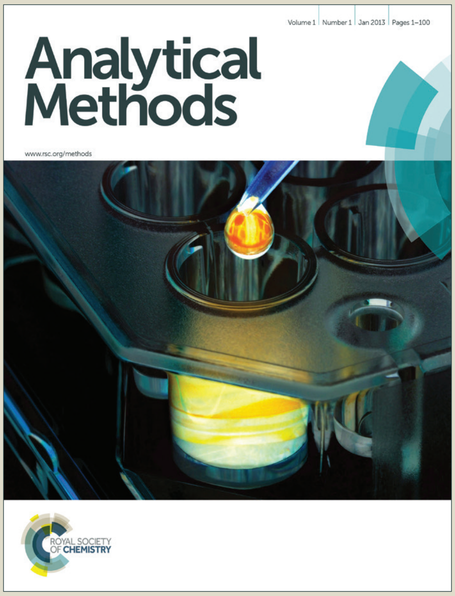
Analytical Methods

Accepted Manuscript


Volume 1 | Number 1 | Jan 2013 | Pages 1–100
Analytical Methods
www.rsc.org/methods
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/methods

Analytical Methods RSCPublishing

ARTICLE

Peak Alignment of One-Dimensional NMR Spectra by Means of an Intensity Fluctuation Frequency Difference (IFFD) Segment-wise Algorithm

Cite this: DOI: 10.1039/x0xx00000x

Received 25th April 2015, Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

K. Wang,[a,b] G. A. Barding[c] and C. K. Larive[a]

The increasing scientific and industrial interest in metabolomics often takes advantage of the high level of qualitative and quantitative information provided by nuclear magnetic resonance (NMR) spectroscopy. However, several chemical and physical factors can affect the frequency of an NMR resonance. Especially in complex biological samples such as biofluids, small perturbations in NMR chemical shifts can complicate the recovery of biomarker information in metabolomics studies using multivariate statistics and pattern recognition tools. Novel segment-wise peak alignment algorithms have been proposed in the literature to correct the misalignment of NMR signals. The approach presented here, the Intensity Fluctuation Frequency Difference (IFFD) algorithm, is a highly efficient method designed to reduce variability in peak positions across the multiple NMR spectra used in metabolomics studies. This automated method refines segmentation using differences in the frequencies of the intensity fluctuations of signals and baseline noise to improve spectral alignment. Alignment is performed sequentially using an open source program, icoshift, employing a fast Fourier transform (FFT) engine to align all spectra simultaneously. The IFFD-icoshift method is illustrated for ^{1}H NMR spectra measured for 50 human urine samples collected from healthy volunteers: 41 samples, including urine from a pregnant female, were collected randomly following a normal dietary routine and 9 samples were collected after dietary supplementation with ibuprofen, alcoholic beverages or an energy drink. We demonstrate the superior performance of IFFD-icoshift alignment over a wide range of peaks and its capacity to enhance the interpretability and robustness of multivariate statistical analysis. This approach is widely applicable for NMR-based metabolic studies and is potentially suitable for many other types of data sets such as chromatographic profiles and MS data.

Introduction

The metabolic analysis of complex biosamples, such as tissues and body fluids, through various analytical methods is a growing area of interest due to the desire to correlate alterations in metabolism with disease, toxicology, and substance abuse.[1-4] Urine and blood plasma or serum are particularly attractive for metabolomics studies because they can be obtained in relatively high volumes through minimally or non-invasive means.[5,6] Urine has the additional advantages of being molecularly stable and provides a relatively simple matrix with minimal levels of protein in most patients.[7,8] As a result, urine is widely used in clinical analyses and has been employed for metabolite profiling of maple syrup urine syndrome,[5] propionate metabolism disorders,[9] toxicological profiling studies,[10-11] inflammatory bowel disease,[12] and biomarker discovery of human disease.[13-15]

Metabolomics (or metabonomics) is the untargeted comprehensive metabolic analysis of an organism.[16] Through experimental design, metabolomics studies compare control and stressed populations, where the stressor (or treatment) may be a disease state, genetic mutation, toxicological insult, or other parameter under the control of the investigator.[16,17] Several analytical techniques have been used in metabolomics studies, with proton (^{1}H) nuclear magnetic resonance spectroscopy (NMR) one of the most widely applied.[18,19] NMR is inherently quantitative, amenable to a variety of sample formats (including intact tissue), and is usually performed without an accompanying separation step.[20] These advantages make NMR a powerful method for metabolic studies of biofluids and it has been applied as such for several decades.[21-24]

Although NMR-based metabolomics is generally highly reproducible and robust, the recovery of biological information is complicated by instrumental imperfections, noise and line shape variations, together with phase and baseline distortions, all of which contribute to quantitative errors.[25-28] Thus, NMR data sets typically require pre-processing prior to statistical modelling. For example, spectroscopic noise is usually reduced by apodization of the free induction decay (FID) or through more sophisticated wavelet-based approaches,[29,30] and manual or automatic baseline/phase correction is generally required.[31-32]

Major challenges in the analysis of biological data sets can arise due to natural variations in sample dilution, especially urine, and because of differences in resonance frequency across spectra, sometimes called "positional noise".[33-38] These effects can impact the interpretation of metabolomics data sets using statistical techniques, e.g., statistical correlation spectroscopy (STOCSY),[39-41] principal component analysis (PCA),[42] and partial least squares (PLS).[43,44] For example, treatment groups can sometimes be discriminated on the basis of irrelevant variations in peak positions rather than changes in sample composition characterizing the metabolic states of an organism.[39] Such difficulties are often observed in ^{1}H NMR spectroscopic studies of urine samples, which can vary in pH, ionic strength, metal ion concentrations, and osmolality, all of which can affect chemical shifts. Correcting variations in peak position is therefore essential for improved spectral information recovery.[24,25,45]

Different approaches have been explored to solve the alignment problem and have provided appropriate solutions for many different experimental cases. The historical basis of this approach relies on a very simple and widely used binning or bucketing method. Binning involves data reduction performed through integration of NMR resonances within standardized spectral regions (bins or buckets) whose width commonly ranges between 0.01 and 0.05 ppm.[46] The major drawbacks of this approach are the loss of spectral resolution and splitting of peaks across adjacent regions, which is especially problematic when chemical shifts vary dramatically with experimental conditions. When high resolution is required, other more sophisticated alignment methods may be considered such as dynamic time warping (DTW) or correlation optimized warping (COW),[47-49] which have been demonstrated to be effective on chromatographic data and also have been used to solve simple NMR alignment issues with satisfactory results.[50] Apart from being computationally intensive, the main problem of these approaches is that alignment is obtained by local stretching or compression. This is not really suitable for NMR spectral data because this correction model works best when there is a positive correlation between peak width and position.

Alternatively, less computationally intensive alignment methods identify the relevant peaks present in an NMR spectrum and convert the spectrum to a list of peaks and relative attributes, thereby dramatically decreasing the dimensionality of the data. Such methods have been primarily developed and introduced by Torgrip and colleagues.[38,51,52] The major drawbacks of this approach are the elimination of the information carried by the fine structure of the signal shape and the need to define meta-parameters for the peak-picking procedure and subsequent alignment.

To make full-resolution alignment of large datasets feasible and avoid data reduction steps such as binning, interval based algorithms represent a key step forward toward a definitive solution to the alignment problem. In practice the chemical shift of each NMR signal depends on several factors and, especially for those signals strongly affected by solution conditions, can independently shift to higher or lower frequencies across a series of spectra. Peaks that are adjacent or even overlapped in one spectrum can be baseline separated in other spectra and, in extreme cases, their relative position may be inverted. Because of the complexity of NMR spectra, it is preferable to reduce the global alignment problem to alignment over a series of smaller and specific spectral intervals. In this way, resonance shifts occurring in opposite directions can be more readily corrected and eventually a global, full resolution aligned NMR data set can be reconstructed.

Savorani et al. addressed the problem of segmental spectral alignment with the open source algorithm icoshift. Icoshift is a peak alignment algorithm based on correlation shifting of spectral intervals to simultaneously align all the spectra in a data set.[53] By combining a rapid optimized FFT engine, icoshift reduces the calculation times for large metabonomic data sets from hours (e.g., with COW), or minutes to seconds.[54] However, in some datasets, the peak positions of resonances shift significantly between the samples. The effect of strongly shifted spectra also challenges the methods based on spectral binning, because peaks could mistakenly be assigned to the wrong bins.

The pivotal operation for automatic segment-wise peak alignment is distinguishing regions of baseline from the regions containing signals, and many approaches can be found in the NMR literature.[55-57] A common approach is to use numeric derivative or Continuous Wavelet transform (CWT) to improve the signal-to-noise ratio (SNR), usually combined with noise reduction routines such as Savitzky-Golay smoothing.[58] Subsequently, threshold values are used to recognize signals.[25] However, baseline distortions occur frequently in NMR spectra, especially in bio-fluid samples, and baseline recognition routines can have difficulty with weak signals. In extreme cases peaks may not be recognized because the baseline distortion may be greater than the peak intensities, even after treatment by CWT.

The new approach presented in this study, Intensity Fluctuation Frequency Difference (IFFD), is designed to automatically produce intervals for spectral alignment using icoshift. The algorithm discriminates between baseline and signal based on the frequency of intensity fluctuations and uses this information to define spectral segments such that regions containing peaks are separated by regions of baseline. The peaks within the intervals defined by the algorithm are then aligned by the icoshift program. The IFFD-icoshift method is evaluated for the alignment of both major and minor peaks in NMR spectra of human urine. The performance of IFFD-icoshift is compared with other widely used alignment methods, particularly the use of icoshift alone. The success of alignment is also evaluated using PCA through increased interpretability and improved information recovery.

Experimental

Materials and reagents

Sodium-3-trimethylsilyl-propanesulfonic acid-d_6 (DSS) was purchased from Isotec (St. Louis, MO). Deuterium oxide was obtained from Cambridge Isotope Laboratories, Inc. (Andover, MA). Mono- and dibasic sodium phosphate and sterile specimen cups (90 mL) were purchased from Fisher Scientific (Pittsburgh, PA) and sodium azide was obtained from Sigma-Aldrich (St. Louis, MO).

Sample collection and preparation

Human urine samples were collected in accordance with UC Riverside's Office of Research Review Board (HS 12-086). The health of the volunteers was based on their own admission. Volunteers were requested to report the use of pharmaceuticals (i.e., pain killers) and alcohol throughout the course of the study and test samples were so indicated by the subjects. To minimize pH-induced chemical shift changes in the ^{1}H NMR spectra of urine samples, a pH 7.3 phosphate buffer was prepared using 200 mM Na_2HPO_4 and 36 mM NaH_2PO_4 in 25% D_2O with DSS and sodium azide added to achieve a final concentration of 1 mM.[7] A 670 µL aliquot of urine was transferred immediately after collection to a 1.5 mL microcentrifuge tube along with 330 µL of buffer and stored at -80 °C until analysis.

NMR experimental parameters

^{1}H NMR spectra were measured using a Bruker Avance NMR spectrometer operating at 599.69 MHz and equipped with a 5-mm inverse broadband probe with xyz gradients. The magnetic field homogeneity was optimized using up to 28 shims and the probe was manually tuned and matched. Solvent suppression was accomplished by excitation sculpting with the transmitter set on the water resonance.[59] Spectra were acquired without spinning at 298 K using digital quadrature detection. The line-width of DSS for all spectra was less than 1 Hz prior to apodization. Free induction decays were collected into 32768 points using a spectral width of 11.67 ppm. Spectra were measured using an 11.0 µs 90° pulse and 16 dummy scans preceded the coaddition of 256 transients with a relaxation delay of 1.5 s for an experiment time of 17.7 min.

NMR data processing

Spectra were processed using ACDlabs Spectrus Processor (Advanced Chemistry Development, Inc., Toronto, Canada) and Matlab (2013b, The Mathworks Inc., Natick, MA). The resonances of the residual water signal, urea, and spectral regions below 0.5 ppm and above 9.5 ppm were set as dark regions and excluded from the analysis. Processed spectra were exported as *.txt files containing peak intensity with the corresponding chemical shift.

IFFD

The IFFD program, described in greater detail in the Results and Discussion section, was constructed in Matlab and utilizes the icoshift toolbox available on the web.[53] The IFFD package is freely available by request to the authors. The IFFD and PCA analyses were conducted blindly to ensure an unbiased evaluation, and data labels were added only after the PCA analysis was complete. For these experiments, samples from volunteers of both genders were included together in the control group.

PCA

Principal Component Analysis (PCA) has been frequently applied to assess alignment quality.[36,38] Application of PCA to unscaled data sets introduces a bias toward the variation of the most intense peaks because of their larger contribution to total variance.[60,61] To overcome this limitation, data are typically scaled using a method such as Pareto or unit variance scaling, which is also known as autoscaling to achieve equal impact of variables in a PC model.[61,62]

PC scores summarize inter-relationships between observations.[60] A subset of metabolites correlated with PC scores can be identified using the corresponding factor loadings. In the case of an unscaled data set, loadings can be directly used for interpretation. If PC scores explain the variation in sample chemical composition, the line shapes of PC loadings will resemble metabolite peaks in the NMR spectrum.

However, the visualization of loadings is difficult with unit-variance scaled models because the line shapes of loadings are not interpretable in the same way as in unscaled models and so cannot be used directly to identify peaks dominating the variance in the data. To correct this deficit, a loading value can be multiplied by the standard deviation of its original spectral variable.[54]

Results and discussion

Though several approaches have been introduced to address peak alignment in large NMR spectral datasets generated for metabolomics studies, the current algorithms suffer from various practical limitations. Key to the success of segment-wise peak alignment methods is the appropriate definition of the frequency intervals over which alignment is to be performed. A popular algorithm, icoshift, can be used in a variety of modes (i.e. manual, automated) to align NMR peaks across a series of spectra, and the icoshift program is incorporated into our IFFD workflow as described below. In practical terms, using the icoshift regular intervals mode, it is possible for resonances to be separated into different segments, which can produce an incorrect alignment, illustrated in Fig. 1. For proper alignment, the interval boundaries must not be set in regions containing peaks. Though the user can define the interval boundaries in the icoshift program, it can be difficult to select the interval boundaries manually across a large series of spectra because baseline region may be hard to define by visual interpretation. As illustrated by the overlaid spectra highlighted in Fig. 1A, even though the three resonances in each spectrum are well-resolved it can be difficult to identify regions of baseline when considering the full set of spectra. Therefore, a segment-wise method is necessary to provide accurate interval boundaries appropriate for each metabolomics data set.

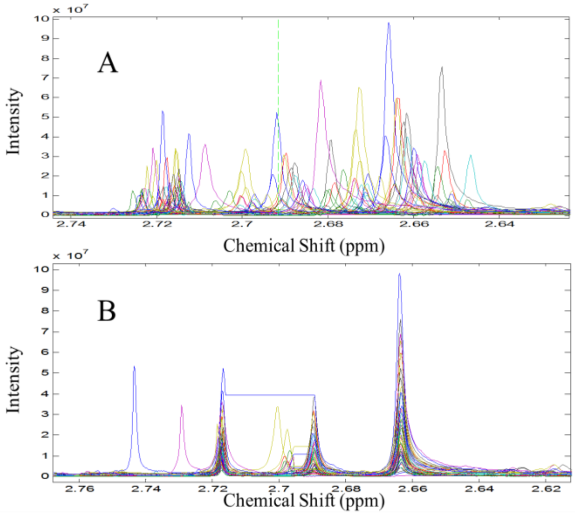


Fig. 1 (A) Expansion of the 600 MHz NMR spectra for 50 urine samples without alignment. The green dashed line illustrates a boundary point using regular intervals (the number of intervals is 50, which means the whole spectrum was divided into 50 segments in equal length) in icoshift. (B) Peak alignment using regular intervals in icoshift. In (B) incorrect alignment resulted from the separation of resonances into different segments.

IFFD aims to align peak positions within segments defined by the frequency of intensity fluctuations. In essence this algorithm efficiently identifies regions of baseline by distinguishing signals from noise based on their intensity fluctuation frequency, an approach that is robust with respect to baseline distortions. In IFFD, the fluctuation frequency is defined as the distance (i.e. number of data points) between adjacent crests. The spacing, ΔHz, between data points is regular and is determined by the spectral width and the number of points acquired in the time domain free induction decay (FID). Crests are identified by comparing the intensity of a given data point with the intensity of the adjacent data points at both higher and lower frequencies. If a data point has an intensity greater than the neighbouring points, it is defined as a crest.

Because NMR spectral noise is dominated by Johnson noise, baseline intensity fluctuations are mostly derived from noise. The interpoint width of baseline noise crests should be significantly smaller than those of relevant signal peaks due to the higher frequency content of the Johnson noise-limited NMR baseline fluctuations. As a result, the frequency of intensity fluctuations in baseline regions is similar and defined by a small number of data points (low ΔHz) between adjacent crests. Peaks, which include both noise and signals, can be identified by their larger intensity fluctuation frequency determined by the distance between the crest of the peak and the adjacent baseline crests (a large number of data points, high ΔHz). Measurement of the fluctuation frequency allows the regions containing baseline and peaks to be discriminated, as shown in Fig. 2. Importantly, this method is effective even when the peak has a low relative intensity.

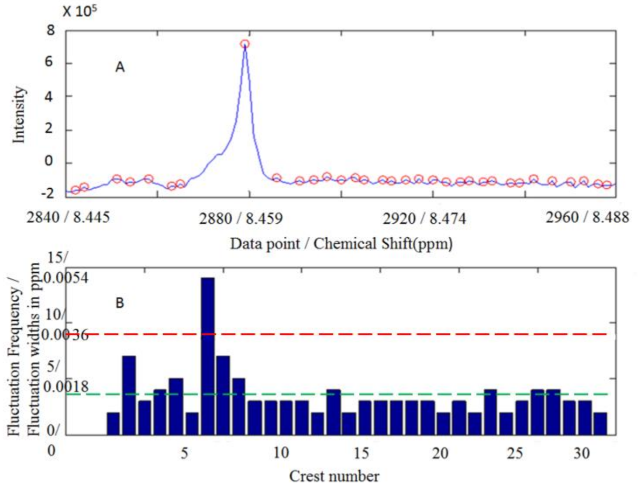


Fig. 2 (A) Expansion of a ^{1}H NMR spectrum containing baseline and a resonance of low relative intensity. The red circles in (A) designate the frequency crests. (B) Histogram showing the distances (i.e. number of data points/ppm) between crests in (A). Because the distance between adjacent crests is greatest for the peak, the value of the fluctuation frequency (in data points/ppm) is larger for this signal than for the corresponding baseline regions. The Min-threshold value, set here as the mean frequency fluctuation over the segment, is indicated by the green dashed line, and the Max-threshold value, which is twice of the mean frequency fluctuation over the segment, is indicated by the red dashed line.

For the 600 MHz NMR spectrum in Fig. 2 with a digital resolution of approximately 0.00036 ppm/point, The peak widths of noise fluctuations averaged 0.0015 ppm (approximately 4.5 data points) while those of the relevant signal is 0.005 ppm (approximately 15 data points). The value of the fluctuation frequency is greater for the NMR signal than for the baseline noise and can be used as a criterion to automatically recognize spectral regions containing baseline and peaks. In this example, two threshold values, once (Min-threshold) and twice (Max-threshold) the mean fluctuation frequency for the segment, were used to distinguish between baseline noise and a resonance of low relative intensity (Fig. 2A). In this application, the goal of discriminating between baseline and peaks is to obtain the appropriate baseline points to using in dividing the spectrum into segments, those points less than the Min-threshold value are identified as baseline points, meanwhile, those points greater than the Max-threshold value are identified as signal points.

In data sets containing spectra from many samples, the regions containing baseline and peaks may be different for each spectrum, and some spectra may contain unique peaks that are not found in other spectra. Therefore as a reference we calculate a maximum spectrum, $d_{\max}$, having the maximum number of features

$$d_{\max} = \max(d) \qquad (1)$$

where d is an $n \times m$ matrix, n is the number of samples and m is the number of real spectral data points. The term max(d) returns a row vector containing the maximum element from each column.

Subsequently, an optimization routine is applied to select the segments for icoshift peak alignment. The optimization routine selects the segments based on three principles: (1) each segment should contain at least one peak, (2) the segment should maximize the distance on either side of the peak(s), and (3) the segment boundaries should be at a local minimum with respect to baseline intensity. After these optimizations, the interval points, must in the absolute baseline points, are obtained.

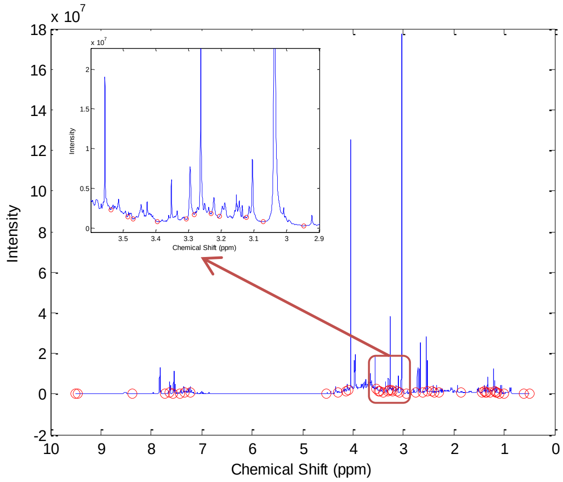


Fig. 3 Interval points (red circles) obtained by IFFD. The expansion from 2.9 - 3.6 ppm shows the segmentation of this region.

As shown in Fig.3, IFFD efficiently avoids peaks in defining interval points. Even in regions where the baseline is not flat, as illustrated the expansion from 3.6-2.9 ppm, IFFD is able to discriminate between peaks and baseline.

Once the intervals are defined using IFFD, the segment boundaries are used as an icoshift parameter to define the regions for peak alignment. As icoshift makes use of relative peak positions, the method generally avoids aligning two unique peaks in different spectra that have very close, but distinct chemical shifts. Following an initial peak alignment, the process can be iterated and provides a built-in assessment of success: if the number of peaks in $d_{\max}$ of the aligned spectra is the same as the maximum number of peaks in all spectra, a successful alignment has been achieved.

The IFFD method avoids the separation of the same peak into different segments across a series of spectra, is computationally efficient, semi-automated, and requires fewer user-defined parameters compared with alternative approaches.

Application of IFFD to compensate for chemical shift variation in urine ^{1}H NMR spectra

A dataset of 50 ^{1}H NMR spectra of human urine samples were analysed using IFFD. Dietary effects often result in subtle changes in metabolic profiles and even small variations in peak positions can significantly limit the recovery of information from a collection of spectra. The dataset analysed in this work included 41 human urine samples collected randomly by healthy volunteers following a normal dietary routine, 8 test samples collected after dietary supplementation with ibuprofen, alcoholic beverages or an energy drink, and a test sample prepared by 50% dilution of a control sample. This data set was previously analysed using an automated metabolic profiling analysis method Visual Interpretation of Z-Score Ratios (VIZR).[63] Though PCA was unable to properly segregate the test and control spectra in our prior study, VIZR analysis successfully discriminated each of the test samples from the control and identified the NMR spectral regions responsible for the disparity between the individual test samples and the control group. The self-normalizing nature of the VIZR calculation proved to be independent of dilution effects, especially important in urine analyses.[63]

As illustrated in the insert in Fig. 1A, in the spectral region 2.54 - 2.76 ppm the extent of peak position variance across the series of 50 NMR spectra makes it difficult to automatically identify baseline regions to use in the definition of segments for peak alignment. To avoid mistakenly assigning peaks to the wrong bins (as in Fig. 1B), the process of alignment was separated into two steps. In the first step we separated the spectra into segments using the principles that one or more peaks should populate each interval and that the frequency distance between intensity fluctuations will be greater for peaks than for baseline noise. The interval boundaries and frequency segments defined using IFFD are shown in Fig. 3, with the expansion again highlighting the region 2.54 - 2.76 ppm.

As shown in Fig. 4, the segments initially defined by IFFD are of variable width. Some segments (e.g., 2.54 - 2.61 ppm) are easily separated from adjacent segments and contain a single peak. Other segments (e.g., 2.61 - 2.76 ppm) contain multiple peaks. Once these initial segments were determined, icoshift was used to align the peaks in each segment. The results are shown in Fig. 5 for the segment from 2.61 - 2.76 ppm. Fig. 5A shows the stacked plot of all 50 spectra demonstrating that each spectrum contains only 3 peaks in this region.

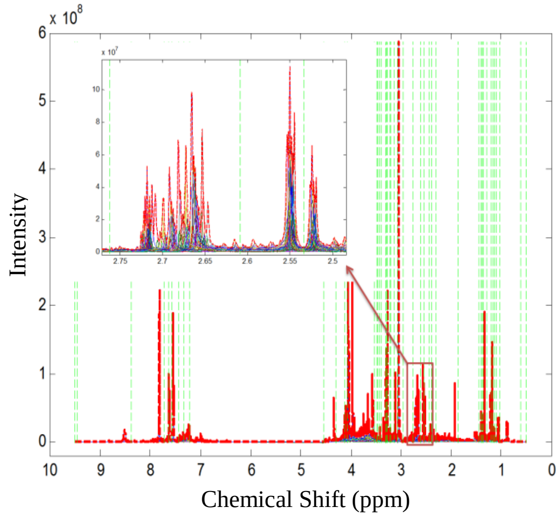


Fig. 4 Interval boundaries (green dashed lines) determined using IFFD for segment wise peak alignment. The expansion from 2.54-2.76 ppm shows the segmentation of this region. The overall maximum spectrum, d_{max}, is indicated by the red dotted line.

Fig. 5B shows the overlay of all 50 spectra prior to alignment. As it is difficult to define the baseline in this region, the segment defined in Fig. 5B was initially aligned by icoshift as a ‘whole’ segment. Fig. 5C shows the result after the first alignment. The right two peaks were aligned to the same position, but the downfield peak was not initially aligned by icoshift. Even though the alignment of the peaks in this segment was incomplete, new baseline regions are produced that can be used in a subsequent step to separate create new sub-segments, as shown in Fig. 5C.

IFFD was performed again with new interval boundaries obtained. Following icoshift analysis for the new segments, all of the peaks were properly aligned (Fig. 5D). The steps in which segments are further refined into sub-segments using IFFD and then aligned using icoshift can be performed repeatedly until the number of peaks in the maximum spectrum, d_{max}, (shown by the red dotted line) is equal to the number of peaks produced by the sum of all spectra. In the final step, all peak alignment segments are combined to recreate each complete spectrum, shown in Fig. 6. The expansion of the region from 2.54 – 2.76 ppm in Fig. 6 shows the results of peak alignment in greater detail.

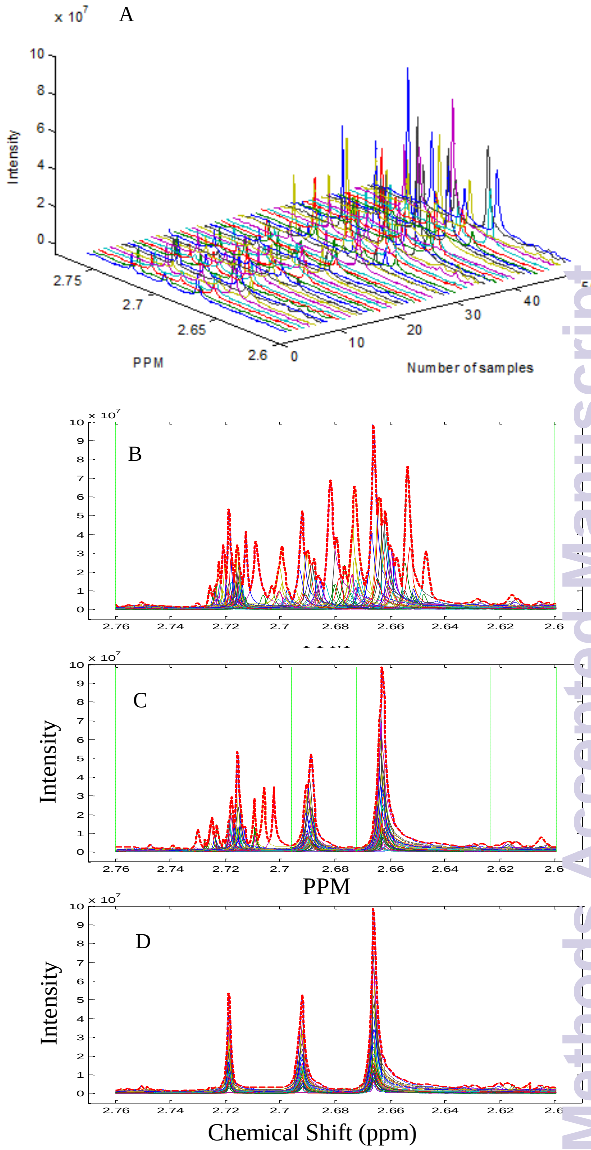


Fig. 5 (A) Stacked plot showing the region from 2.61 to 2.76 ppm for all 50 spectra. Each spectrum contains only 3 peaks but their frequencies are highly variable. (B) Overlay of all 50 spectra in the region 2.61 to 2.76 ppm prior to alignment. The maximum spectrum defined by d_{max} (red dotted line), contains many more than three peaks in this region due to the variability in peak position. (C) Results of icoshift alignment for the initial segment (2.61-2.76 ppm) and the new interval boundaries (green dashed lines) produced by IFFD. (D) Application of icoshift to the segments shown in (C) produces complete peak alignment.

Evaluation of the variation in sample composition using PCA

To evaluate the effectiveness of the alignment methods applied herein, we used PCA to recover the latent components associated with metabolic differences. The data set was subjected to unit-variance scaling prior to PCA. Unit-variance scaling normalizes the data using the standard deviation as the scaling factor, thereby reducing the impact of the most intense resonances. In Fig. 7A we examine the effect of unit-variance scaling on the PCA results without peak alignment.

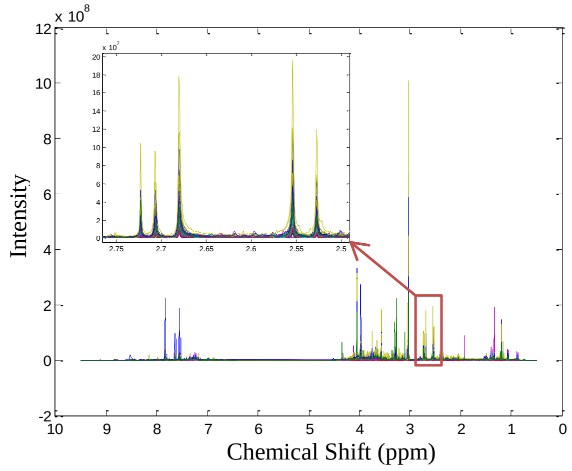


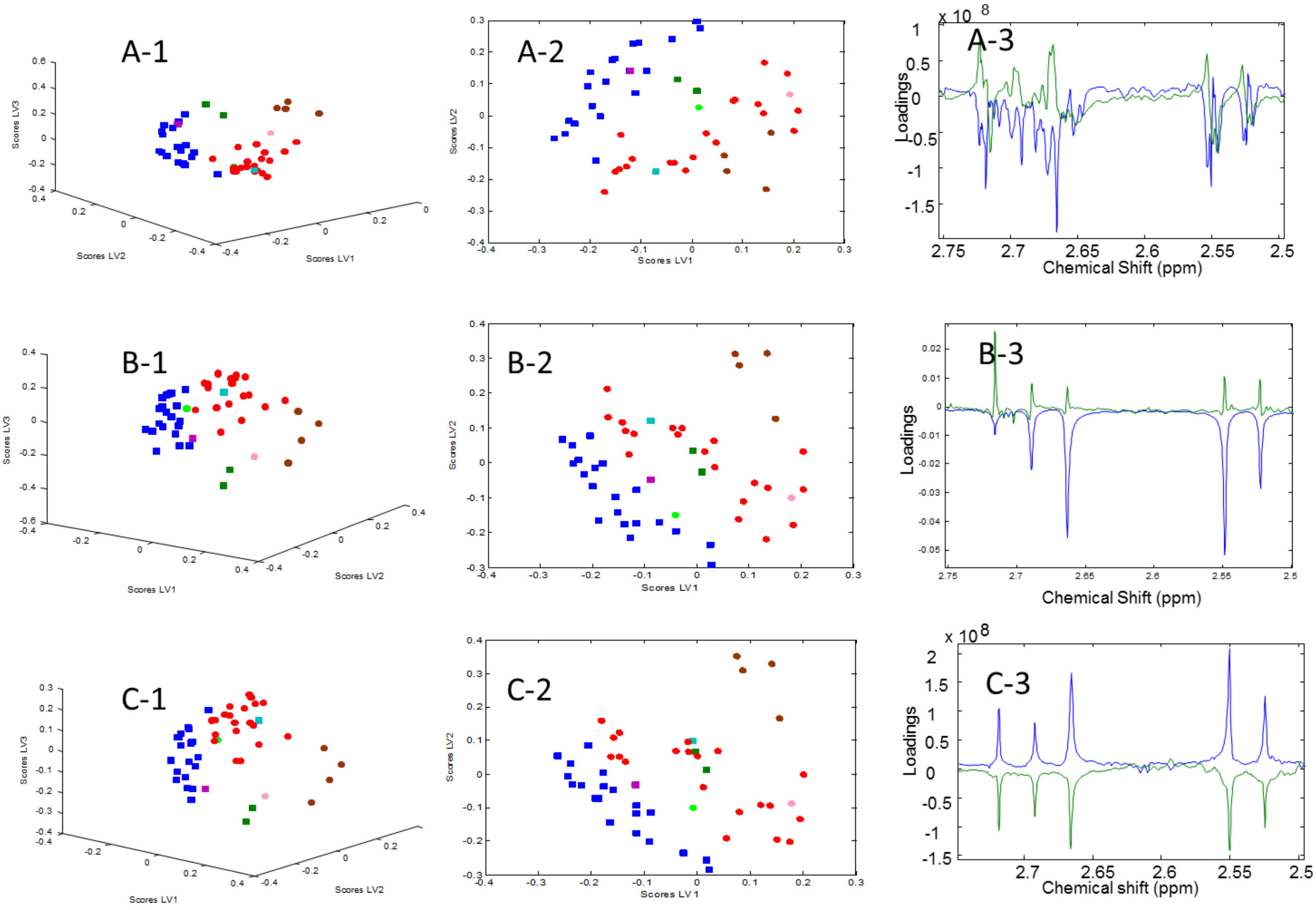
Fig. 6 Overlay of the spectra reconstituted after segment-wise alignment. The expansion shows the effect of alignment on the spectral region between 2.54 and 2.76 ppm.

Fig. 7A-1,2 shows the PC scores of the unit-variance scaled data without peak alignment. The corresponding loadings plots for PC1 and PC2 are comprised almost entirely of first-derivative-like line shapes, as illustrated by the region 2.54 – 2.76 ppm (Fig. 7A-3). The derivative-like line shape in the loadings plot indicates that variable peak positions contribute substantially to the PC scores. As a result, the principal components reflect the variation caused by variable peak position as well as sample composition.

In Fig. 7B, the result of unit-variance scaling and icoshift peak alignment by 50 equal segments was displayed. After icoshift peak alignment, a better separation of treatment and control samples was obtained (Fig. 7B-1,2). Though the line shape of the loadings plots (Fig. 7B-3) are improved compared with those in Fig. 7A, the distortions indicate that variable peak positions still contribute to the PC scores.

Finally, Fig. 7C illustrates the benefit derived from the combination of chemical shift alignment by IFFD-icoshift. As a result of the improved peak alignment, the PC loadings (Fig. 7C-3) are not distorted and exhibit the line shapes expected for an NMR spectrum. The segregation of treatment and control samples is improved in Fig. 7C-1,2, The control samples are largely separated by gender, with the pregnant female sample(emerald circle) lying well within the group of female control samples.[20] Though the goal of this work was to use

Fig. 7 The PCA scores and loadings results for unit-variance scaled data without peak alignment (A), after icoshift peak alignment (B),



and following peak alignment with IFFD-icoshift (C). In the scores plots (A,B,C-1,2) the control samples are shown as dark blue squares (male) and red circles (female), with one sample from a pregnant female indicated by an emerald circle. The dietary supplementation samples are identified by green squares (ibuprofen, male), brown circles (ibuprofen, female), purple square (alcoholic beverages, male), and a light blue square (AMP energy drink, male). In the loadings plots shown in A,B,C-3 the blue line indicates the loadings for PC1 while the green line shows the PC2 loadings.

IFFD-icoshift to produce aligned NMR spectra as an input for PCA, Cloarec et al. have demonstrated that such positional variation of peaks can provide additional biochemical information through chemometric models incorporating back-scaled loadings.[64]

Conclusions

The combination of IFFD and icoshift provides a fast and accurate spectral alignment using full spectral information. The alignment quality was evaluated by the application of unit-variance PCA. IFFD-icoshift alignment produced considerable improvement in the lineshape of the PC loadings, removing the contributions of variable peak position to the PC scores.

The major improvement provided by IFFD-icoshift is that peaks of low relative intensity are easily identified and aligned, it avoids the separation of peaks across multiple segments, and needs fewer user-defined parameters, providing more accurate alignment. IFFD-icoshift is computationally fast and therefore can be used to align large data sets including hundreds or thousands of sample profiles. The method is widely applicable for spectral alignment in NMR-based studies involving complex mixtures, including metabolite profiling, natural product identification and food analysis and may be suitable for application to many other types of data sets such as chromatographic profiles.

Acknowledgements

This research was supported by State Scholarship Fund of China Scholarship Council (CSC) (No. 201306565052), Shaanxi National Science Foundation (No. 2014JM2045), the Fundamental Research Funds for the Central Universities (No. 0009-2014G1311085) and the Fundamental Research Funds for the Central Universities (No. 0009-2014G1121099) to K.W. Support by the National Science Foundation (CHE-1213845 and IOS-1121626) and the U.S. Department of Agriculture, National Institute of Food and Agriculture - Agriculture and Food Research Initiative (2011−04015) to C.K.L. is gratefully acknowledged.

Notes and references

[a] Department of Chemistry, University of California – Riverside, Riverside, CA 92521.

[b] College of Science, Chang'an University, Xi'an, 710064, China.

[c] Department of Chemistry and Biochemistry, California State Polytechnic University, Pomona, CA 91768.

1 C. Napoli, N. Sperandio, R.T. Lawlor, A. Scarpa, H. Molinari and M. Assfalg, Urine metabolic signature of pancreatic ductal adenocarcinoma by (1)h nuclear magnetic resonance: identification, mapping, and evolution, *J. Proteome Res.*, 2012, 11(2), 1274-1283.

2 A. Zhang, H. Sun and X. Wang, Serum metabolomics as a novel diagnostic approach for disease: A systematic review, *Anal. Bioanal. Chem.*, 2012, 404(4), 1239-1245.

3 J. Meng, X. Zhang, H. Wu, J. Bu, C. Shi, C. Deng and Y. Mao, Morphine-induced conditioned place preference in mice: Metabolomic profiling of brain tissue to find "molecular switch" of drug abuse by gas chromatography/mass spectrometry, *Anal. Chem. Acta.*, 2012, 710, 125-130.

4 M. Aimetti, S. Cacciatore, A. Graziano and L. Tenori, Metabonomic analysis of saliva reveals generalized chronic periodontitis signature, *Metabolomics*, 2012, 8(3), 465-474.

5 E. Holmes, P.J. Foxall, M. Spraul, R.D. Farrant, J.K. Nicholson and J.C. Lindon, 750 MHz ^{1}H NMR spectroscopy characterisation of the complex metabolic pattern of urine from patients with inborn errors of metabolism: 2-hydroxyglutaric aciduria and maple syrup urine disease, *J. Pharm. Biomed. Anal.*, 1997, 15, 1647-1659.

6 T. De Meyer, D. Sinnaeve, B. Van Gasse, E.R. Rietzschel, M.L. De Buyzere, M.R. Langlois, S. Bekaert, J.C. Martins and W. Van Criekinge, Evaluation of standard and advanced preprocessing methods for the univariate analysis of blood serum ^{1}H-NMR spectra, *Anal. Bioanal. Chem.*, 2010, 398(4), 1781-1790.

7 O. Beckonert, H.C. Keun, T.M. Ebbels, J. Bundy, E. Holmes, J.C. Lindon and J.K. Nicholson, Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts, *Nat. Protoc.*, 2007, 2(11), 2692-2703.

8 M.E. Bollard, E.G. Stanley, J.C. Lindon, J.K. Nicholson and E. Holmes, NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition, *NMR Biomed.*, 2005, 18(3), 143-162.

9 W.R. Wikoff, J.A. Gangoiti, B.A. Barshop and G. Siuzdak, Metabolomics identifies perturbations in human disorders of propionate metabolism, *Clin. Chem.*, 2007, 53(12), 2169-2176.

10 T.M. Ebbels, H.C. Keun, O.P. Beckonert, M.E. Bollard, J.C. Lindon, E. Holmes and J.K. Nicholson, Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: The consortium on metabonomic toxicology screening approach, *J. Proteome Res.*, 2007, 6(11), 4407-4422.

11 J. Nicholson, H. Keun and T. Ebbels, COMET and the challenge of drug safety screening, *J. Proteome Res.*, 2007, 6(11), 4098-4099.

12 R. Schicho, R. Shaykhutdinov, J. Ngo, A. Nazyrova, C. Schneider, R. Panaccione, G.G. Kaplan, H.J. Vogel and M. Storr, Quantitative metabolomic profiling of serum, plasma, and urine by (1)H NMR spectroscopy discriminates between patients with inflammatory bowel disease and healthy individuals. *J. Proteome Res.*, 2012, 11(6), 3344-3357.

13 T. Gebregiworgis and R. Powers, Application of NMR metabolomics to search for human disease biomarkers, *Comb. Chem. High Throughput Screen*., 2012, 15(8), 595-610.

14 T. Kind, V. Tolstikov, O. Fiehn and R.H. Weiss, A comprehensive urinary metabolomic approach for identifying kidney cancer, *Anal. Biochem.*, 2007, 363(2), 185-195.

15 H.J. Issaq, T.J. Waybrighe and T.D. Veenstra, Cancer biomarker discovery: Opportunities and pitfalls in analytical methods, *Electrophoresis*, 2011, 32(9), 967-975.

16 O. Fiehn, Metabolomics--the link between genotypes and phenotypes, *Plant Mol. Biol*., 2002, 48(1-2), 155-171.

17 K.A. Hamersky, C.E. Merrywell, F. Fang and C.K. Larive, "Metabolic Profiling" in *NMR Spectroscopy in Pharmaceutical Analysis*, U. Holzgrabe, I. Wawer and B. Diehl, Eds. Elsevier, Oxford, 2008, pp.233-267.

18 A. Zhang, H. Sun, P. Wang, Y. Han and X. Wang, Modern analytical techniques in metabolomics analysis, *Analyst*, 2012, 137(2), 293-300.

19 C.K. Larive, G.A. Barding, Jr. and M.M. Dinges, NMR spectroscopy for metabolomics and metabolite profiling, *Anal. Chem.*, 2015, 87(1), 133-146.

20 G.A. Barding, Jr., R. Salditos and C.K. Larive, Quantitative NMR for bioanalysis and metabolomics, *Anal. Bioanal. Chem*., 2012, 404(4), 1165-1179.

21 F.F. Brown, I.D. Campbell, P.W. Kuchel and D.L. Rabenstein, Human erythrocyte metabolism studies by ^{1}H spin echo NMR, *FEBS Lett.*, 1977, 82(1), 12-16.

22 J.R. Bales, D.P. Higham, I. Howe, J.K. Nicholson and P.J. Sadler, Use of high-resolution proton nuclear magnetic resonance spectroscopy for rapid multi-component analysis of urine, *Clin. Chem.*, 1984, 30(3), 426-432.

23 J.K. Nicholson, M.J. Buckingham and PJ. Sadler, High resolution ^{1}H n.m.r. studies of vertebrate blood and plasma, *Biochem. J*., 1983, 211(3), 605-615.

24 J.M. Fonville, A.D. Maher, M. Coen, E. Holmes, J.C. Lindon and J.K. Nicholson, Evaluation of full-resolution J-resolved ^{1}H NMR projections of biofluids for metabonomics information retrieval and biomarker identification, *Anal. Chem*., 2010, 82(5), 1811-1821.

25 T.R. Brown and R. Stoyanova, NMR spectral quantitation by principal-component analysis. II. Determination of frequency and phase shifts, *J. Magn. Reson. B*, 1996, 112(1), 32-43.

26 J.C. Cobas, M.A. Bernstein, M. Martin-Pastor and P.G. Tahoces, A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data, *J. Magn. Reson.*, 2006, 183(1), 145-151.

27 R. Stoyanova, A. W. Nicholls, J. K. Nicholson, J. C. Lindon, T. R. Brown, *J. Magn. Reson.*, 2004, 170, 329.

28 C.G. Tang, An analysis of baseline distortion and offset in NMR spectra, *J. Magn. Reson*., 1994, 109(2), 232-240.

29 A.K. Leung, F. Chau and J. Gao, A review on applications of wavelet transform techniques in chemical analysis: 1989–1997, *Chemom. Intell. Lab. Syst*., 1998, 43, 165-184.

30 N. Trbovic, F. Dancea, T. Langer and U. Günther, Using wavelet de-noised spectra in NMR screening, *J. Magn. Reson*., 2005, 173(2), 280-287.

31 J. C. Hoch and A.S. Stern, *NMR Data Processing*, Wiley-Liss, New York, 1996.

32 H. Witjes, W.J. Melssen, H.J.A. in t'Zandt, M. van der Graaf, A. Heerschap and L.M.C. Buydens, Automatic correction for phase shifts, frequency shifts, and lineshape distortions across a series of single resonance lines in large spectral data sets, *J. Magn. Reson.*, 2000, 144(1), 35-44.

33 H.Witjes, M. van den Brink, W.J. Melssen and L.M.C. Buydens, Automatic correction of peak shifts in Raman spectra before PLS regression, *Chemom. Intell. Lab. Syst.*, 2000, 52(1), 105-116.

34 A. Craig, O. Cloarec, E. Holmes, J.K. Nicholson and J.C. Lindon, Scaling and normalization effects in NMR spectroscopic metabonomic data sets, *Anal. Chem*., 2006, 78(7), 2262-2267.

35 F. Dieterle, A. Ross, G. Schlotterbeck and H. Senn, Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics, *Anal. Chem*., 2006, 78(13), 4281-4290.

36 J. Forshed, I. Schuppe-Koistinen and S. P. Jacobsson, Peak alignment of NMR signals by means of a genetic algorithm, *Anal. Chim. Acta*, 2003, 487, 189-199.

37 E. Holmes, P.J.D. Foxall, J.K. Nicholson, G.H. Neild, S.M. Brown, C.R. Beddell, B.C. Sweatman, E. Rahr, J.C. Lindon, M. Spraul and P. Neidig, Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states, *Anal. Biochem*., 1994, 220(2), 284-296.

38 R.J.O. Torgrip, M. Åberg, B. Karlberg and S.P. Jacobsson, Peak alignment using reduced set mapping, *J. Chemom*., 2003, 17(11), 573-582.

39 O. Cloarec, M.E. Dumas, A. Craig, R. H. Barton, J. Trygg, J. Hudson, C. Blancher, D. Gauguier, J. C. Lindon, E. Holmes and J.K. Nicholson, Statistical total correlation spectroscopy: an exploratory approach for latent biomarker identification from metabolic ^{1}H NMR data sets *Anal. Chem*., 2005, 77(5), 1282-1289.

40 E. Holmes, O. Cloarec and J.K. Nicholson, Probing latent biomarker signatures and in vivo pathway activity in experimental disease states via statistical total correlation spectroscopy (STOCSY) of biofluids: application to HgCl2 toxicity, *J. Proteome Res*., 2006, 5(6), 1313-1320.

41 L.M. Smith, A.D. Maher, O. Cloarec, M. Rantalainen, H. Tang, P. Elliott, J. Stamler, J.C. Lindon, E. Holmes and J.K. Nicholson, Statistical correlation and projection methods for improved information recovery from diffusion-edited NMR spectra of biological samples, *Anal. Chem*., 2007, 79(15), 5682-5689.

42 S. Wold, K. Esbensen and P. Geladi, Principal component analysis, *Chemom. Intell. Lab. Syst*., 1987, 2, 37-52.

43 S. Wold, M. Sjostrom and L. Eriksson, PLS.regression: A basic tool of chemometrics, *Chemom. Intell. Lab. Syst*., 2001, 58, 109-130.

44 S. Wei, J. Zhang, L. Liu, T. Ye, G.A. Gowda, F. Tayyari and D. Raftery, Ratio analysis nuclear magnetic resonance spectroscopy for selective metabolite identification in complex samples, *Anal. Chem.*, 2011, 83(20), 7616-7623.

45 L. Csenki, E. Alm, R. Torgrip, K.M. Aberg, L.I. Nord, I. Schuppe-Koistinen and J. Lindberg, Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional ^{1}H NMR data, *Anal. Bioanal. Chem*., 2007, 389(3), 875-885.

46 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 and J.C. Lindon, Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples, *J. Pharm. Biomed. Anal.*, 1994, 12(10), 1215-1225.

47 N.P.V. Nielsen, J.M. Carstensen and J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, *J. Chromatogr. A*, 1998, 805, 17-35.

48 G. Tomasi, F. van den Berg and C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, *J. Chemom*., 2004, 18(5), 231-241.

49 D. Bylund, R. Danielsson, G. Malmquist and K.E. Markides, Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography–mass spectrometry data, *J. Chromatogr. A*, 2002, 961, 237–244.

50 F.H. Larsen, F. van den Berg and S.B. Engelsen, An exploratory chemometric study of ^{1}H NMR spectra of table wines, *J. Chemom*., 2006, 20(5), 198-208.

51 R.J. Torgrip, J. Lindberg, M. Linder, B. Karlberg, S.P. Jacobsson, J. Kolmert, I. Gustafsson and I. Schuppe-Koistinen, New modes of data partitioning based on PARS peak alignment for improved multivariate biomarker/biopattern detection in ^{1}H-NMR spectroscopic metabolic profiling of urine *Metabolomics*, 2006, 2(1), 1-19.

52 E. Alm, R.J. Torgrip, K.M. Aberg, I. Schuppe-Koistinen and J. Lindberg, A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support, *Anal. Bioanal. Chem*., 2009, 395(1), 213-223.

53 F. Savorani, G. Tomasi and S.B. Engelsen, icoshift: A versatile tool for the rapid alignment of 1D NMR spectra, *J. Magn. Reson.*, 2010, 202(2), 190-202.

54 K.A. Veselkov, J.C. Lindon, T.M. Ebbels, D. Crockford, V.V. Volynkin, E. Holmes, D.B. Davies and J.K. Nicholson, Recursive segment-wise peak alignment of biological (1)h NMR spectra for improved metabolic biomarker recovery, *Anal. Chem.*, 2009, 81(1), 56-66.

55 D.E. Brown, Fully automated baseline correction of 1D and 2D NMR spectra using Bernstein polynomials, *J. Magn. Reson. A*, 1995, 114, 268-270.

56 S. Golotvin, A. Williams, Improved baseline recognition and modeling of FT NMR spectra, *J. Magn. Reson.* 2000, 146(1), 122-5.

57 G. Schulze, A. Jirasek, M.M. Yu, A. Lim, R.F. Turner and W.B. Michael, Investigation of selected baseline removal techniques as candidates for automated implementation, *Appl. Spectrosc.*, 2005, 59(5), 545-574.

58 A. Savitzky and M.J.E. Golay, Smoothing and differentiation of data by simplified least-squares procedures, *Anal. Chem.* 1964, 36(8), 1627-1639.

59 T.L. Hwang and A.J. Shaka, Excitation sculpting using arbitrary waveforms and pulsed field gradients, *J. Magn. Reson. A*, 1995, 112, 275-279.

60 R. Bro and A.K. Smilde, Principal component analysis, *Anal. Methods*, 2014, 6, 2812-2831.

61 B. Worley, R. Powers, Multivariate analysis in metabolomics, *Curr. Metabolomics*, 2013, 1, 92-107.

62 R. A. van den Berg, H.C.J. Hoefsloot, J.A. Westerhuis, A.K. Smilde and M.J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, *BMC Genomics*, 2006, 7, 142.

63 G.A. Barding, D.J. Orr, S. Sathnur and C.K. Larive, VIZR – An automated chemometric tool for metabolic profiling, *Anal. Bioanal. Chem.*, 2013, 405, 8409-8417.

64. O. Cloarec, M. E. Dumas, J. Trygg, A. Craig, R.H. Barton, J.C. Lyndon, J.K. Nicholson, E. Holmes, Evaluation of the orthogonal projection on latent structure model limitations caused by chemical shift variability and improved visualization of biomarker changes in ^{1}H NMR spectroscopic metabonomic studies, *Anal. Chem.* 2005, 77 (2), 517-526.