Journal of Magnetic Resonance 202 (2010) 190–202

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

*i*coshift: A versatile tool for the rapid alignment of 1D NMR spectra

F. Savorani, G. Tomasi, S.B. Engelsen *

Quality & Technology, Department of Food Science, Faculty of Life Sciences, University of Copenhagen, Rolighedsvej 30, 1958 Frederiksberg C, Denmark

ARTICLE INFO

Article history:
Received 25 August 2009
Revised 3 November 2009
Available online 18 November 2009

Keywords:
Preprocessing
NMR
Alignment
Metabonomics
Chemometrics
Algorithm

ABSTRACT

The increasing scientific and industrial interest towards metabonomics takes advantage from the high qualitative and quantitative information level of nuclear magnetic resonance (NMR) spectroscopy. However, several chemical and physical factors can affect the absolute and the relative position of an NMR signal and it is not always possible or desirable to eliminate these effects *a priori*. To remove misalignment of NMR signals *a posteriori*, several algorithms have been proposed in the literature. The *i*coshift program presented here is an open source and highly efficient program designed for solving signal alignment problems in metabonomic NMR data analysis. The *i*coshift algorithm is based on correlation shifting of spectral intervals and employs an FFT engine that aligns all spectra simultaneously. The algorithm is demonstrated to be faster than similar methods found in the literature making full-resolution alignment of large datasets feasible and thus avoiding down-sampling steps such as binning. The algorithm uses missing values as a filling alternative in order to avoid spectral artifacts at the segment boundaries. The algorithm is made open source and the Matlab code including documentation can be downloaded from www.models.life.ku.dk.

1. Introduction

In recent years a constantly increasing interest has been devoted towards the – omics sciences in which Nuclear Magnetic Resonance Spectroscopy (NMR) plays a central role since it is able to provide, in a short time, a reliable and unique metabolic fingerprint of complex chemical and/or biological matrices. However, the large amount and the high complexity of the acquired data make it challenging to disentangle the sought meaningful information, and a major effort has been made lately for providing the researchers with mathematical and statistical tools able to cope, in a reasonable time, with such overwhelming information. Multivariate data analysis methods have been developed and tailored for dealing with this new complex problem, providing a step forward into data handling and interpretation of metabolomic and metabonomic studies which are becoming more and more common. Multivariate data-mining in spectroscopy has a long history in near infrared spectroscopy but has a weaker tradition in nuclear magnetic resonance spectroscopy. While the first application of multivariate chemometric analysis to NMR spectra appeared in the early eighties [1], it was not until the early nineties that the field of metabonomics emerged and the highly complex metabolic fingerprints in NMR spectra of body fluids made the need for powerful multivariate data analysis obvious [2]. Chemometrics is now rapidly gaining momentum in the analysis of NMR spectra [3]. However, NMR data are not always readily suitable for the analysis by so-called bilinear chemometric methods. NMR spectra are generally a sequence of response signals, closely spaced Lorentzian peaks, differing in shape, intensity and position. While all these entities carry important information, changes in peak positions of the same analyte signal between samples does not conform to the bilinearity assumptions and thus deteriorate the chemometric modeling. In algebraic terms, bilinearity requires that each column (if samples are stored in the rows of a matrix) contain information about a signal originating from an identical common compound along all the samples in order the statistical approach to be able to work properly. However, especially due to small pH changes and intermolecular interactions, this is rarely the case with biological samples, wherefore it is imperative for multivariate exploratory metabonomics investigations aimed at biomarker profiling or pattern recognition studies, that the data be aligned before chemometric analysis. Of course, any effort should be made prior to and during the NMR analysis for assuring that the samples are being collected and prepared as homogeneously as possible (preparation protocols) and that the instrumental conditions and parameters are identical [4]. Despite standardization, spectral misalignments still occur and this is the reason why *a posteriori* aligning methods are required. The aim of this study is to provide the NMR spectroscopist with a simple, versatile and efficient algorithm for performing *a posteriori* alignment which uses the newest and fastest algorithmic techniques.

* Corresponding author. Fax: +45 3533 3245.
E-mail addresses: frsa@life.ku.dk (F. Savorani), gto@life.ku.dk (G. Tomasi), se@life.ku.dk (S.B. Engelsen).

1090-7807/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmr.2009.11.012

Different approaches have been explored for solving the alignment problem and have supplied appropriate solutions for many different experimental cases. The historical bases of such an approach relies on a very simple, but still largely used method, *binning* or *bucketing*, which involves a data reduction, performed through NMR signals integration within standardized spectral regions whose width commonly ranges between 0.01 and 0.05 ppm (*bins* or *buckets*) [5]. The major drawback of this solution is the loss of spectral resolution. When high resolution is required, other and more sophisticated alignment methods need to be considered such as dynamic time warping (DTW) or correlation optimized warping (COW) [6–8], which have been demonstrated to be effective on chromatographic data and also have been employed for solving simple NMR alignments with satisfactory results [9]. Apart from being computationally intensive, the main problem of these two approaches is that alignment is obtained by local stretching or compression. This is not really suitable for NMR signals because this model for correction works best when there is a positive correlation between peak width and shift.

Another important class of alignment methods finds all the relevant peaks present in an NMR spectrum. This converts spectra into a list of peaks and relative attributes thereby dramatically decreasing the dimensions of the dataset. Torgrip et al. have introduced and developed these methods [10–12]. The major drawbacks of these methods are the elimination of the information carried by the fine structure of the signal shape and the need to define some meta-parameters for the peak-picking procedure and subsequent alignment.

An alternative, more advanced, approach concerns designing algorithms able to perform an automatic NMR peak alignment with no or limited user intervention, trying meanwhile to keep all the relevant spectral information. The first attempts to develop such an algorithm involved the application of a genetic algorithm to align segments of spectra [13,14], the application of partial linear fit to align the spectra [15] and a search for the misaligned spectral region by a PCA based algorithm followed by a rigid shift [16]. None of these methods have been broadly adopted due to lack of alignment performance and/or high computational costs. Wong et al. [17] solved the problem of the computational inefficiency by employing a Fast Fourier Transformation (FFT) correlation engine to boost the algorithmic speed (PAFFT) and at the same time introduced the use of regular spectral intervals to be individually and rigidly aligned. Veskelov et al. [18] combined the properties of the peak picking methods with the FTT and interval features of PAFFT. The result is a rapid fully automated aligning process, able to recursively split the NMR spectra in meaningful intervening regions and to align their signal until a certain degree of goodness is reached. Similarly, a recently published method using a fuzzy Hough transform [12] is able to perform an automatic full spectral alignment. Although promising, these methods rely on a peak picking algorithm which need the user to set some meta-parameters which can dramatically affect the final result and especially those based on the Hough transform are computationally very expensive.

The trend towards interval based algorithms represents a key step forward in the research of a definitive solution to the alignment problem. As a characteristic, the chemical shift of each NMR signal (or pattern of them) depends on several factors and can independently change in any direction. Peaks that are adjacent or even overlapped in one spectrum can be baseline separated in other spectra and, as an extreme consequence, their relative position inverted. When dealing with the spectral features it is therefore preferable to reduce the global problem to smaller localized ones that can be found in specific spectral intervals. In this way, the shift occurring in opposite directions can be easily solved and eventually a global, full resolution aligned NMR dataset can be reconstructed. In practice, it is often only a small percentage of the total spectral width that suffers for misalignment problems and, besides increasing the chances of bad corrections, it is clearly inefficient to align what it is already aligned. An automatic search for the meaningful regions of intervention can be a desirable option for an aligning method, but what it is apparent for the average NMR spectroscopist can be very difficult (and time consuming) to implement algorithmically. In effect, it is rarely worth the effort of optimizing several meta-parameters in order to make a method work compared to manually select the regions of intervention. Fully automated procedures also normally do not allow user intervention and, if the achieved result is not satisfying, there are no alternatives.

The new algorithm presented in this study, *i*coshift, has been designed to provide the user with a versatile open source tool for spectral alignment which is based on rigid shift of intervals and which uses the FFT to boost the simultaneous alignment of all spectra in a dataset. *i*coshift combines a rapid optimized FFT engine, which brings the calculation times for large metabonomic datasets from hours (e.g., with COW), or minutes [18] to seconds, with optional interactive facilities for interval definitions and for alignment automation. It introduces the use of missing values for solving the interpolation problems that still represent an open issue for other algorithms. Plot and interactive facilities are also provided for the user to be able to immediately evaluate the achieved result. Applications of *i*coshift to solve different misalignment problems in real experimental NMR datasets will be illustrated in this paper. The Matlab® (2008b, The Mathworks Inc., Natick, MA, USA) code of *i*coshift, including a help section and a demo, is freely available for download from www.models.life.ku.dk.

2. Results and discussion

2.1. The icoshift algorithm

The *i*coshift algorithm, namely interval-correlation-shifting, which derives its name form the basic coshift algorithm [19], independently aligns each NMR signal to a target (which can optionally be an actual signal or a synthetic one like the average, or the median) by maximizing the cross-correlation between user-defined intervals. Fig. 1 shows an overview of *i*coshift results when applied to a misaligned set of human urine NMR spectra zoomed into a strongly misaligned region. The algorithm is comprised of three essential parts: (1) interval definition, (2) maximization of the cross-correlation of each interval by an FFT engine and (3) signal reconstruction (Fig. 2). For the three parts, several options are available depending on how the alignment problem is to be solved. Therefore, it is possible to define intervals automatically (e.g., allowing a full spectrum alignment with regularly spaced intervals, or adjacent intervals of user-defined length, or customized interval boundaries), set a boundary for the maximum local correction allowed for each interval and define the fill-in value for the reconstruction part (viz., a missing value or the first/last point in the segment).

The basic principle of *i*coshift is quite similar to other published methods for the alignment of spectral and chromatographic signals: Peak Alignment by FFT (PAFFT [17]), Recursive Peak Alignment by FFT (RAFFT [17]) and Recursive Segment-wise Peak Alignment (RSPA [18]) and can in fact be used as a computation engine for these methods as it is efficient and numerically sound. The algorithm was designed having in mind the wide range of misalignment problems that can be encountered in NMR spectroscopy (e.g., it allows the registration to a reference signal and completely customized-interval definitions) and can help reduce the emergence of artifacts related to the insertion/deletion model used for the alignment (namely, by inserting missing values instead of

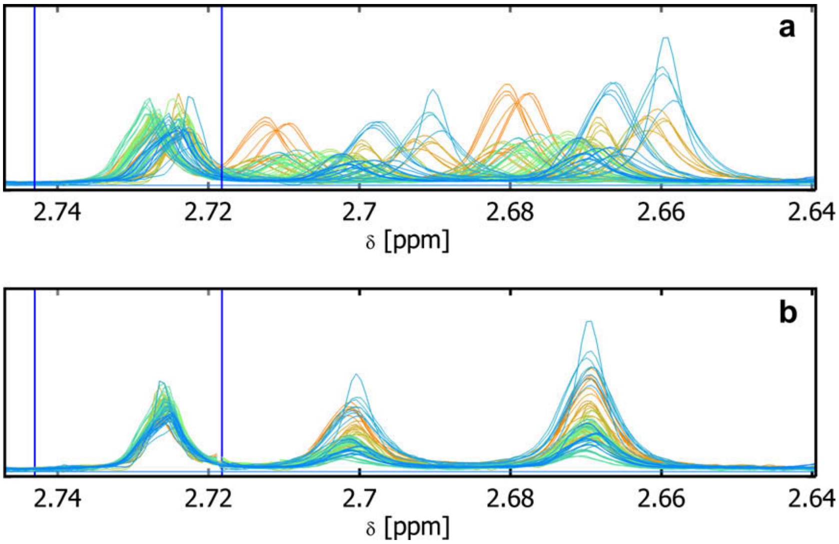

Fig. 1. Overview of the algorithm results. *i*coshift provides aligned NMR datasets working on user-defined intervals.

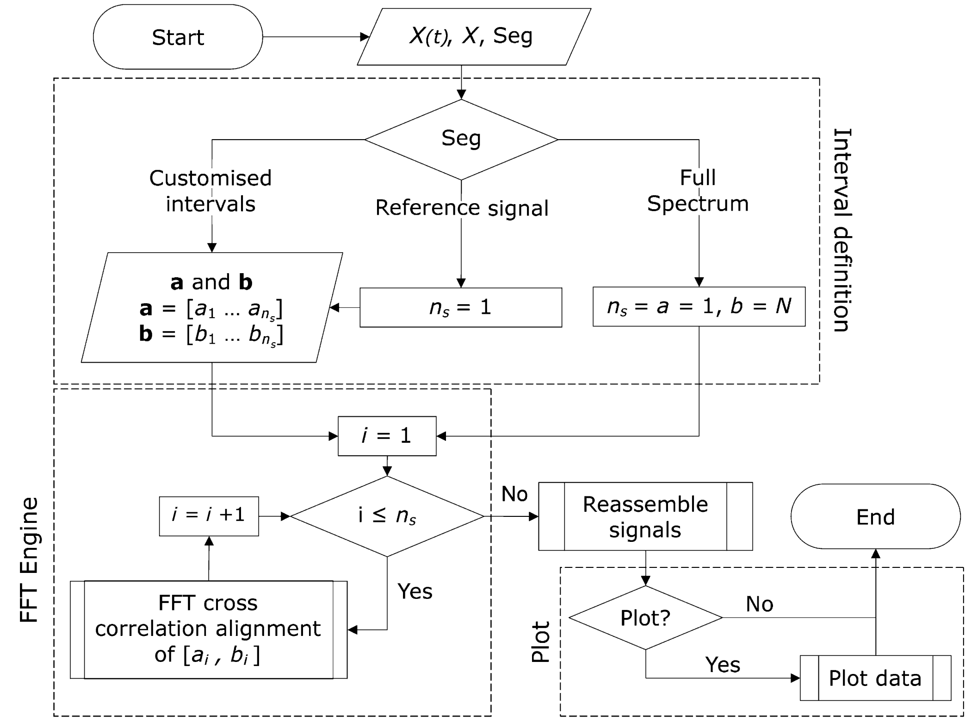

Fig. 2. Simplified flow chart of the algorithm. X is the dataset to be aligned. The n_s intervals, whose optimal lag is found through the FFT engine, have form $[a_i, b_i]$ for $i = 1, \ldots, n_s$. and $X(t)$ is the target. The Seg variable defines which type of alignment is used. N is the signal length.

repeating the value on the boundary). However, like the majority of current alignment methods, it cannot correct for a change in the order of peaks. The details of the implementation of *i*coshift are given in the experimental section.

Three experimental NMR datasets, covering a multiplicity of alignment problems to be solved, were used to assess and illustrate the performance of the *i*coshift algorithm. The achieved results are presented and discussed case by case. The figures presented are based on the actual plot facilities of the *i*coshift program.

2.2. Case 1: the wine data

The wine dataset consists of 40 spectra of different table wines [9] in which the spectra were processed and reduced to 8712 data points covering a 5.5 ppm spectral region (from 6.00 to 0.50 ppm). The wine samples included red, white and rosé wines that were not buffered and/or pH adjusted before NMR analysis. This introduced a broad range of shifts of all the pH dependent signals found in the organic acids/amino acids region. Simple alignment according to the TSP (3-(trimethylsilyl)-propionic acid-d_4) reference signal cannot correct for the pH dependent shifts nor for the dominant ethanol signals. This can be observed in Fig. 3a for example by inspection of ethanol's ^{13}C satellites (signals No. 1′, 1″, 8′ and 8″). In the original publication Larsen et al. [9] used COW [20] for solving the misalignment problem, but since COW was only able to correct for the shift of the dominant ethanol signals, it was necessary to use an *ad hoc* multistep interval procedure, utilizing coshift, for improving the alignment of a much smaller signal such

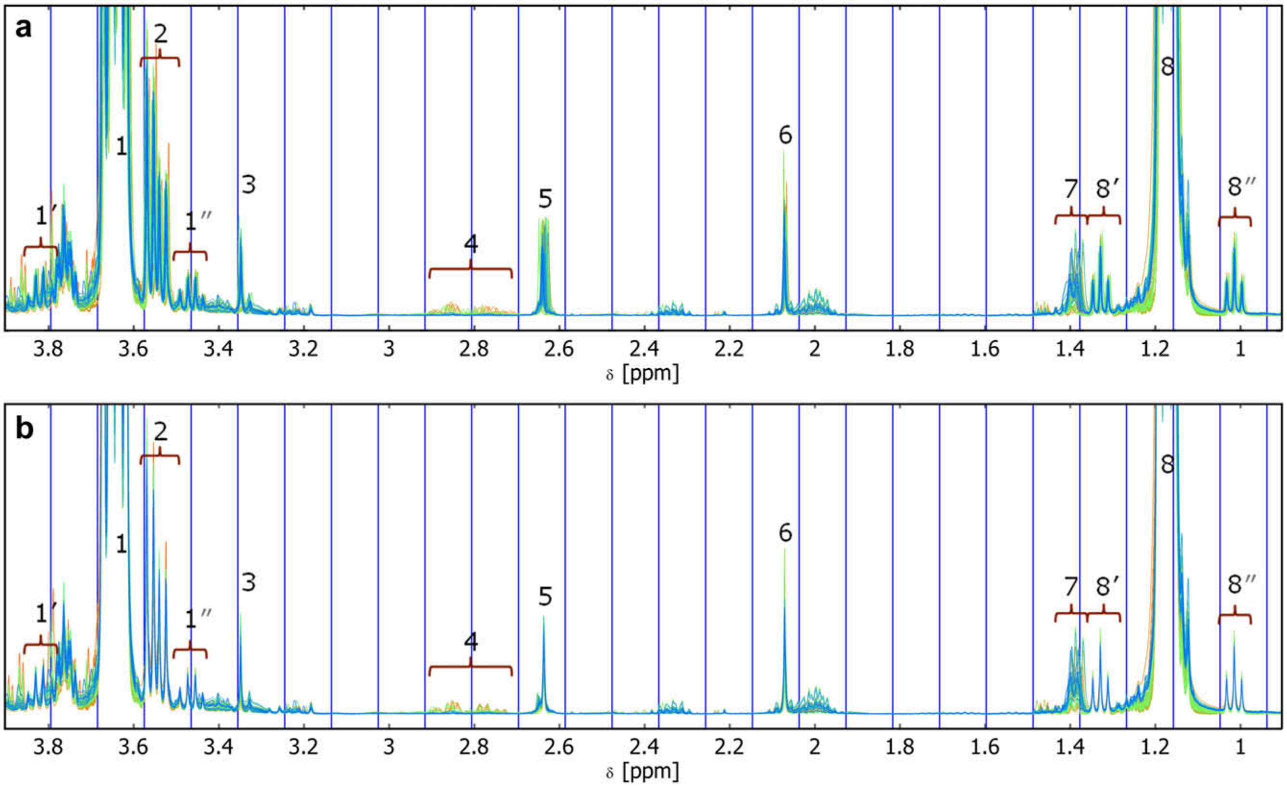

Fig. 3. Comparison between raw (a) and *i*coshift aligned (b) NMR spectra for the table wine dataset. A regular splitting into 50 intervals was performed by the algorithm and the full alignment process took less than 1 s. The main resonance peaks are: 1. ethanol (–CH_2–); 1′. and 1″. ^{13}C satellites of peak 1; 2. glycerol; 3. methanol; 4. malic acid; 5. succinic acid; 6. acetic acid; 7. lactic acid; 8. ethanol (–CH_3); 8′. and 8″. ^{13}C satellites of peak 8.

as the lactic acid one (signal No. 7 in Fig. 3). The subsequent multivariate quantitative modeling by *i*PLS (interval partial least squares) regression [21] showed greatly improved results making the authors able to distinguish among the different types of wine and to obtain reliable calibration models for the prediction of some indigenous wine metabolites. Just like the *i*PLS algorithm, the *i*coshift algorithm can split the dataset into a user-defined number of intervals (or into regular spaced intervals). Fig. 3 shows the 40 superimposed wine spectra before and after the *i*coshift correction using regular intervals. Evidently, the spectral alignment problems are corrected for both the dominant and the smaller peaks. The alignment of the spectra is produced by the *i*coshift algorithm by the Matlab command line #1 in Table 1, which applies *i*coshift to the spectra in WineData using the average spectrum as a reference spectrum and dividing the spectra up into 50 regularly spaced intervals. The result is shown in Fig. 3b with the 50 regular intervals marked by straight vertical lines. For this dataset, the *i*coshift alignment takes less than 1 s and the AlignedWineData is ready for subsequent multivariate data analysis. As apparent from the figure, almost all the pH dependent broadly misaligned signals are efficiently shift corrected. However, one signal remains evidently misaligned by this procedure (the lactic acid doublet centered at 3.95 ppm; #7), which illustrates one fundamental problem with unsupervised alignment algorithms: in this case, the spectral splitting into regular intervals left one part of the lactic acid signal in one interval and another part of the signal in the adjacent interval, which made it impossible for the algorithm to provide an effective correction for the shift. In *i*coshift, this problem can easily be overcome by using a different number of intervals. However, in many cases, a custom interval solution is preferable for optimal division of the sample spectra into baseline separated intervals. This can be done by providing the *i*coshift algorithm with a vector ("*custom_intervals*" in Table 1) containing the boundaries of the desired intervals. The selected intervals do not necessarily have to be adjacent and can be of flexible size. The alignment of the WineData set to the average spectrum target was applied using the Matlab command line #2 in Table 1. The third element in the options parameter makes the *i*coshift algorithm performing an automatic

Table 1
Simple command lines in Matlab for applying *i*coshift to the NMR datasets in the three cases. Command line #1 applies *i*coshift to the spectra in WineData using the average spectrum as a reference spectrum and dividing the spectra up into 50 regularly spaced intervals. Command line #2 applies *i*coshift to the WineData using the average spectrum as a reference spectrum, dividing the spectra up into intervals defined by the vector "custom_intervals" and using the *x*-axis definition (ppm) described in the vector "ppm_scale" for plotting. The fourth parameter 'f' is optional and makes the *i*coshift algorithm automatically determine the maximum allowed correction for each interval. Command line #3 applies *i*coshift to the spectra in NibathData using the average spectrum as a reference and the interval settings described by the "*custom_intervals*" vector. The automatic fast search for the suitable maximum allowed shift for each interval does not need to be specified since it is set as the default. Command line #4 applies *i*coshift to the PlasmaData using the average spectrum as a reference and using the signal contained in the interval described in "reference_region" as a guide for shifting the entire spectra.

	Command line
#1	AlignedWineData = icoshift('average', WineData, 50);
#2	AlignedWineData = icoshift('average', WineData, custom_intervals, 'f', [2 0 1], ppm_scale);
#3	AlignedNibathData = icoshift('average', NibathData, custom_intervals);
#4	AlignedPlasmaData = icoshift('average', PlasmaData, reference_region);

full-spectrum correlation-shift before the interval alignment. This additional step is an advantage (more in diminishing the risk of misalignment than computationally) in cases with large shifts as it allows the spectra to be roughly pre-aligned according to the dominant signals. The results of both alignment steps are shown in Fig. 4 for the challenging lactic acid region. The full-spectrum correlation-shifting step was effective in aligning the intense ethanol peaks as well as their ethanol ^{13}C satellites (since they are shifted exactly like their main ^{1}H signals) (Fig. 4b), the lactic acid doublet was not aligned because of its pH dependency. In order to align these signals the customized interval-correlation-shifting step was necessary for obtaining fully-aligned signals (Fig. 4c). Many efforts have been dedicated to optimize the algorithm speed and for this dataset the time of calculation for this two-step procedure is only about 0.2 s. On the contrary, computation time for COW is significantly higher, especially when the optimization step is taken into account. In terms of quality of the alignment, the results obtained with COW are acceptable, although the lactic acid region is not handled correctly (cf. Supplementary Fig. 2). Similar observations can be made about RSPA, which is also slower than *i*coshift because of the computation overhead related to the peak picking and the recursion. RSPA results are comparable with *i*coshift, but not quite as good for the lactic acid region. However, this might depend on the choice of the meta-parameters, and other values might lead to improved results. The optimization of these parameters and a detailed comparison between *i*coshift and RSPA is beyond the scope of this paper and is left for future research.

A simple but illustrative way to demonstrate the successful alignment of the spectra is to compare the performances of multivariate PLS regression models to the content of chemical constituents. In this case, PLS prediction of lactic acid was calculated using mean-centered data and repeated-random cross-validation for determining the number of significant PLS components. Fig. 5 shows a comparison between the calibration curves obtained using either the unaligned raw spectra (Fig. 5a) or the *i*coshift-aligned spectra (Fig. 5b). Clearly, the signal alignment has improved the PLS model significantly by reducing its complexity and obtaining a better correlation between measured and predicted values of lactic acid content. A similar, but more dramatic improvement is obtained on the PLS model built for the ethanol calibration. In this case the Root Mean Square Error of Cross-Validation (RMSECV) lowered from 0.351 to 0.177 and the R^2 improved from 0.79 to 0.95 (Supplementary Figs. 1a and b). It is important to bear in mind that spectral alignment also can be a destructive process as it can remove useful physical information related to the signal shifts in the spectra. This can be demonstrated by the ability of a full-spectrum PCA model (mean-centered data) to differentiate among the three different types of wine, which is clearly lost after the alignment process (compare Fig. 5c and d). The shifts induced by different acidity (pH) of the three kind of wine, carrying the information about their nature, are removed after the alignment. This is a fundamental characteristic of any spectral alignment process that always has to be taken into account. However, it is possible to preserve the information about the shift corrections for all samples and intervals using an optional output (“ind”) which collects it into a table. If desired, this table can be appended to the NMR dataset for a multivariate data analysis that preserves and isolates the shift information. The wine data demonstrates how spectral alignment can largely be performed in a fully automated operator-blinded mode (unsupervised) but also may lead to erroneous results and/or to deterioration of the information sought. In *i*coshift, these problems have been solved by offering an interactive mode that allows the user to define the intervals that are to be aligned leaving the rest unaffected. The use of the customized-interval definition has the benefit that user-written algorithms for automatic interval definition [18] can be directly interfaced to the *i*coshift program.

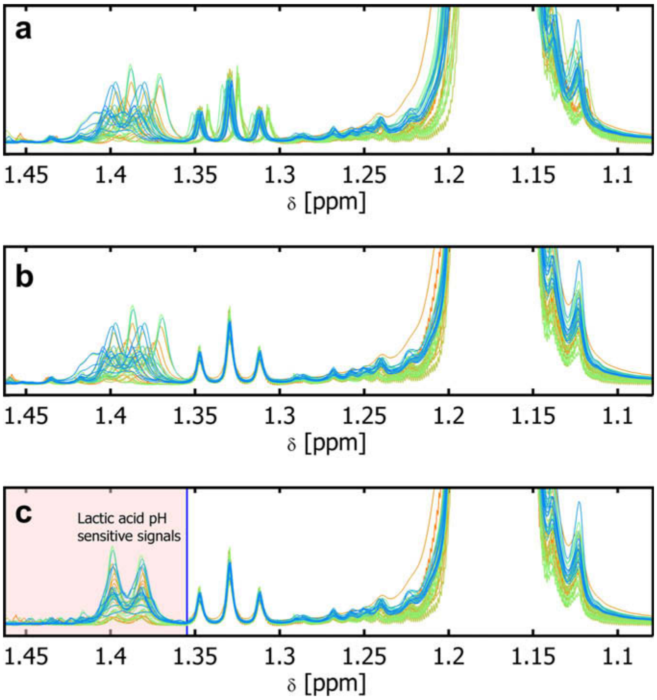

Fig. 4. Customized interval alignment of the wine dataset zoomed into the lactic acid NMR spectral region. The figure clearly illustrates the two steps of the alignment, starting from the raw data (a), passing through the full-spectrum correlation-shifted spectra (b) and obtaining the customized interval-correlation-shifted spectra (c). The intervals selected for the *i*coshift are highlighted with a background color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

2.3. Case 2: the nickel-bath data

In order to test the speed and performance of the *i*coshift algorithm under more realistic metabonomic conditions, an unusually large NMR dataset (460 × 40,905) was investigated. The nickel-bath dataset is part of a study aimed at following the evolution of organic additives in an electroplating nickel bath (unpublished data). The superimposed raw spectra (Fig. 7) show strong and broad signal misalignments along the whole spectral region. Succinic acid was added as an internal standard and its largely misaligned singlet can be found at approximately 2.5 ppm in the aliphatic region. The aqueous solutions were not buffered prior to NMR analysis causing also the lactic acid resonance peak to move within an almost 0.3 ppm wide region which is unusually broad for normal NMR datasets. Moreover, significant misalignments are also present in the aromatic region for the signals of some additives that were of great interest for the study of these nickel-bath samples. This dataset is particularly challenging because the multiple signal shifts present are of very different magnitude depending on the pH susceptibility of the molecules. This feature represents an insurmountable problem for the alignment algorithms based on FFT cross-correlation published so far [17,18,20] because they use a predetermined and invariable allowed maximum shift for the spectra to be aligned. The maximum allowed shift is thus identical for each interval in order to prevent excessive shifting and misalignment on a local scale. However, in practice, since the intervals can be of different size, the required maximum allowed shift for one interval can be wider than the size of another interval. This will seriously limit the algorithm’s flexibility and hamper its ability to achieve satisfactory results. In order to overcome this limitation, the *i*coshift algorithm is implemented with an (optional) automatic procedure for finding the maximum

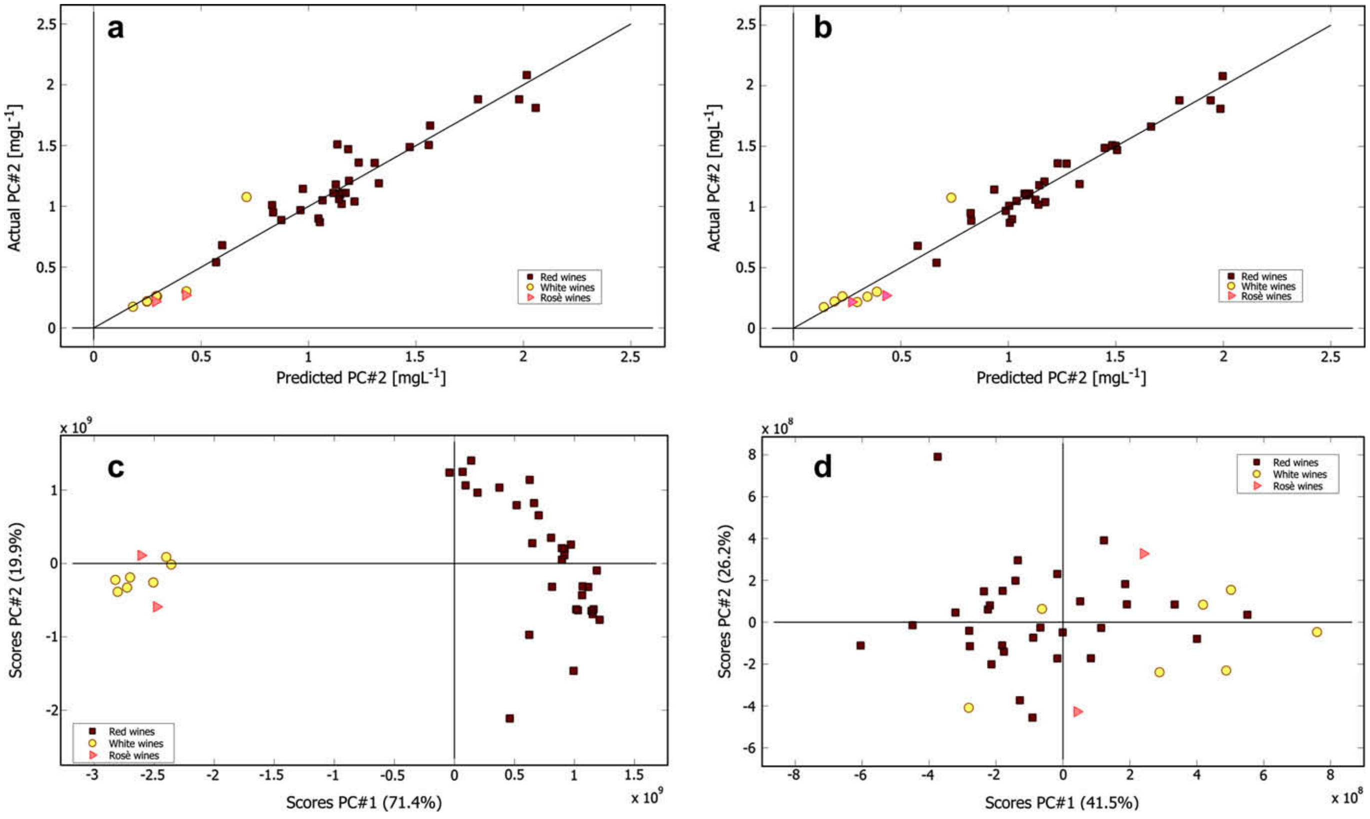

Fig. 5. PLS performances and PCA score plots for the wine dataset. Subplots (a) and (b) show the full-random cross-validated PLS calibration curves for the content of lactic acid obtained before (a, RMSECV = 0.137 R^2 = 0.93) and after (b, RMSECV = 0.104 R^2 = 0.96) the *i*coshift alignment when only the NMR region containing the lactic acid doublet (1.45–1.35 ppm) is taken into consideration. Subplots (c) and (d) show the PCA score plots calculated on the whole NMR region before (c) and after (d) the *i*coshift alignment.

allowed shift of each interval. The results of the *i*coshift alignment to the average spectrum for the Ni-bath dataset (NibathData in line #3 of Table 1) using custom intervals are illustrated in Fig. 6. Despite the large dimension of the dataset, the *i*coshift alignment was completed in only 4.74 s. After the *i*coshift alignment all the selected intervals appear fairly well aligned and in particular interval No. 11, containing the singlet peak of the succinic acid, is now well aligned (Fig. 6c). Such a broad misalignment could not be optimally corrected for using other aligning methods such as COW [20] and RSPA [18] (Fig. 7). The poor alignment performance of COW and RSPA are related to the automated segmentation (both) and in the compression/expansion model for alignment (COW). The problem is that the succinic acid peak does not always end up in the same segment in the target spectrum and in the sample to be aligned. For COW, the problem cannot be solved in an efficient way just by using custom segment boundaries because the compression/expansion model seeks the alignment of the succinic acid interval by intervening on the previous intervals. Therefore, the user-defined segmentation would also affect the other intervals making the overall procedure lengthy, tedious and without a guaranteed success. The problem of RSPA is essentially that the interval boundaries for the succinic acid are very close to the peak. When the intervals are not matched between sample and target, they cannot be corrected because no peaks are present in the target to estimate the correct shift.

2.4. Case 3: the rat plasma data

It is not always optimal to align a dataset by splitting it into smaller intervals since complex peak shape can be fundamental to the sought information. This is normally the case in typical metabonomic investigations of human (or animal) body biofluids (blood, urine, fecal liquid, etc.) in which the complexity of the pattern of resonances makes it difficult to define customized intervals without increasing the risk of deteriorating the relevant information. As a common practice metabonomic datasets are acquired under strictly defined experimental conditions, listed by a protocol, limiting as much as possible the risk of acquiring noisy spectra [22]. Still, some unwanted variations that may arise both during the sample preparation and during NMR acquisition can lead to slightly misaligned spectra even when they have been referenced to an internal standard. Indeed, for this kind of metabonomic datasets (plasma and serum samples) the common referencing molecules TSP or DSS (2,2-dimethyl-2-silapentane-5-sulfonate) have proved to be unreliable because unpredictable interactions with blood proteins make their signal vary. Therefore, it has become common practice to align the spectra according to pH unaffected signals present in a suitable concentration in all the spectra [22]. A perfect candidate for animal blood samples is the α-D-glucopyranose anomeric doublet, centered at 5.23 ppm, just on the right side of a broad lipid olefinic resonance at about 5.27 ppm. A plot of the complete rat plasma spectral dataset is shown in Fig. 8. The α-D-glucose region is highlighted and amplified in the insets a and b representing the raw and the aligned spectra, respectively (alignment result produced by command line #4 of Table 1). In practice, *i*coshift searches for the best local cross-correlation for the provided reference region and then shifts the whole spectra left or right according to the found shift indexes. Although the correction is small, it improves the dataset homogeneity and helps the interpretation of the results of a subsequent multivariate data analysis. The relatively large metabolomic dataset, including 72 plasma samples and consisting of 21,727 data points covering a 6.86 ppm wide spectral region (from 5.86 to −1.00 ppm) required only a fraction of second to be aligned (Table 2).

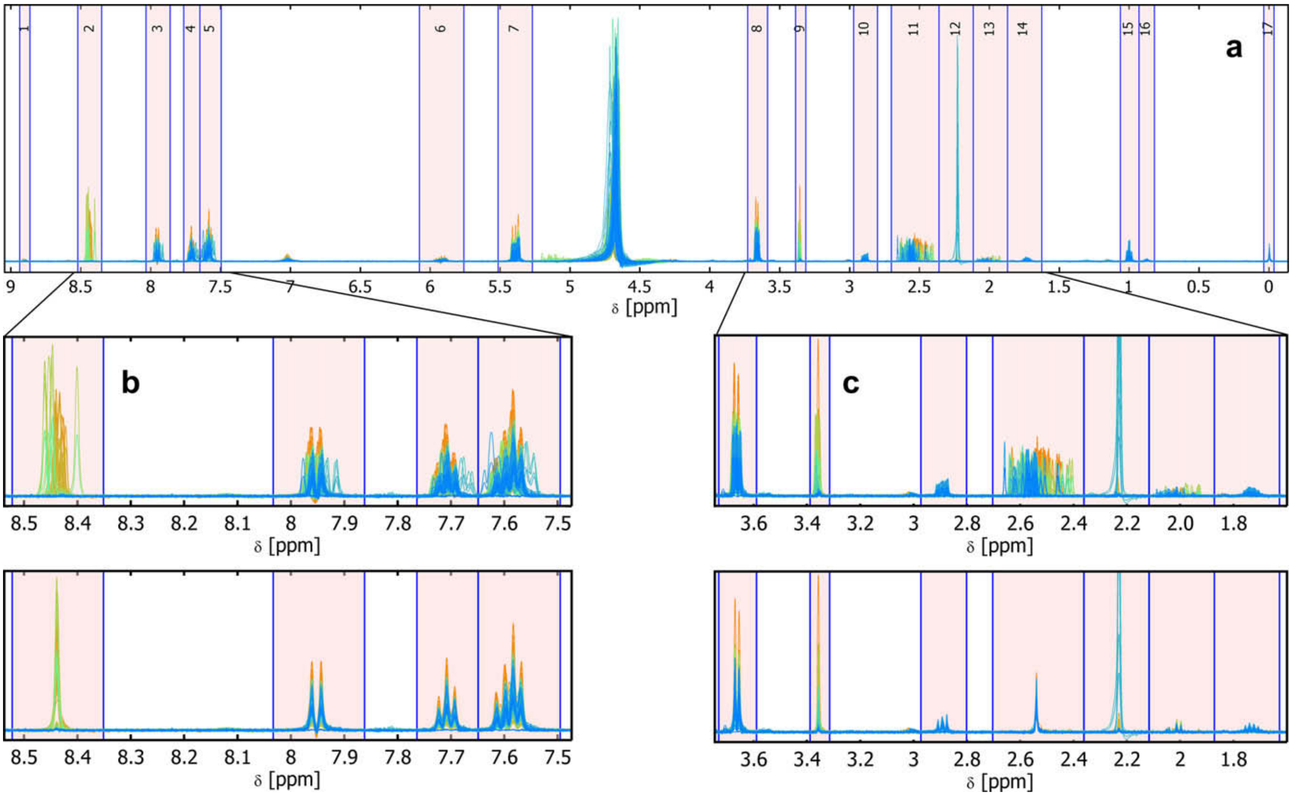

Fig. 6. Customized interval alignment of the nickel-bath data. The dataset is split up into 17 intervals with automatic determination of the maximum allowed shift for each interval. The 17 user-defined selected intervals are numbered on top of the figure and their background is colored. The peak pointed out by $*$ is the singlet resonance peak of succinic acid which was used as an internal standard. (b) and (c) show details of the aromatic and the aliphatic NMR regions, respectively. By mouse-clicking on one spectrum in either the raw spectra window or in the aligned spectra window, the spectrum becomes highlighted in both windows allowing the user to inspect the alignment result for the individual selected spectrum. This feature is demonstrated in the inset of the aromatic region (b) in which spectrum No. 409 was selected. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

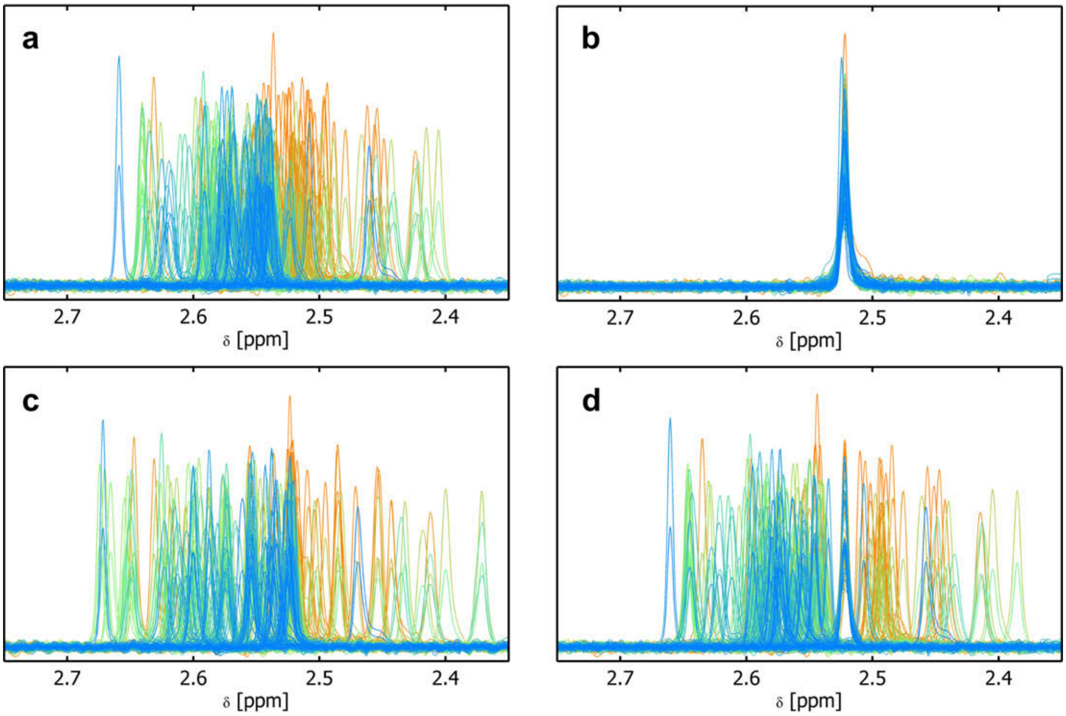

Fig. 7. Comparison of the alignment results of three different algorithms. The succinic acid region of the nickel-bath dataset is plotted after the full dataset has been aligned using different methods: (a) raw data; (b) *i*coshift; (c) COW ($\ell = 175$ points and $t = 10$) and (d) RSPA. Maximum allowed shift = 550 points for all methods.

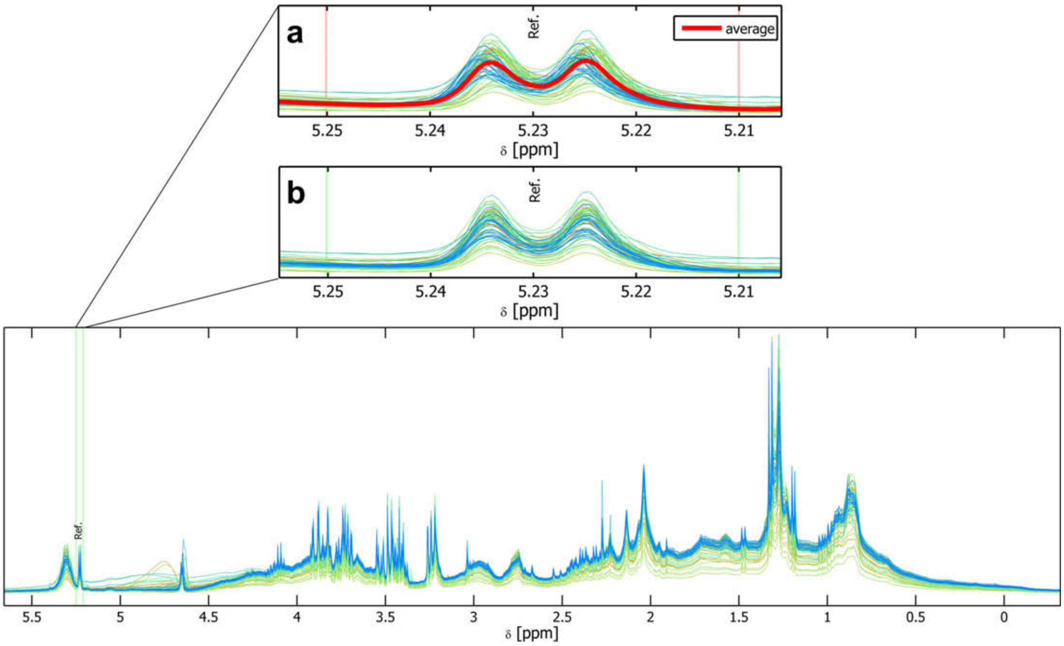

Fig. 8. *i*coshift alignment based on a reference signal of the rat plasma metabolomic dataset. The chosen reference region (5.25–5.21 ppm) contains the anomeric doublet of α-D-glucopyranose which was chosen for the algorithm to drive the alignment of the dataset according to a rigid shift of whole spectra. The average spectrum used as the target is highlighted in inset (a) which shows the raw data zoomed into the selected reference region as well as inset (b) shows the *i*coshift aligned data.

Because of the compression/expansion model for the warping, COW is not capable of correcting the spectra according to one signal: the alignment is always global and the best compromise is found given the segment length and the slack. This can be seen in Supplementary Fig. 3, which clearly shows that the results for this method are suboptimal. RSPA, on the other hand, performs well on the reference signal, but, similarly to COW, it also aligns other parts of the spectra in order to maximize the local correlation.

2.5. Computational aspects

Table 2 shows the results for the alignment of the three different datasets presented above using the optimal definition of the intervals for each dataset. As it can be seen, the time consumption for COW is as expected much larger than for FFT based methods. Particularly expensive was the search for the optimal warping parameters: around 6 min were necessary to complete the alignment for the Ni-bath data and another 37 were employed to find the optimal ℓ, and t using the simplex procedure. This is manifestly non-optimal, especially in view of the relatively poor alignment that was obtained. Computation times for COW and RSPA on dataset 3 are not meaningful (and therefore not reported) because both methods operate on a much larger problem (the shift is sought only on the reference signal in *i*coshift) and cannot handle this case correctly.

Table 2
Speed comparison among different algorithms on the alignment of the presented datasets. Times are in seconds.

	Wine	Nibath	Plasma
COW (opt)[a]	1.30 (38.3)	348 (2.23 × 10^3)	–
*i*coshift	0.089	1.9	0.067
RSPA	2.23	22.32	–

[a] The value in parentheses is the time consumption for the simplex search of the optimal ℓ and t.

Table 3
Calculation times of *i*coshift for random simulated datasets ($n_s = 100$). Times are in seconds.

No. samples	Sample length			
	4096	8192	16,384	32,768
50	0.58	1.10	2.06	4.01
100	0.68	1.24	2.32	4.47
200	0.83	1.44	2.74	5.54
400	1.10	1.98	3.89	8.75

The fact that RSPA requires higher computation times is both related to the sample-wise operation, without padding to $N^{\uparrow}$ (cf. Section 4.2.1), and to the peak picking and automatic interval definition. Obtaining the most efficient RSPA implementation was not the aim of this work and it is expected that zero padding to $N^{\uparrow}$ will improve the performance. Table 3 shows the computation times for a set of artificial sets of specific size, both in terms of number of samples and length of the signal (100 contiguous intervals were used). From the table it is apparent that *i*coshift can be used as a real time method even for the alignment of very large datasets.

3. Conclusions

The *i*coshift algorithm presented in this work provides a computationally efficient and versatile tool for the one-dimensional alignment of NMR signals in large spectral datasets. The speed of alignment algorithms have previously been a hindrance for their extensive application, but in all the tests made so far *i*coshift proved to be sufficiently fast to allow the NMR spectroscopist to work with full-resolution datasets almost in real time and to introduce automation improvements. Although only three input parameters are required for solving the majority of alignment problems, a number of optional meta-parameters can be provided to allow the algorithm to handle special misalignment problems (see Table 1). The speed of *i*coshift allows the user to swiftly and interactively investigate different operative combinations of interval settings for achieving the optimal result; the program is designed to be user friendly and to provide helpful and interactive plot facilities. Paradoxically, the whole idea behind spectral alignment is to render the large NMR datasets bilinear and thus suitable for subsequent multivariate chemometric models such as PCA, PLS and multivariate curve resolution (MCR) [3], but with the wine example, we have also shown that a perfect alignment of NMR spectra might remove part of the information sought. This advocates for using a customizable tool such as *i*coshift when aligning NMR spectra. While *i*coshift can be operated in a fully automated mode, it has been specifically designed for allowing user-defined intervals boundaries. Indeed, the time spent for adjusting the meta-parameters required for tuning a fully-automated interval definition step is often badly spent compared to the time required for the spectroscopist to manually select the intervals. However, the *i*coshift algorithm has also been designed to be used as an engine for fully automated interventions as it has been demonstrated on this work providing that part of programming code (as a separate routine) that applies a recently published method [18]. Although no direct comparisons were made, the *i*coshift algorithm appears to be faster than other recent methods as seconds are required where others report minutes for a smaller set on an equivalent computer [6,17,20]. In order for the software to be widely tested and used and to make further customizations and improvements possible, the Matlab source code is freely available for download at www.models.life.ku.dk/algorithms. Preliminary results have shown that *i*coshift performs equally well on chromatographic data, an investigation that will be pursued in a future publication.

4. Experimental

The primary aim of this study was to describe the new alignment algorithm and its performance on realistic NMR datasets. In the following a short description of each NMR dataset used in the performance test is provided. The second part of this section describes the detailed algorithmic aspects of *i*coshift (a stand-alone collection of algorithms) and provides comparisons with the previously published alignment algorithms highlighting advantages and drawbacks. Finally a comparison of the calculation speed is provided for all the investigated algorithms.

4.1. NMR datasets

4.1.1. Case 1: the wine data

The wine data include the NMR spectra of 40 table wines of different origin and color (red, white and rosé) and is taken from a previous study [9]. The NMR samples were prepared from 495 μl wine and 55 μl of TSP-d_4 (5.8 mM) in D_2O. ^1H NMR spectra were acquired on a Bruker Avance 400 spectrometer (Bruker Biospin GmbH, Rheinstetten, Germany), operating at 400.13 MHz Larmor's frequency for protons, equipped with a 5 mm BBI probe with *Z*-gradients. All experiments were performed at 298 K suppressing the water resonance by pre-saturation followed by a composite pulse. All samples were individually tuned, matched and shimmed and then acquired using a recycle delay of 5 s, 1536 scans and a dwell time of 60.4 ms for acquisition of 32 k data points, resulting in a total acquisition time of 1.979 s. Before Fourier transformation, the FIDs (Free Induction Decay) were zero-filled to 64 k points and apodized by Lorentzian line-broadening of 0.3 Hz. All spectra were individually baseline- and phase corrected using XWin-NMR (Bruker Biospin). The wine dataset has dimensions (samples × variables): 40 × 8712.

Chemical analyses were also performed on the same wines in order to determine, among the others, their content of ethanol and lactic acid. These data have been used in the present study for calculating PLS regression models. The table wine dataset, along with information concerning the *i*coshift intervals and the Y reference chemical values used for PLS calibrations, is available for download at http://www.models.life.ku.dk/research/data/Wine_NMR/index.asp.

4.1.2. Case 2: the nickel-bath data

The nickel-bath dataset is part of a study aimed at following the evolution of organic additives in an electroplating nickel bath (unpublished data). The samples analyzed were aqueous solutions containing approximately 10% of D_2O. ^1H NMR spectra were acquired on a Bruker Avance 500 spectrometer DRX-500 operating at a Larmor frequency of 500.13 MHz for ^1H (11.75 T). All experiments were performed at 298 K suppressing the water resonance by using the standard Watergate pulse sequence (Bruker Biospin). All samples were individually tuned, matched and shimmed. A spectral window of 15.646 ppm was acquired in 2.2 s. The relaxation delay was 9 s. The FIDs were collected into 32 k complex data points and 128 scans were accumulated for each spectrum. Total acquisition time for each spectrum was about 25 min. Zero-filling to 64 k points and a 0.3 Hz line-broadening multiplication was performed prior to Fourier transformation. The spectra were baseline- and phase corrected using MestRe-C 4.9.8.0 (www.mestrec.com, Mestreab Research SL, Santiago de Compostela, Spain). TSP was used as a chemical-shift reference ($\delta = 0$ ppm) and succinic acid was used as internal standard. In total 460 samples were analyzed and their NMR spectral region between 9.5 and −0.5 ppm was used for the present study. The Ni-bath dataset has dimensions (samples × variables): 460 × 40,905.

4.1.3. Case 3: the rat plasma data

This metabonomic dataset is from the study by Kristensen et al. [23] in which a combination of NMR analysis and chemometrics was used as a method for rapidly assessing the lipoprotein subfractions in rat plasma [24]. In that study 100 μl of plasma were transferred to a 5 mm NMR tube and 450 μl D_2O were added. ^1H NMR spectra were then acquired on a Bruker Avance 400 spectrometer (9.4 T) operating at 400.13 MHz. All samples were individually tuned, matched and shimmed and measured using a broad band inverse detection probe equipped with *Z*-gradients designed to 5 mm NMR tubes. Since the average body temperature of rats is 311 K, all experiments were performed at that temperature. Water pre-saturation was employed during the recycle period by a composite 90° pulse in order to suppress the intense water resonance. A spectral window of 8278.15 Hz was acquired in 1.97 s. The relaxation delay was 20 μs. The FIDs were collected into 32 k complex data points and 128 scans were accumulated for each sample. Total acquisition time was 15 min and prior to Fourier transformation the data were zero-filled to 64 k points and multiplied by a 0.3 Hz line-broadening function. The spectra were baseline- and phase corrected manually using Topspin™ (Bruker Biospin). A batch of this study, consisting of 72 plasma samples of rats fed with different diets, was used in the present study. The rat plasma data has the dimensions (samples × variables): 72 × 21,727.

4.1.4. Human urine data

The human urine dataset is the target of a focused metabonomic study in which *i*coshift is applied. This study will be the subject of a forthcoming paper. The challenging region (because of the strong misalignment) between 2.74 and 2.64 ppm, containing an NMR doublet of citric acid together with a singlet of TMA (TriMethylAmine), is shown in Fig. 1 of the present study, demonstrating the capability of *i*coshift of solving intricate misalignments of signals belonging to different molecules.

^{1}H NMR spectra of buffered urine samples were acquired on a Bruker Avance Ultra Shield 500 spectrometer (Bruker Biospin) operating at 500.13 MHz (11.75 T) using a broadband inverse detection probe head equipped with a 120 µl flow-cell. Data were accumulated at 298 K employing a pulse sequence composed by a pre-saturation of the water resonance during the recycle period followed by a composite 90° pulse with an acquisition time of 1.57 s, a recycle delay of 5 s, 128 scans and a sweep width of 10416.67 Hz, resulting in 16 k complex data points. All samples were individually and automatically tuned, matched and shimmed. Prior to Fourier transformation, each FID was zero-filled to 64 k points and apodized by Lorentzian line-broadening of 0.30 Hz. The resulting spectra were manually phased and automatically baseline corrected using Topspin™ (Bruker Biospin) and the ppm scale was referenced towards the TSP peak at 0.00 ppm. Since the composition and the concentration were very similar for every sample, the receiver gain was initially set at a fixed value equal to 287.4 for all the experiments. The human urine dataset has the dimensions (samples × variables): 98 × 30,000.

4.2. Signal processing

The flow chart of the algorithm is shown in Fig. 2. In this section, the three parts of the *i*coshift algorithm: interval definition (cf. Section 4.2.2), FFT engine (cf. Section 4.2.1) and reconstruction (cf. Section 4.2.4) are described. Other alignment methods related to *i*coshift and used for comparison are described in the subsequent sections (viz., PAFFT in Section 4.2.5, RSPA in Section 4.2.6 and COW in Section 4.2.7).

4.2.1. Cross-correlation by FFT

The theory of the calculation of the cross-correlation coefficient using the Fourier Transform is well known and will only be outlined here [25]. Given two continuous functions $X(t)$ and $Y(t)$, the cross-correlation for a lag u between them can be defined as in Eq. (1)

$$C_Y^X(u) = \int_{-\infty}^{\infty} Y(t+u)X(t)dt \tag{1}$$

$$x(f) = \mathscr{F}(X(t)) = \int_{-\infty}^{\infty} X(t)e^{2\pi ift}dt \tag{2a}$$

$$X(t) = \mathscr{F}^{-1}(x(f)) = \int_{-\infty}^{\infty} x(f)e^{-2\pi ift}df \tag{2b}$$

where t and f are generally referred to as time and frequency and $x(t)$ and $X(f)$ are the representations of the same process in the time and in the frequency domain, respectively. However for this specific application and representation of the Fourier transform, t denotes the chemical shift (i.e., a frequency shift expressed in ppm). Correspondingly, f is another variable that can be expressed in ppm^{-1}, and therefore in some time measurement unit [25], but $X(f)$ is not the time domain representation of $x(t)$ as generally intended by the NMR community.

If $C_Y^X(u)$ has a maximum at $\tilde{u}$ in the domain $\mathrm{W} = [-w, w]$, one can define a function $\widetilde{Y}(t) = Y(t+\tilde{u})$ that is shifted along t and has maximum cross-correlation to $X(t)$ at $t = 0$. $\widetilde{Y}(t)$ is said to be the aligned signal, $X(t)$ the alignment target, $\tilde{u}$ the optimal shift and w is the width of the search space for the optimal correction.

The Fourier transform ($\mathscr{F}$ – Eq. (2a)) and its inverse ($\mathscr{F}^{-1}$ – Eq. (2b)) allow the calculation of all the cross-correlation values for arbitrarily large values of w using the fact that $C_Y^X(u)$ can be expressed as a function of $\mathscr{F}^{-1}(x(f)y(f))$ [25]. This holds also for discrete samples (**x** and **y**, respectively) of finite length N of the two functions, so long as $X(t)$ and $Y(t)$ are periodic and that their period is of length N; that is, as long as they are completely determined by **x** and **y**. However, this is not the case for sections of NMR spectra and, in order to avoid contamination from the end sections, it is necessary to pad **x** and **y** with a number of zeros equal to the largest correction allowed w [25]. The FFT algorithm makes it particularly efficient to calculate the $C_Y^X(u)$ in the discrete case and is therefore often used for the signal alignment [17,18].

Several algorithmic techniques can be used to improve the efficiency of the FFT algorithm. In particular, a significant gain is achieved by aligning several samples at once rather than one at a time (Fig. 4 in the Supplementary Material). Note that this is currently possible in *i*coshift only if the boundaries of the intervals are common between spectra and that, for very large N and number of intervals, it may be too costly memory-wise to treat all the samples at once.

It is also well known that the FFT algorithm is faster when N is equal to a power of two [25]. Therefore, in some cases, it is beneficial to pad the samples with zeros to a length of $N^{\uparrow} = 2^{\lceil \log_2 N \rceil}$ points, where $\lceil x \rceil$ indicates the smallest integer larger than x. However, some numerical experiments showed that, while this is almost always true (in the Matlab environment) when samples are treated individually, it is no longer valid when samples are treated in blocks (cf. Supplementary Material Figs. 4 and 5). In particular, in the latter case the advantage in computation time appears to be appreciable (>5%) only when the length of this padding is limited compared to N or when N is small (below 64 points, Fig. 5 in the Supplementary Material). Because of the difficulty in predicting when zero padding to the next power of two is beneficial in the block case, this feature is currently not implemented in *i*coshift.

4.2.2. Interval definition and recursion

The FFT engine can operate on the entire spectrum as well as on intervals of arbitrary length in sequence. Depending on how the intervals are defined, there are essentially three options available: to align the entire spectrum, to align separate, non-overlapping intervals independently, or to align the entire spectrum based on a reference interval. The first and the last options are indeed trivial as it is sufficient to shift the signal by $\tilde{u}$ points, where $\tilde{u}$ is calculated on the entire spectrum or on the reference interval, respectively. The custom interval case is slightly more complex and, differently from other FFT-based alignment algorithms, adjacent intervals can share the common boundary but need not be contiguous. Moreover, if some regions in the samples are not included in any interval, they are left untouched (i.e., $\tilde{u} = 0$ for them).

When automated options for interval definition are used and either the number of intervals, n_s, or their length ℓ are fixed by the user, adjacent intervals share the common boundary, similarly to what is done with COW [20]. However, the treatment of the remainders is different in the two cases: if n_s is fixed, ℓ is equal to $c = \lfloor Nn_s^{-1} \rfloor$ (i.e., the largest integer smaller than Nn_s^{-1}) for the last $c(\ell+1) - N$ intervals and $c+1$ for the first $N - n_s c$ (that is, the remainders are equally split between intervals); on the contrary, if ℓ is fixed, the remainders are attached to the last interval, which has length $\ell + N - n_s(\ell - 1)$, where n_s is $\lfloor (N-1)(\ell-1)^{-1} \rfloor$.

In order to keep the *i*coshift algorithm as general and versatile as possible, no elaborate segmentation routine based on, e.g., peak picking was implemented. This also implies that, when automated interval definition is employed, there is no safeguard against having a boundary occurring in the middle of a peak.

*i*coshift implements a basic (optional) recursion for the customized interval case; namely, it is possible to perform a full-spectrum correction before the single intervals are aligned. This has proven to provide an improved alignment performance in certain cases at a limited additional computational cost.

4.2.3. Data interpolation using missing values

The *i*coshift algorithm is capable of handling missing values that occur at the beginning or at the end of the discrete intervals of $X(t)$ and $Y(t)$. This is possible because the length of $\mathbf{x}$ and $\mathbf{y}$ need not be the same. In particular, let $N_{\mathbf{x}} + w$ and $N_{\mathbf{y}} + w$ be the lengths, after the zero padding to avoid end section contamination, of $\mathbf{x}$ and $\mathbf{y}$, respectively, and, without lack of generality, $N_{\mathbf{x}} > N_{\mathbf{y}}$; then, in order to calculate $C^{\mathbf{x}}_{\mathbf{y}}(u)$ in W, it is sufficient to pad $\mathbf{y}$ with $N_{\mathbf{x}} - N_{\mathbf{y}}$ zeros [25]. Thus, missing values are simply treated by removing them, zero padding the shortest interval to the same length as the longest, and by correcting the resulting optimal shift with the factor $\Delta = m_{\mathbf{x}} - m_{\mathbf{y}}$ (where $m_{\mathbf{x}}$ and $m_{\mathbf{y}}$ is the number of missing values in the leading sections of $\mathbf{x}$ and $\mathbf{y}$, respectively). Namely, if $\tilde{u}_{\text{miss}}$ is the optimal shift after the removal of the missing values and the padding to the same length, the optimal shift for the original intervals is $\tilde{u} = \tilde{u}_{\text{miss}} + \Delta$. The number of trailing missing values is irrelevant in this respect. Fig. 9 illustrates how the use of missing values can improve the reliability of the aligned spectra avoiding the introduction of artifacts that can mislead the successive data analysis.

4.2.4. Reconstruction

Similarly to the other FFT-based algorithms described herein, an insertion/deletion model is used by *i*coshift; therefore, the warping path (i.e., the function relating the chemical shift axis in the sample and target spectra) obtained by *i*coshift is piece-wise linear, made up of segments of slope one that are connected by vertical, horizontal or empty sections (cf. Fig. 6 in the Supplementary Material). In particular, having the chemical shift axes for the target and for the sample ($\mathbf{x}$ and $\mathbf{y}$, respectively) on the abscissae and the ordinates, respectively, an horizontal segment corresponds to replicating the boundary point or to inserting missing values (*insertion*), whereas a vertical one entails a *deletion*. The reconstruction of the samples is performed independently for each spectrum and interval, but no global optimality criterion is used. Hence, unlike COW, but similarly to the other FFT-based alignment algorithms, there is no measure of the goodness of the fit after the deletion/insertion occurs. This is clearly suboptimal but works well under the assumption that the boundaries are located in regions that do not contain any relevant signal. Algorithmic techniques such as dynamic programming, breadth first search, beam search or genetic algorithms have been suggested for alignment algorithms that do not use FFT and cross-correlation. The best options for such global optimization are currently being investigated.

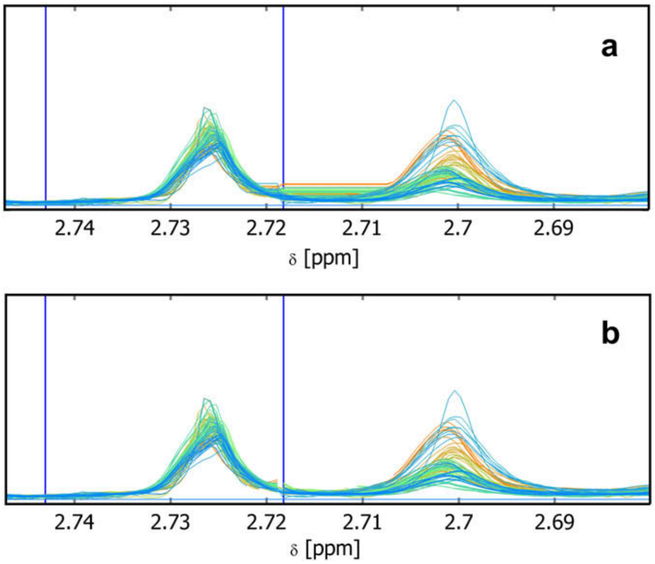

Fig. 9. Effect of the imputed value on the quality of the aligned signals. (a) Imputation of first/last boundary point in each interval, (b) imputation of missing values.

4.2.5. Peak Alignment by FFT (PAFFT) and Recursive Peak Alignment by FFT (RAFFT)

The PAFFT algorithm corresponds to the *i*coshift with automatic segmentation into intervals of equal length and insertion using the first/last point in each interval. From the computational point of view, PAFFT treats the samples separately (one at a time) and implements padding to the nearest power of two for each interval. However, compared to *i*coshift, there is no explicit zero padding to avoid end section contamination; in practice, in the current implementation (http://physchem.ox.ac.uk/~jwong/specalign/data/PAFFT.m, download: 18th August, 2009), the contamination is avoided only if $N^{\uparrow} - N \geqslant w$, but this need not always be the case.

The RAFFT algorithm recursively calls PAFFT with intervals of decreasing length until some lower limit for N is reached or the similarity does not improve. This recursion step is not explicitly implemented in *i*coshift, which only allows the shift of the full spectrum as a preliminary step. RAFFT suffers from the same drawbacks described for PAFFT.

4.2.6. Recursive Segment-wise Peak Alignment (RSPA)

The RSPA method, like PAFFT and RAFFT operates on one sample at a time and includes an automated interval definition part and a recursion routine. The automated segmentation is based on peak picking and on some user-defined constants that depend on the dataset (e.g., the noise level, the minimum distance between peaks that are merged to the same interval, the height of peaks that are considered intense within the set of all the detected peaks). The principals of the algorithm are only outlined in this section, as more details are available in the original publication [18]; more attention will be given to describing the choices made for the implementation as the source code of RSPA is currently not publicly available.

As a first step a full spectrum shift is performed. Secondly, the interval boundaries are defined based on peak positions and on the positions of the corresponding bracketing minima. These are found as the points where the numerical derivative changes sign (i.e., no smoothing is used) of the signal. Subsequently, the height of each peak is determined as the difference between the value at the maxima and the smallest value at the bracketing minima and initial estimates of the intervals are taken as the bracketing minima of intense peaks. The result, at this stage is an ordered set of intervals $\mathscr{A}_m = \{[a_1, b_1], \ldots, [a_{i_m}, b_{i_m}]\}$ and heights $\mathscr{H}_m = \{h_1, \ldots, h_{n_m}\}$ for each sample m and the target. Intervals relative to intense peaks (here determined as those whose height exceeds the 30th percentile of $\mathscr{H}_m$) that are closer than a predefined value s are merged; that is, if $a_{i+1} - b_i \leqslant s$, then a new interval $[a_i, b_{i+1}]$ is formed in substitution of the two original ones.

In a second stage, the initial estimates for the intervals for each sample are compared to those for the target and vice versa. Intervals in a spectrum are considered overlapping with a segment in the target, if their mean falls within the central part of the target's interval (that is if $0.5(a_{k,\text{sam}} + b_{k,\text{sam}}) \in [0.75a_{\text{tar}} + 0.25b_{\text{tar}}, 0.25a_{\text{tar}} + 0.75b_{\text{tar}}]$) and if overlap occurs for more than one k, then they are merged and a new segment is formed for the sample as $[a_{\min(k),\text{sam}}, b_{\max(k),\text{sam}}]$. The same procedure is then repeated for the target to merge target intervals matched to the same sample interval. Finally, the boundaries of the segments in the sample and the target are compared and matched (that is, if overlapping intervals in target and sample are $[a_{\text{tar}}, b_{\text{tar}}]$ and $[a_{\text{sam}}, b_{\text{sam}}]$, respectively, the final boundaries are $[\min(a_{\text{sam}}, a_{\text{ref}}), \max(b_{\text{sam}}, b_{\text{ref}})]$). This expansion of the intervals may lead to some overlap between adjacent intervals within either the sample or the target, which is eliminated by merging the overlapping intervals. In the optimal case, the result of this procedure is a set of matched intervals containing relevant peaks and a series of intervening regions with unmatched peaks or noise. In the simulation of the RSPA based on *i*coshift, the intervening regions are not aligned as there would be no information useful to establish the correctness of the alignment.

The recursion part includes a function that evaluates the goodness of the alignment of each interval in terms of the piece-wise correlation coefficient $\rho_t(\mathbf{x}_i, \mathbf{y}_i)$ (i.e., the normalized sum of the Pearson's correlation coefficients between segments of length t for the ith interval). If this value is lower than a predefined threshold, the interval is split. The splitting is done at the minimum in the spectrum being aligned closer to the minimum of the vector $\mathbf{x}_i * \mathbf{y}_i$, where $*$ denotes the element-wise product. If a segment resulting from the split is smaller than the minimum length allowed (taken as equal to s, i.e., the minimum distance between merged peaks), it is disregarded. The segments resulting from the splitting are subject to further (separate) alignment in the following recursion step. In the implementation of the RSPA used herein, an additional constraint is applied to the number of recursions, as it was observed that artifacts could emerge if too many recursions are applied to markedly different intervals that are sufficiently long to allow several splitting without reaching the lower limit for the segment length.

4.2.7. Correlation optimized warping (COW)

The COW algorithm has been extensively used for the alignment of chromatographic data and has been tested also on NMR signals [19,20,7]. The methods also finds piece-wise linear warping paths, but, rather than using a insertion/deletion model it uses a compression/expansion model. That is, the warping paths are formed by segments whose slope is allowed to take only a limited number of values determined by the length of the corresponding segments (ℓ) in the sample, and by the so-called slack parameter t. The COW algorithm aligns the samples with a target by dividing them in an equal number of segments, whose boundaries (a_i for $i = 1, \ldots, n_s$), in the target, are allowed to take all the integer values within predefined intervals (namely $a_{i+1} - a_i + 1 \in [\ell - t, \ell + t]$, where t is the so-called slack parameter). The optimal warping is the one that maximizes the sum of the Pearsons correlation coefficient for all the intervals after the segments in the target are interpolated to the same length as the corresponding interval in the sample (typically ℓ). In particular, dynamic programming is used to ensure that the global maximum, given the local constraints, is attained. The COW algorithm works spectrum-wise, but from the algorithmic point of view, it is possible to treat the spectra in blocks with considerable reduction in computational complexity [20].

4.3. Computation tests

All the calculations have been performed on a personal computer equipped with i7 core, 965 chipset, 3.2 GHz, 12 GB of memory, Windows XP-64 SP3 and Matlab® 7.7 (2008b, The Mathworks Inc., Natick, MA, USA). In-house optimized Matlab routines derived from the code available at www.models.life.ku.dk have been used. The optimal segment length and slack parameter for COW have been obtained using the simplex optimization routine by Skov et al. [20]. In the corrected version of COW, the calculations on the nodes in the initial grid search are computed in parallel on the 8 available cores on the machine. This allowed a ~50% reduction in computation time compared to the non-parallel version. The RSPA routine (cf. Section 4.2.6) for Matlab has been implemented in-house using *i*coshift as a computation engine. The functions for peak picking, automatic segmentation and recursion are optimized to the largest possible extent and will be available for download at www.models.life.ku.dk.

The synthetic datasets (Table 3) are made of uniformly distributed random numbers as the computational complexity of *i*coshift is only dependent on the length of the intervals, on their number n_s, on the maximum allowed correction w, and on the number of samples M (that is, $O(M\sum_i^{n_s}(N_i + w)\log_2(N_i + w))$, where $N_i + w$ is the length of the ith interval zero padded to handle the end section contamination). Automated segmentation to 100 segments was used setting the maximum allowed correction as the minimum between 100 and half the interval length.

Acknowledgments

Flemming H. Larsen, Maider Vidal, Mette Kristensen are acknowledged for providing the datasets used for assessing the algorithm's features. The Danish Ministry of Food, Agriculture and Fisheries is acknowledged for sponsoring the project "Urinary biomarkers for eating patterns using NMR spectroscopy and Chemometrics" (3304-FVFP-060706-01) under which the present research has been developed. The Villum Kann Rasmussen project (Metabonomic Cancer Diagnostic) and The EU FP6 ToK project "Food informatics" (EC Contract No: MTKI-CT-2005-030056) and gratefully acknowledged for support to Francesco Savorani and. Giorgio Tomasi. Eventually, Rasmus Bro, Frans van den Berg and Lars Nørgaard are acknowledged for valuable scientific discussions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmr.2009.11.012.

References

[1] D. Johnels, U. Edlund, E. Johansson, S. Wold, A multivariate method for carbon-13 NMR chemical shift predictions using partial least-squares data analysis, J. Magn. Reson. 55 (1983) 316–321.

[2] K.P. Gartland, C.R. Beddell, J.C. Lindon, J.K. Nicholson, Application of pattern recognition methods to the analysis and classification of toxicological data derived from proton nuclear magnetic resonance spectroscopy of urine, Mol. Pharmacol. 39 (1991) 629–642.

[3] H. Winning, F.H. Larsen, R. Bro, S.B. Engelsen, Quantitative analysis of NMR spectra with chemometrics, J. Magn. Reson. 190 (2008) 26–32.

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

[5] 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, Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples, J. Pharm. Biomed. Anal. 12 (1994) 1215–1225.

[6] N.P.V. Nielsen, J.M. Carstensen, J. Smedsgaard, Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping, J. Chromatogr. A 805 (1998) 17–35.

[7] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemom. 18 (2004) 231–241.

[8] D. Bylund, R. Danielsson, G. Malmquist, 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 961 (2002) 237–244.

[9] F.H. Larsen, F. van den Berg, S.B. Engelsen, An exploratory chemometric study of H-1 NMR spectra of table wines, J. Chemom. 20 (2006) 198–208.

[10] R.J.O. Torgrip, M. Aberg, B. Karlberg, S.P. Jacobsson, Peak alignment using reduced set mapping, J. Chemom. 17 (2003) 573–582.

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

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

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

[14] G.C. Lee, D.L. Woodruff, Beam search for peak alignment of NMR signals, Anal. Chim. Acta 513 (2004) 413–416.

[15] J.T.W.E. Vogels, A.C. Tas, J. Venekamp, J. Van Der Greef, Partial linear fit: a new NMR spectroscopy preprocessing tool for pattern recognition applications, J. Chemom. 10 (1996) 425–438.

[16] R. Stoyanova, A.W. Nicholls, J.K. Nicholson, J.C. Lindon, T.R. Brown, Automatic alignment of individual peaks in large high-resolution spectral data sets, J. Magn. Reson. 170 (2004) 329–335.

[17] J.W.H. Wong, C. Durante, H.M. Cartwright, Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets, Anal. Chem. 77 (2005) 5655–5661.

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

[19] F. van den Berg, G. Tomasi, N. Viereck, Warping: investigation of NMR preprocessing and correction, in: S.B. Engelsen, P.S. Belton, H.J. Jakobsen (Eds.), Magnetic Resonance in Food Science: The Multivariate Challenge, Royal Society of Chemistry, Cambridge, 2005, pp. 131–138.

[20] T. Skov, F. van den Berg, G. Tomasi, R. Bro, Automated alignment of chromatographic data, J. Chemom. 20 (2006) 484–497.

[21] L. Norgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. 54 (2000) 413–419.

[22] J.T.M. Pearce, T.J. Athersuch, T.M.D. Ebbels, J.C. Lindon, J.K. Nicholson, H.C. Keun, Robust algorithms for automated chemical shift calibration of 1D H-1 NMR spectra of blood serum, Anal. Chem. 80 (2008) 7158–7162.

[23] M. Kristensen, F. Savorani, G. Ravn-Haren, M. Poulsen, J. Markowski, F.H. Larsen, L. Dragsted, S.B. Engelsen, NMR and interval PLS as reliable methods for determination of cholesterol in rodent lipoprotein fractions, Metabolomics (2009), in press, doi:10.1007/s11306-009-0181-3.

[24] M. Petersen, M. Dyrby, S. Toubro, S.B. Engelsen, L. Norgaard, H.T. Pedersen, J. Dyerberg, Quantification of lipoprotein subclasses by proton nuclear magnetic resonance-based partial least-squares regression models, Clin. Chem. 51 (2005) 1457–1461.

[25] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C, Cambridge University Press, Cambridge, UK, 2002.