are put on the same scale as buckets containing information.

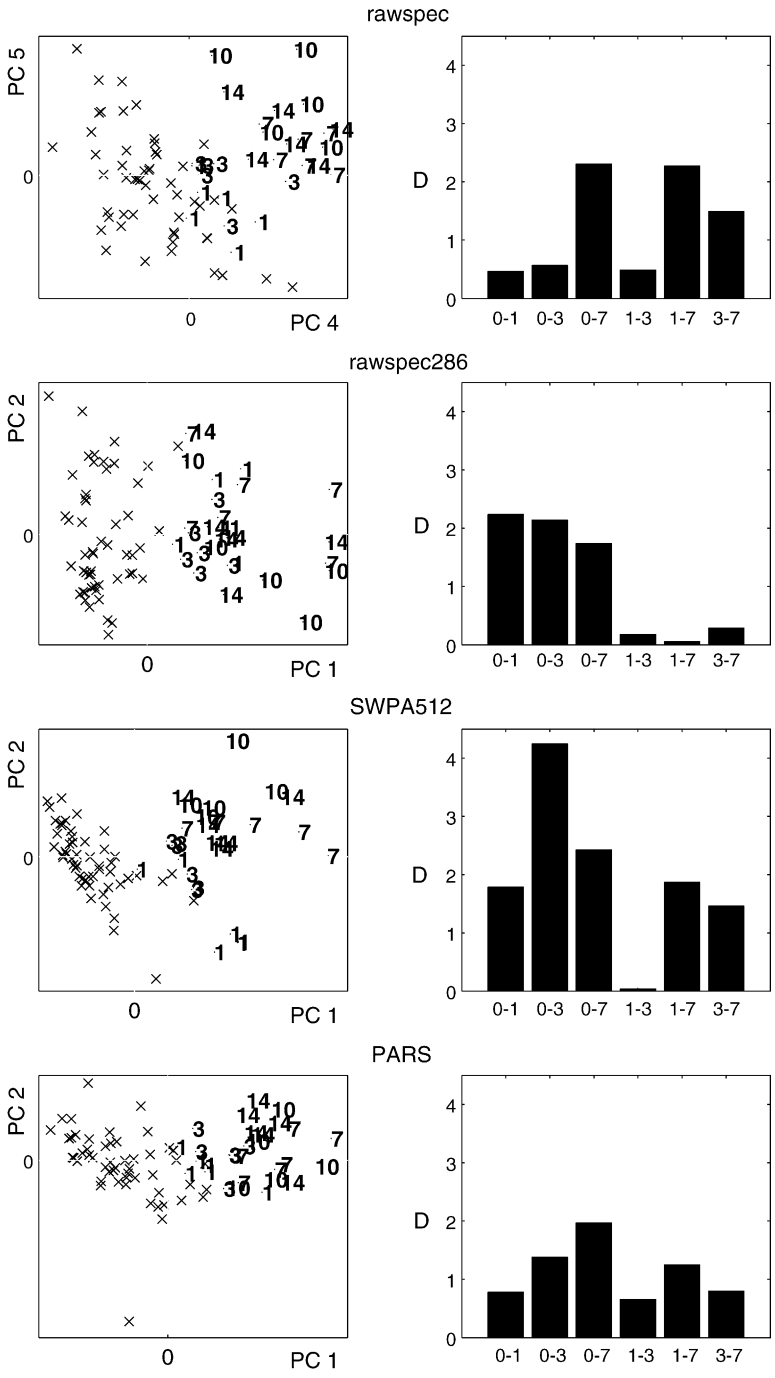

Fig. 5. The left column depicts the best PCA scores plot from each data set reflected by the class separation and right column show distances by means of the measure of class separation (MCS) between all the classes: 0–1: control to day 1; 0–3: control to day 3; 0–7: control to days 7–14; 1–3: day 1 to day 3; 1–7: day 1 to day 7–14; 3–7: day 3 to days 7–14; D depicts the MCS distance, rawspec: raw data, rawspec286: "classical bucketing" with 286 buckets, i.e. 0.04 ppm/bucket, SWPA512: segment-wise peak aligned data (beam search algorithm and shift correction alignment only) bucketed to 512 buckets, PARS: peak alignment using reduced set mapping.

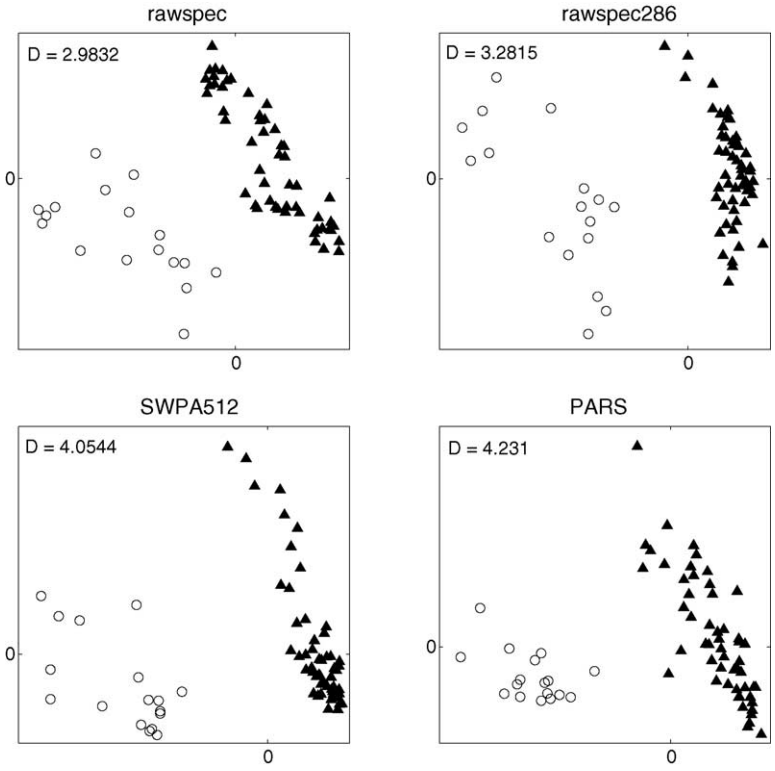

Fig. 6. Scores of the first two PLS components from PLS-DA of control and dosed groups from day 7 to day 14 are plotted, PLS component 1 on the *x*-axis and component 2 on the *y*-axis. D depicts the MCS distance, rawspec: raw data, rawspec286: "classical bucketing" with 286 buckets, i.e. 0.04 ppm/bucket, SWPA512: segment-wise peak aligned data (beam search algorithm and shift correction alignment only) bucketed to 512 buckets, PARS: peak alignment using reduced set mapping.

The plots in Fig. 5 show the possibility of separating all groups from each other: control, day 1, day 3 and days 7–14, after PCA analysis, using the MCS. These results will vary with the representation—shown is the best scores plot from the two PCs representing the best overall results ranked by the MCS. The bar plots to the right indicate an overall better separation between almost all classes when SWA is performed. The mean MCS distances are 1.26, 1.10, 1.97 and 1.14 for raw spectra, classical bucketing, SWA and PARS, respectively. The most important feature of these plots may be their interpretability. In the segment-wise peak aligned case, samples from day 1 are spreading in one direction separated from the directions obtained from the later days, a phenomenon which may contain valuable information on early formation of metabolites. Also worth noting is that bucketing and peak alignment reduce a lot of variation, not related to this analysis, which is explained by PC 1–3 in the raw data. Another important observation in these plots is that no information of interest seems to be destroyed or lost by the peak alignment. With the classical bucketing, the separation between dosed classes is indefinite, while SWA shows the same pattern as in raw data but with a better separation of the classes.

In Fig. 6 the results from the PLS-DA analysis are given. A good separation between the control group and dosed group for days 7–14 is shown. However, PARS produces the best results, followed by SWA, classical bucketing with 0.04 ppm per bucket, and raw data, in that order, ranked by the MCS.

Differing numbers of buckets and projections of PCA scores will show varying results as well as other distance measures or evaluations. Several score representations and different inter- and intra-group distance measures, such as different combinations of Mahalanobis and Euclidean distances have been tried out and the proposed MCS captures the information provided by the grouping of the scores. The score sets in this work have been chosen to show the best separation in each case and are considered to reflect the class distances fairly.

4. Conclusions

The two dedicated peak alignment methods examined in this work produce better results than classical bucketing or raw data considering the measure of class separation (MSC) of scores from PCA or PLS-DA. This is due to the removal of variations originating from instrumental instabilities, a background sample matrix and preservation of the variability from minor peaks. There are always risks of introducing errors when manipulating spectra; however, for the two proposed peak alignment methods, the advantages outweigh the disadvantages. Although it is hard to validate the alignment meth-

ods by numbers derived from the spectral domain, we can find no proof that the alignment methods destroy the latent information in the data.

No attempt to interpret the origin of the observed class separation has been made. The class differences can probably be accounted for by the NMR signals reflecting the excretion of the exogenous compound and its metabolites or by general changes in metabolism due to acute toxicity.

References

[1] M. Spraul, P. Neidig, U. Klauck, P. Kessler, E. Holmes, J.K. Nicholson, B.C. Sweatman, S.R. Salman, R.D. Farrant, E. Rahr, C.R. Beddell, J.C. Lindon, J. Pharm. Biomed. Anal. 12 (1994) 1215–1225.
[2] E. Holmes, J.K. Nicholson, A.W. Nicholls, J.C. Lindon, S.C. Connor, S. Polley, J. Connelly, Chemom. Intell. Lab. Syst. 44 (1998) 245–255.
[3] B.C. Potts, A.J. Deese, G.J. Stevens, M.D. Reily, D.G. Robertson, J. Theiss, J. Pharm. Biomed. Anal. 26 (2001) 463–476.
[4] B.M. Beckwith-Hall, J.T. Brindle, R.H. Barton, M. Coen, E. Holmes, J.K. Nicholson, H. Antti, Analyst 127 (2002) 1283–1288.
[5] M. Defernez, I.J. Colquhoun, Phytochemistry 62 (2003) 1009–1017.
[6] J. Forshed, I. Schuppe-Koistinen, S.P. Jacobsson, Anal. Chim. Acta 487 (2003) 189–199.
[7] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (1985) 491–500.
[8] S. Wold, H. Antti, F. Lindgren, J. Ohman, Chemom. Intell. Lab. Syst. 44 (1998) 175–185.
[9] R. Siuda, G. Balcerowska, D. Aberdam, Chemom. Intell. Lab. Syst. 40 (1998) 193–201.
[10] T.R. Brown, R. Stoyanova, J. Magn. Reson. Ser. B 112 (1996) 32–43.
[11] T. Brekke, O.M. Kvalheim, E. Sletten, Anal. Chim. Acta 223 (1989) 123–134.
[12] J. Jaumot, M. Vives, R. Gargallo, R. Tauler, Anal. Chim. Acta 490 (2003) 253–264.
[13] J. Forshed, F.O. Andersson, S.P. Jacobsson, J. Pharm. Biomed. Anal. 29 (2002) 495–505.
[14] R. Stoyanova, A.W. Nicholls, J.K. Nicholson, J.C. Lindon, T.R. Brown, J. Magn. Reson. 170 (2004) 329–335.
[15] J.T.W.E. Vogels, A.C. Tas, J. Venekamp, J. Van Der Greef, J. Chemom. 10 (1996) 425–438.
[16] H. Witjes, W.J. Melssen, H. Zandt, M. van der Graaf, A. Heerschap, L.M.C. Buydens, J. Magn. Reson. 144 (2000) 35–44.
[17] R.J.O. Torgrip, M. Åberg, B. Karlberg, S.P. Jacobsson, J. Chemom. 17 (2003) 573–582.
[18] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[19] J.K. Nicholson, I.D. Wilson, Prog. Nucl. Magn. Reson. Spectrosc. 21 (1989) 449–501.
[20] G.-C. Lee, D.L. Woodruff, Anal. Chim. Acta 513 (2004) 413–416.
[21] J.H. Holland, Adaption in Natural and Artificial Systems, 1st ed., The University of Michigan, 1975.
[22] R. Wehrens, L.M.C. Buydens, Trends Anal. Chem. 17 (1998) 193–203.
[23] C. Houck, J. Joines, M. Kay, A Genetic Algorithm for Function Optimization: A Matlab Implementation, NCSU-IE TR 95-09, 1995.
[24] R. Bisiani, in: S.C. Shapiro (Ed.), Encyclopedia of Artificial Intelligence, Wiley, New York, 1992, pp. 1467–1468.
[25] MATLAB 6.5, The MathWorks Inc., Natick, MA, 2002.
[26] E.C. Harrington Jr., Ind. Quality Contr. 21 (1965) 494–498.
[27] F.O. Andersson, R. Kaiser, S.P. Jacobsson, J. Pharm. Biomed. Anal. 34 (2004) 531–541.
[28] W.J. Dunn III, S. Wold, J. Med. Chem. 21 (1978) 1001–1007.
[29] M. Barker, W. Rayens, J. Chemom. 17 (2003) 166–173.
[30] M. Åberg, Variance Reduction in Analytical Chemistry, Thesis, Stockholm University, Dept. of Analytical Chemistry, 2004.