Published in final edited form as:
Anal Chem. 2012 June 19; 84(12): 5372–5379. doi:10.1021/ac301327k.

Variable Reference Alignment: an improved peak alignment protocol for NMR spectral data with large inter-sample variation

Neil MacKinnon[1,2], **Wencheng Ge**[2], **Amjad P. Khan**[3], **Bagganahalli S. Somashekar**[1,2], **Pratima Tripathi**[1,2], **Javed Siddiqui**[3], **John T. Wei**[4,5], **Arul M. Chinnaiyan**[3,4,5,6], **Thekkelnaycke M. Rajendiran**[3], and **Ayyalusamy Ramamoorthy**[1,2,*]

[1]Biophysics, University of Michigan, Ann Arbor, MI, USA

[2]Department of Chemistry, University of Michigan, Ann Arbor, MI, USA

[3]Michigan Center for Translational Pathology, Department of Pathology, University of Michigan, Ann Arbor, MI, USA

[4]Department of Urology, University of Michigan Medical School, Ann Arbor, MI, USA

[5]Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI, USA

[6]Howard Hughes Medical Institute, University of Michigan Medical School, Ann Arbor, MI, USA

Abstract

In an effort to address the variable correspondence problem across large sample cohorts common in metabolomic/metabonomic studies, we have developed a pre-alignment protocol that aims to generate spectral segments sharing a common target spectrum. Under the assumption that a single reference spectrum will not correctly represent all spectra of a data set, the goal of this approach is to perform local alignment corrections on spectral regions which share a common 'most similar' spectrum. A natural beneficial outcome of this procedure is the automatic definition of spectral segments, a feature that is not common to all alignment methods. This protocol is shown to specifically improve the quality of alignment in ^{1}H NMR data sets exhibiting large inter-sample compositional variation (e.g. pH, ionic strength). As a proof-of-principle demonstration, we have utilized two recently developed alignment algorithms specific to NMR data, recursive segment-wise peak alignment and interval correlated shifting and applied them to two data sets comprised of 15 aqueous cell line extract and 20 human urine ^{1}H NMR profiles. Application of this protocol represents a fundamental shift from current alignment methodologies that seek to correct misalignments utilizing a single representative spectrum, with the added benefit that it can be appended to any alignment algorithm.

Keywords

Metabolomic; alignment; NMR; urine

[*]Corresponding Author: Prof. Ayyalusamy Ramamoorthy, Biophysics and Department of Chemistry, University of Michigan, 930 North University Avenue, Ann Arbor, Michigan 48109-1055, USA, Phone: (734) 647-6572, Fax: (734) 764-3323, ramamoor@umich.edu.

Supporting Information Available: Selected method details, scree plots of PCA models generated, cell line plots, PC 2 loadings, and alignment assessment metric comparisons after constant segment size icoShift alignment. This material is available free of charge via the Internet at http://pubs.acs.org

Introduction

Studies involving the metabolome have become invaluable, spanning a broad spectrum of research fields including (but not limited to) food quality testing[1,2], environmental monitoring[3], and human disease detection[4,5]. Commonly, metabolomic/metabonomic investigations will employ either chromatography-coupled mass spectrometry (MS) or nuclear magnetic resonance (NMR) as the analytical technique. A problematic characteristic common to these techniques is signal positional fluctuation across samples. For example, while NMR is a robust, highly reproducible analytical technique, recovery of relevant biological information is often complicated by resonance variation across the samples comprising the study cohort.

The issue of inter-sample peak variation due to sample conditions (e.g. pH, ionic strength) is a problem common to NMR spectra. Chemical shift variation between samples is particularly important when multivariate statistical analyses (e.g. PCA, PLS-DA, OPLS-DA[6], STOCSY[7]) are performed. The fundamental assumption of a one-toone correspondence between the variable (chemical shift) and peak intensity across all samples fails, resulting in a reduction of important biological information recovery. One of the first attempts to correct for small chemical shift variations was to create “bins” of data, integrating intensities within a defined chemical shift range (e.g. 0.04 ppm)[8]. While spectral binning was certainly an important development, the resulting loss of spectral resolution concomitant with the emphasis of large intensity signals at the cost of (potentially important) low intensity signals required the development of sophisticated alignment methods amenable to high resolution spectra with large dynamic ranges.

Recently, there have been numerous procedures proposed that attempt to correct the misalignment problem in diverse data sets obtained from near-infrared, NMR, and mass spectroscopies. Early procedures, including dynamic time warping[9], covariance optimized warping[10], partial linear fit[11], and peak alignment by a genetic algorithm[12] have led to families of alignment methodologies[13–15]. Generally, the differences among the procedures can be described in terms of the number of user-defined parameters required, the expected introduction of signal distortion, and the computational requirements[16–20]. Since the alignment of a full resolution data set is desirable, algorithms exploiting the fast Fourier transform (FFT) were developed to ease the computational burden. While initially developed for MS data, both the peak alignment by FFT and recursive alignment by FFT methods are easily adapted to NMR spectra[21,22].

Specific to NMR spectra, recursive segment-wise peak alignment (RSPA)[23] and interval correlated shifting (icoShift)[24] utilize the efficient FFT engine to handle the large data sets typical of NMR-based metabolomics/metabonomics studies. RSPA identifies the ‘most-similar’ spectrum of a data set (reference spectrum) and aims to align the remaining spectra by first automatically calculating spectral segments based on a peak-picking routine, designed to confine multiplet signals to a common segment. Test segments are shifted at their boundaries to maximize correlation with the reference spectrum, with boundaries linearly interpolated. Each segment is then recursively divided to correct misalignment of signals not necessarily a member of a multiplet signal. The icoShift method also relies on maximizing the correlation between target and test spectra on a segment-wise basis, however it does not rely on peak-picking since segment boundaries are user controlled.

Two requirements common to most recent alignment algorithms are the identification of a suitable reference spectrum and the definition of suitable spectral regions (segments) where local alignment occurs. Before the development of RSPA, the choice of reference spectrum was somewhat arbitrary, often resulting in multiple target spectra tested until satisfactory alignment was achieved. Similarly, choice of segment boundaries is often only dictated by the total number of segments desired, potentially resulting in a boundary dividing a resonance leading to significant peak distortion. Alternatively, segment boundaries may either be explicitly defined as in the case of icoShift, requiring careful examination of the data set to ensure suitable limits are selected, or segment boundaries may be automatically calculated as in RSPA.

This report details a pre-segmentation procedure of spectral data sets based on satisfaction of a segment-wise ‘closeness’ condition, resulting in the simultaneous automatic definition of segment boundaries and identification of spectral regions sharing a common reference spectrum. The pre-selection of global segments is shown to improve alignment utilizing both the icoshift and RSPA alignment algorithms. In the case of icoshift, the improvement is a result of an automated selection of segments with non-constant length, while in the case of RSPA the improvement results from alignment towards a segment-specific reference spectrum. The modified alignment algorithms are tested on two data sets chosen to represent typical data exhibiting relatively small (aqueous extracts from cell lines) and large (human urine) inter-sample variation. We anticipate that this additional calculation will be valuable in extracting additional biologically important spectral information.

Materials and Methods

Materials

2,2-dimethyl-2-silapentane-5-sulfonate-d_6 (DSS), 3-(Trimethylsilyl)propionic-2,2,3,3-d_4 acid sodium salt (TSP), D_2O, potassium phosphate (monobasic, dibasic), methanol, and chloroform were obtained from Sigma/Aldrich (Milwaukee, USA) and used as received. All water was of MilliQ purity.

Prostate Cell Line Growth and Harvesting

Five prostate cell lines were cultured in triplicate and 2×10^6 cells of each cell line were stored at −80 °C until used for analysis. Details of the growth conditions are available in the supporting information.

Urine

A total of 20 urine samples were collected according to protocol approved by the Institutional Review Board. Briefly, urine samples were collected in preservative-free urine collection cups (minimum 10 mL) and centrifuged at 4000 rpm for 15 minutes at 4 °C. The resulting supernatants were carefully separated from the pellets and stored separately at −80 °C until used for analysis.

Sample Preparation and NMR Spectroscopy

Aqueous metabolites were extracted from cell lines as described previously[25]. Briefly, the cell pellet was extracted with a 1/1/1 v/v/v of ice-cold methanol, chloroform, and water. The aqueous methanolic layer was collected and dried under vacuum before storing at −80 °C until used. The dried samples were brought to room temperature and re-dissolved in 400 μL phosphate buffer (100 mM, pH 7.2 prepared with D_2O) containing 1 mM DSS prior to NMR measurement.

Urine samples were thawed and 450 μL of each sample was added to 50 μL of 1 M phosphate buffer (pH 7.2) containing TSP ($[TSP]_f = 3$ mM), prepared in D_2O. The sample was subsequently centrifuged to remove residual insoluble material and then transferred to a 5 mm NMR tube for measurement.

A Bruker AVANCE™ 900 MHz (Bruker Biospin, Germany) spectrometer operating at 889.79 MHz equipped with a TCI cryoprobe was used. A single-pulse experiment including WATERGATE water suppression was utilized. For each cell line spectrum, 1024 transients each containing 32K data points and a spectral width of 16 ppm were collected, using a 30° flip angle[26] and a 2 s recycle delay. For each urine spectrum, 64 transients each containing 32K data points and a spectral width of 16 ppm were collected, using a 30° flip angle and a 2 s recycle delay[27]. NMR experiments were performed on a separate set of 20 urine samples using a Bruker Avance 500 MHz spectrometer operating at proton resonance frequency of 500.13 MHz equipped with a 5 mm TXI SB probe. For these experiments, a single-pulse of 30° flip angle with water presaturation and a 3 s recycle delay was used. Each spectrum with a 15 ppm spectral width was obtained by co-adding 128 transients each containing 32K data points.

All measurements were performed at 298 K. Offline spectral processing was done using ACD/NMR Processor 12.01 (ACD/Labs, Toronto, Canada). The FID were zero-filled to 64K, subjected to exponential multiplication equivalent to 0.3 Hz line broadening prior to Fourier transformation, and referenced to DSS/TSP at 0 ppm.

Data Preparation

All spectra were phase and baseline corrected in ACD/NMR Processor prior to export as an ASCII file. All further processing was performed in Matlab R2010b (Mathworks, Natick, MA). Cell line and urine spectral data sets were normalized using the Probabilistic Quotient Normalization (PQN) procedure, choosing the respective average spectrum of each data set as the reference[28]. Signal intensity in the water region (4.5 – 5.0 ppm) and water + urea region (4.5 – 6.0 ppm) was set to 0 in the spectra of all cell line and urine samples, respectively. The alignment algorithm based on the RSPA alignment method[23] was written in-house (Matlab), while the publicly available icoShift Matlab function was utilized[24,29].

Spectral Alignment

RSPA was implemented based on the report of Veselkov et al.[23], and is briefly summarized in the supporting information. For our purposes, reference spectra were calculated according to a closeness index taking $\alpha = 0.02$ ppm (Eq. S-1, Supporting Information). Signals were identified through the zero-crossing of the first order derivative calculated utilizing a 3rd order polynomial according to the Savitsky-Golay filtering method[30,31]. Peaks were grouped such that the inter-peak distance was less than 20 Hz, and the maximum shift threshold was set to 25 Hz. Test segment alignment was accepted subject to an improvement in the alignment quality parameter (Eq. 1, $\alpha = 0.05$ ppm).

The icoShift algorithm was used as available[24,29], with a brief description available in the supporting information. In this work, the segment boundaries were set automatically, the maximum shift threshold was set to 25 Hz, and shifted boundaries were filled with the first/last point of the segment.

In both alignment algorithms there is a traditional requirement that all test spectra be aligned towards a single reference spectrum. We propose a slight modification to the protocol whereby the restriction of a single reference spectrum is relaxed. A pre-alignment module was written such that global spectral segments sharing a common reference spectrum, as calculated according to Eq. S-1, are identified. The global segments are generated by incremental growth of a test segment followed by calculation of the reference spectrum, repeated until a different reference spectrum is identified. The global segment boundaries are thus defined and incremental growth of a new segment occurs in an identical fashion. The global segments are then individually subjected to either the RSPA or icoShift with alignment taking place towards the segment-specific reference spectrum. The fully aligned spectrum is finally reassembled after each global segment alignment is complete (see schematic, Fig. 1).

For demonstration purposes, four alignment procedures were performed. In the case of RSPA, alignment was performed without modification of the original algorithm (single-reference RSPA, sr-RSPA). RSPA was also performed after generation of the global segments, with alignment occurring towards the segment-specific reference spectrum (variable-reference RSPA, vr-RSPA). In vr-RSPA, there was no restriction on the algorithm creating local segments within the global segment. In the case of icoShift, the global segment boundaries defined by our module were used and alignment was performed towards a single reference identified utilizing Eq. S-1 (single-reference icoShift, sr-icoShift). Finally, both the global segment boundaries and segment-specific reference spectrum were input and is referred to as variable-reference icoShift (vr-icoShift).

Assessment of Alignment

Application of an alignment procedure is expected to increase the similarity amongst all spectra, and thus the correlation coefficient. The mean correlation coefficient was calculated for each data set before and after application of the alignment algorithms.

An alignment quality parameter[23] may be calculated for a data set containing *n* spectra according to Equation 1:

$$aq_\alpha = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{i-1} cc_\alpha(\mathbf{s_i}, \mathbf{s_j}) \qquad (1)$$

where cc_α ($\mathbf{s_i}, \mathbf{s_j}$) is the scaled correlation coefficient matrix of spectra s_i and s_j (with bin size α = 0.05 ppm for this work), and aq_α ranges from 0 (unaligned) to 1 (perfectly aligned). The alignment quality may be tested over the entire spectrum (aq_{global}) or within each segment sharing a common reference spectrum (aq_{local}).

An alternative measure of alignment is to monitor the amount of variation described by the first components of a principal components analysis (PCA) of the data. The simplicity value for a data matrix **X** provides such a measure, which takes the sum of the singular values (derived from singular value decomposition, SVD, of the data), scaled to unit variance, and raised to the fourth power (Equation 2):

$$\text{Simplicity} = \sum \left(SVD \left(\frac{x}{\sqrt{\sum_{i,j} (x_{ij}^2)}} \right) \right)^4 \qquad (2)$$

As the simplicity value approaches one, a greater proportion of the data variation is described by the first principal components (PC), which ideally describes the biological variation of interest. In the case of misalignment, the additional variation would result in an increase in the number of required PCs (i.e. a decrease in the amount of variation described by the first PCs), and thus a decrease in the simplicity value[14,20].

A problem common among many alignment algorithms is the introduction of spectral artifacts resulting from boundary interpolation. This may result in peak distortion (area and shape) and therefore the increased variation is not inherent to the data. The peak factor[14,20] is one measure to quantify such distortions and is calculated based on the Euclidean distance (norm) of the data before ($x_{i,raw}$) and after ($x_{i,aligned}$) alignment (Equation 3,4):

$$\text{Peak Factor}=\frac{\sum_{i=1}^{I}(1-\min(c(i),1)^2)}{I} \tag{3}$$

and

$$c(i)=\left|\frac{\text{norm}(x_{i,aligned})-\text{norm}(x_{i,raw})}{\text{norm}(x_{i,raw})}\right| \tag{4}$$

Ideally, peak distortion will be minimal after spectral alignment and thus the norm will be similar resulting in a peak factor of one. If distortion occurs there will be a difference in the norm post-alignment and the peak factor will decrease.

Principal Components Analysis—PCA was performed on the unaligned, sr/vr-icoShift, and sr/vr-RSPA aligned cell line and urine data sets after mean-centering, with n-1 PCs calculated for each model. The resulting scree plots of the unaligned, sr/vr-icoShift, and sr/vr-RSPA aligned PCA models are given in Fig. S-1 (Supporting Information). A qualitative measure of improved information recovery was obtained using a pseudo-variable importance to projection (VIP) score. The VIP score is applicable to supervised multivariate analysis techniques (e.g. PLS), and is essentially a sum of the weighted latent variable loadings with each loading weighted by the fraction of variation described by the latent variable. Here, a pseudo-VIP score was calculated, with the only difference to the traditional VIP being an unsupervised multivariate method was used (i.e. PCA). Since the average VIP over all variables is 1, the total number of variables with a pseudo-VIP score > 1 (i.e. number of variables considered significant) was determined for each PCA model. We refer to the VIP score throughout, with the implication that the pseudo-VIP was actually calculated.

Results and Discussion

Spectral alignment is often critical in maximizing the information retrieval from various statistical analysis techniques. The magnitude of alignment is often dependent on the pH and ionic strength variability that contribute to the chemical shift variation of the samples. Notably, such variations have been explored as a new source of biological information[32]. For this study, we have chosen aqueous extracts from prostate cell lines and urine as two sample classes spanning this variability range, from relatively well controlled (cell extracts) to large (urine samples) chemical shift variation. The normalized unaligned spectra are given in Fig. S-2C (cell extracts, Supporting Information) and Fig. 2C (urine). In the case of cell line extracts, the chemical shift variation before application of an alignment algorithm is already minimal, however this is not the case for the urine spectra (e.g. citrate region).

In this work, the prerequisite selection of a common reference spectrum representative of the full spectrum has been relaxed such that multiple spectra from a data set have the potential to act as a segment-specific reference spectrum. A natural result of this procedure is the definition of segments that are “most similar”, with the calculated boundaries potentially input into procedures such as icoShift that otherwise require subjective boundary definition (procedure schematic, Fig. 1). As proof-of-principle demonstration, we have applied this procedure to both aqueous cell line extract and urine ^{1}H NMR data using two alignment procedures, RSPA and icoShift.

Case 1 – Aqueous Cell Line Extract Data Set

In the case of the cell line data, 71 segments were identified with the resulting segment specific reference spectrum and distribution of reference spectra given in Fig. 3. For comparison, the reference spectrum chosen considering the full spectral data set was number 10 (single reference). Interestingly, the mean correlation between the unaligned spectra was quite high (0.81, Table 1), which is reflected in the distribution of reference spectra where each spectrum contributes to the alignment process. As expected, a highly correlated data set is consistent with the observation that each spectrum has the potential to act as a reference. Under this condition it is anticipated that global alignment towards a single reference would be satisfactory, which is indeed what was observed.

With identification of segment-specific reference spectra, alignment was performed using the RSPA algorithm. For comparison, the RSPA method was first applied to the cell line data using a single reference spectrum (number 10, sr-RSPA). The algorithm was then applied utilizing the individual global segments, choosing the segment-specific reference spectrum as the target (vr-RSPA). The alignment assessment parameters calculated as a result of the two alignment methods are summarized in Table 1. Both sr-RSPA and vr-RSPA methods performed equally well, which was further confirmed by visual inspection of the ^{1}H NMR spectra depicted in Fig. S-2A–C (Supporting Information).

Alignment was then performed utilizing the recently developed icoShift algorithm. As with RSPA, the icoShift algorithm was run under two conditions: alignment towards a single reference spectrum with the segment boundaries defined by our pre-segmentation process (71 segments, sr-icoShift), and alignment towards a segment-specific reference spectrum using the same boundaries as in sr-icoShift (vr-icoShift). All assessment parameters are listed in Table 1. Once again, regardless of the parameter chosen, both methods performed equally well (visually confirmed in Fig. S-2C–E). The benefit of pre-calculating segment boundaries can be appreciated when icoShift alignment is performed using segments of constant length. Increasing the total number of segments (i.e. decreasing the segment length) correlates with an increasing alignment quality at the expense of introducing spectral distortion (Fig. S-5, supporting information). Automatic calculation of segment boundaries consistently strikes a balance between maximizing alignment while minimizing spectral distortion, regardless of the data (cell extract or urine).

The generation of global segments for alignment towards a local common reference warrants examination of the resulting local alignments (aq_{loc}) after applying both RSPA and icoShift. As demonstrated in Fig. S-3 (Supporting Information), the alignment qualities on a per-segment basis for the variable reference method are nearly identical to the single reference case, regardless of the alignment algorithm applied. Segments may be identified where vr-RSPA slightly outperforms vr-icoShift (segment centered at 6.10 ppm) or conversely, vr-icoShift performs slightly better (segments between 6.22 – 6.60 ppm). Interestingly, the segments generated between 6.22 – 6.60 ppm encompass primarily noise, with only a single resonance appearing within this spectral region (fumarate, 6.50 ppm). The improved performance of vr-icoShift in this region is a direct result of a fundamental difference between the two alignment algorithms; RSPA utilizes a peak-picking algorithm in order to segment a spectral region prior to alignment. Thus, RSPA will not align a global segment containing purely noise, whereas icoShift will still maximize the spectral correlation within the same segment. The fact that icoShift performs better than RSPA within this range (or any other segment containing only noise) therefore would not be experimentally significant from the perspective of metabolomic information recovery.

Case 2 – Urine Data Set

In the case of the urine data collected at 900 MHz, 40 segments were identified in the ^{1}H NMR spectra with the resulting segment specific reference spectrum and distribution of reference spectra given in Fig. 3. The reference spectrum calculated including the entire spectral data set was number 9 (single reference). Once again, it is clear that a single reference does not satisfactorily represent all segments, however in contrast to the cell line example, several spectra are never chosen as a reference. This observation is also reflected in the decrease of the mean correlation among all unaligned spectra (0.60, Table 1).

Alignment was performed using the RSPA method as described in the cell line extract example above. In the case of the 900 MHz urine spectra, alignment was found to be superior utilizing vr-RSPA (Table 1), particularly in terms of overall alignment quality, simplicity factor, and final spectral correlation, also confirmed by visual inspection of Fig. 2A–C. Both methods tested had minimal impact of peak distortion regardless of the sample data set.

Alignment was also performed using the icoShift algorithm, as described in the cell line extract example. The resulting alignment assessment parameters were nearly equivalent regardless of which icoShift alignment procedure was applied (Table 1, Fig. 2C–E). Interestingly, the simplicity factor resulting from the vr-RSPA method was equivalent to the value calculated using either of the icoShift methods (as was the variance described by the first and second PCs), however the overall alignment quality for the vr-RSPA method was greater than either of the icoShift values.

The local alignment qualities (aq_{loc}) for the 900 MHz urine data after RSPA and icoShift alignment are given in Fig. 4. The observed aq_{loc} values are nearly identical in the case of sr-icoShift versus vr-icoShift, however vr-RSPA aq_{loc} values are consistently greater than the sr-RSPA case. Additionally, the overall pattern of aq_{loc} values between icoShift and RSPA is quite similar, with notable differences occurring between 2 – 3 ppm and 6.5 – 7 ppm, where vr-RSPA outperforms both of the icoShift methods. In fact, the increased global alignment quality of vr-RSPA versus the icoShift methods (Table 1) can be explained by these segment-specific alignment improvements.

To demonstrate the applicability of the variable reference method across multiple magnetic field strengths, ^{1}H NMR spectra of an alternative set of 20 urine samples were collected at 500 MHz. The alignment assessment results pre- and post-alignment are summarized in Table 1, with spectra presented in Fig. S-7 (supporting information). The lower resolution unaligned data (500 MHz) had higher mean correlation and aq_{glob} values compared to the high resolution data (900 MHz). Once again vr-RSPA yielded the greatest improvement of the aq_{glob} parameter, with both sr/vr-icoShift causing minimal spectral distortion. Interpreting the simplicity value and associated PCA parameters is complicated in the case of the low resolution data set due to the presence of strong glucose signals in two of the twenty samples. Any improvement as a result of alignment is masked since the glucose signals strongly dominate the PCA models. Notably, the spectral distortion artificially improves the PCA-related parameters, and thus emphasizes the importance in considering multiple alignment assessment parameters.

Metabolomic Information Retrieval

The goal of any alignment procedure is to ultimately improve the interpretability of results derived from multivariate analyses. Misalignment of spectral data will confound potentially important biological information and thus decrease the overall significance of the measurements. Therefore we subjected all data sets to PCA and examined the first and second PC loading plots for evidence of enhanced information recovery.

Principal component 1 loadings for the cell line spectra before and after application of both icoShift and RSPA are given in Fig. 5A (PC 2 loadings are available in the supporting information, Fig. S-4). Visual inspection reveals minimal difference between the pre-alignment data compared to any of the post-aligned data regardless of the alignment method applied. As an alternative measure of information recovery, variables were ranked according to their VIP score and the total number of variables with a VIP > 1 was obtained for each PCA model (Table 1). The total number of variables with VIP > 1 in the post-alignment models fluctuate within ± 1% of the pre-aligned data. This result reinforces the general observation that in the case of data sets exhibiting small inter-sample chemical shift variation, improvement of information recovery is independent of the alignment algorithm chosen.

The loadings for PC 1 generated for the 900 MHz urine spectra before and after application of icoShift and RSPA are given in Fig. 5B. Visual inspection reveals several spectral regions with very different, alignment protocol dependent, PC 1 loadings. In general, data sharing similar alignment assessment parameters (aq_{glob} and simplicity factor, sr/vr-icoShift and vr-RSPA) share common loading patterns. Additionally, the number of variables with VIP > 1 increased approximately 10 – 20 % relative to the unaligned data set (Table 1), demonstrating that the decreased chemical shift variation increased the significant variables contributing to the PCA models. While outside the scope of this report, this result suggests that the increase in significant signals post-alignment could translate to a more comprehensive understanding of the underlying biology.

Conclusion

As is the case with all alignment techniques, judicious choice of algorithm and respective parameters remains dependent on both the sample set and requirements of the user. As demonstrated in this report, all algorithms tested on samples requiring minimal adjustment in alignment yielded essentially equivalent results. The issue becomes important when considering samples exhibiting large phenotypic variation such as human urine biospecimens. Not surprisingly, a single reference spectrum correctly representing the entire expected chemical composition of all samples does not exist.

We have introduced a protocol that can be appended to any alignment algorithm and applied to a variety of data sets. As demonstrated, our protocol performed at least as well as standard alignment methodologies, with increased benefit observed for data sets exhibiting large inter-sample variation. Similar to 1D ^{1}H NMR spectral data, such a protocol could conceivably be applied to any data requiring alignment, including multi-dimensional NMR spectra, mass spectrometry spectra or chromatographic data.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was supported by funds from NIH (RR023597 to A.R.) and partly by GM095640 to A.R.

References

1. Da Silva Neto HG, da Silva JBP, Pereira GE, Hallwass F. Magn Reson Chem. 2009; 47:S127–S129. [PubMed: 19810052]
2. Spraul M, Schütz B, Humpfer E, Mörtter M, Schäfer H, Koswig S, Rinke P. Magn Reson Chem. 2009; 47:S130–S137. [PubMed: 19899106]

3. Sardans J, Peñuelas J, Rivas-Ulbach A. Chemoecology. 2011; 21:191–225.

4. Griffin JL, Shockcor JP. Nat Rev Cancer. 2004; 4:551–561. [PubMed: 15229480]

5. Issaq HJ, Fox SD, Chan KC, Veenstra TD. J Sep Sci. 2011; 34:3484–3492. [PubMed: 22102289]

6. Trygg J, Wold S. J Chemometr. 2002; 16:119–128.

7. Cloarec O, Dumas ME, Craig A, Barton RH, Trygg J, Hudson J, Blancher C, Gauguier D, Lindon JC, Holmes E, Nicholson J. Anal Chem. 2005; 77:1282–1289. [PubMed: 15732908]

8. Spraul M, Neidig P, Klauck U, Kessler P, Holmes E, Nicholson JK, Sweatman BC, Salman SR, Farrant RD, Rahr E, Beddell CR, Lindon JC. J Pharmaceut Biomed. 1994; 12:1215–1225.

9. Kassidas A, MacGregor JF, Taylor PA. AlChE Journal. 1998; 44:864–875.

10. Nielsen NPV, Carstensen JM, Smedsgaard J. J Chromatogr A. 1998; 805:17–35.

11. Vogels JTWE, Tas AC, Venekamp J, Van Der Greef J. J Chemometr. 1996; 10:425–438.

12. Forshed J, Schuppe-Koistinen I, Jacobsson SP. Anal Chim Acta. 2003; 487:189–199.

13. Eilers PHC. Anal Chem. 2004; 76:404–411. [PubMed: 14719890]

14. Skov T, van den Berg F, Tomasi G, Bro R. J Chemometr. 2006; 20:484–497.

15. Clifford D, Stone G, Montoliu I, Rezzi S, Martin FP, Guy P, Bruce S, Kochhar S. Anal Chem. 2009; 81:1000–1007. [PubMed: 19138127]

16. Pravdova V, Walczak B, Massart DL. Anal Chim Acta. 2002; 456:77–92.

17. Tomasi G, van den Berg F, Andersson C. J Chemometr. 2004; 18:231–241.

18. Forshed J, Torgrip RJO, Åberg KM, Karlberg B, Lindberg J, Jacobsson SP. J Pharmaceut Biomed. 2005; 38:824–832.

19. Åberg KM, Alm E, Torgrip RJO. Anal Bioanal Chem. 2009; 394:151–162. [PubMed: 19198812]

20. Giskeødegàrd GF, Bloemberg TG, Postma G, Sitter B, Tessem MB, Gribbestad IS, Bathen TF, Buydens LMC. Anal Chim Acta. 2010; 283:1–11.

21. Wong JWH, Durante C, Cartwright HM. Anal Chem. 2005; 77:5655–5661. [PubMed: 16131078]

22. Wong JWH, Cagney G, Cartwright HM. Bioinformatics. 2005; 21:2088–2090. [PubMed: 15691857]

23. Veselkov KA, Lindon JC, Ebbels TMD, Crockford D, Volynkin VV, Holmes E, Davies DB, Nicholson JK. Anal Chem. 2009; 81:46–66.

24. Savorani F, Tomasi G, Engelsen SB. J Magn Reson. 2010; 202:190–202. [PubMed: 20004603]

25. MacKinnon N, Khan AP, Chinnaiyan AM, Rajendiran TM, Ramamoorthy A. Metabolomics. 201210.1007/s11306-012-0398-4

26. Ackerstaff E, Phug BR, Nelson JB, Bhujwalla ZM. Cancer Res. 2001; 61:3599–3603. [PubMed: 11325827]

27. Beckonert O, Keun HC, Ebbels TMD, Bundy J, Holmes E, Lindon JC, Nicholson JK. Nat Protoc. 2007; 2:2692–2703. [PubMed: 18007604]

28. Dieterle F, Ross A, Schlotterbeck G, Senn H. Anal Chem. 2006; 78:4281–4290. [PubMed: 16808434]

29. [Accessed October 20, 2011.] www.models.life.ku.dk

30. Savitzky A, Golay MJE. Anal Chem. 1964; 36:1627–1639.

31. Steinier J, Termonia Y, Deltour J. Anal Chem. 1972; 44:1906–1909. [PubMed: 22324618]

32. Cloarec O, Dumas ME, Trygg J, Craig A, Barton RH, Lindon JC, Nicholson JK, Holmes E. Anal Chem. 2005; 77:517–526. [PubMed: 15649048]

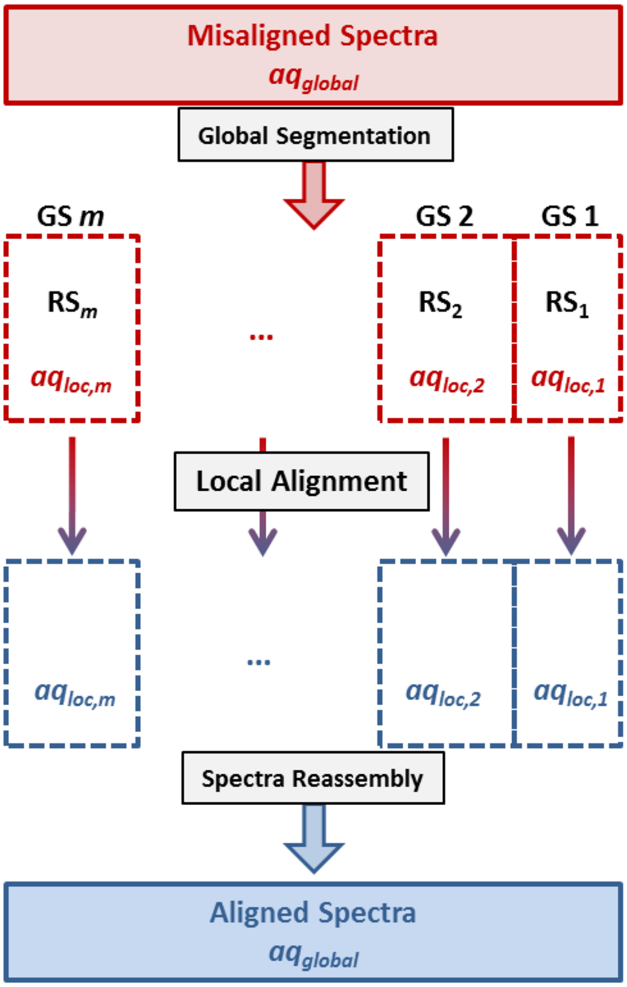

Figure 1.

A schematic representation of the alignment protocol incorporating the variable reference procedure. A misaligned spectral dataset (with associated global alignment quality, aq_{global}) is first divided into a total of m global segments (GS), under the condition that all spectra share a GS-specific reference spectrum (RS_m). Each GS is subjected to local alignment. A local alignment quality ($aq_{loc,\ m}$) can be calculated for each of the m GS before and after local alignment. The aligned GS are then reassembled to yield the globally aligned full spectra (with associated aq_{global}).

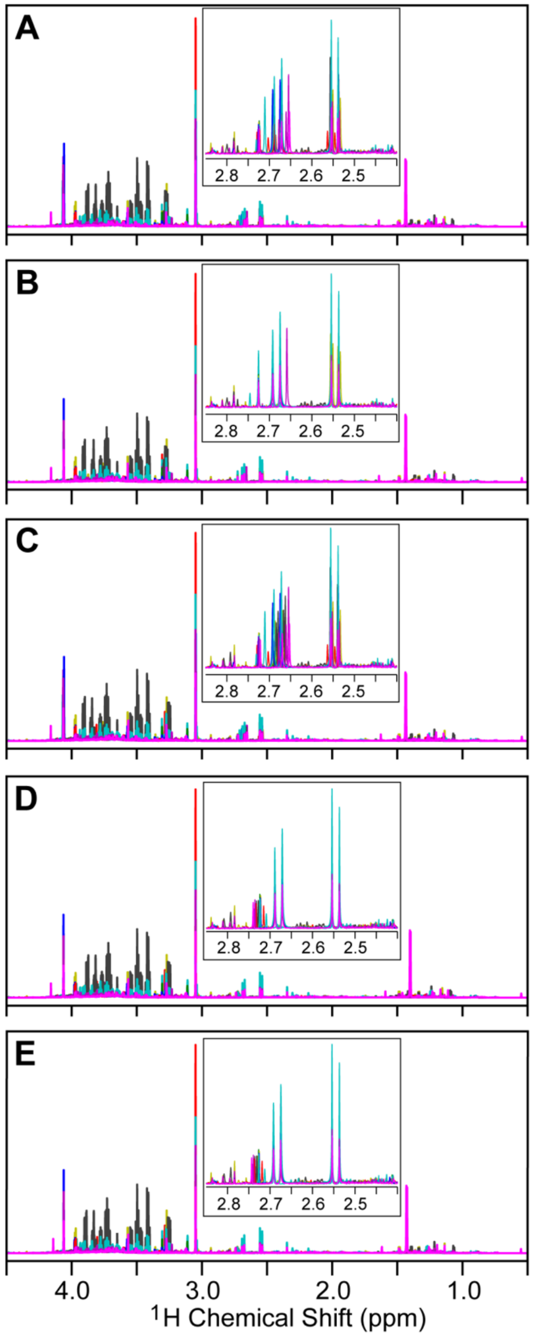

Figure 2.

900 MHz ^{1}H NMR spectra of human urine pre- and post-alignment. Overlay of all 20 spectra of aligned single reference (sr)-RSPA (**A**), aligned variable reference (vr)-RSPA (**B**), unaligned raw spectra (**C**), aligned sr-icoShift (**D**), and aligned vr-icoShift (**E**). Inset: expansion between 2.4 – 2.85 ppm highlighting the citrate chemical shift region.

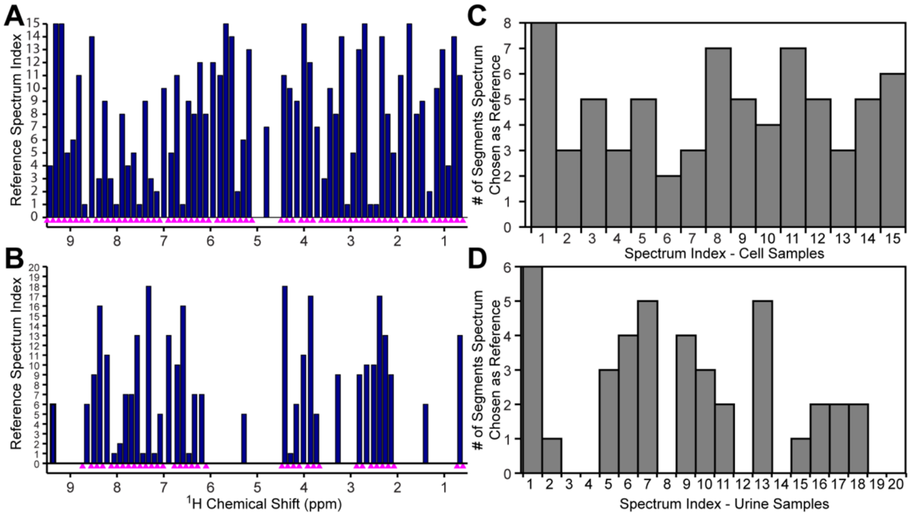

Figure 3.

Calculated segments and corresponding reference spectra for cell line and 900 MHz urine spectra. Reference spectra identified for each segment with corresponding segment boundaries (magenta triangles) plotted along the chemical shift axis for the cell line (**A**) and urine (**B**) data. Reference spectrum distributions across the global segments for **C**) cell line spectra and **D**) urine spectra. For comparison, spectrum # 9 and 10 was chosen as the cell line and urine single reference, respectively.

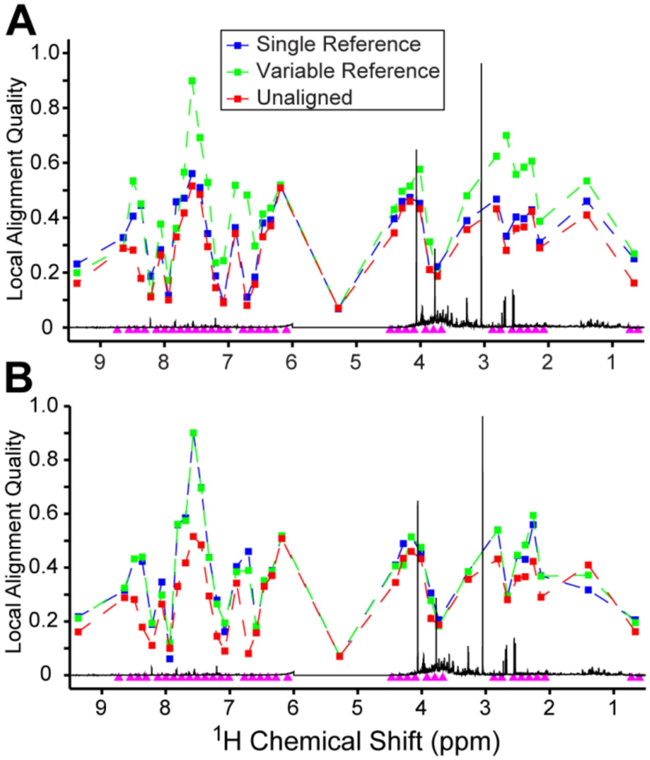

Figure 4.

Segment-specific local alignment qualities (aq_{loc}) calculated before and after **A**) RSPA and **B**) icoShift alignment of 900 MHz ^{1}H NMR spectra of human urine. Local alignment qualities were calculated for each segment on the unaligned spectra (red) and after both single reference and variable reference alignment RSPA/icoShift strategies (blue, green respectively). Magenta triangles mark calculated segment boundaries. In total, 40 segments were identified for the urine spectra. A representative spectrum is plotted for reference.

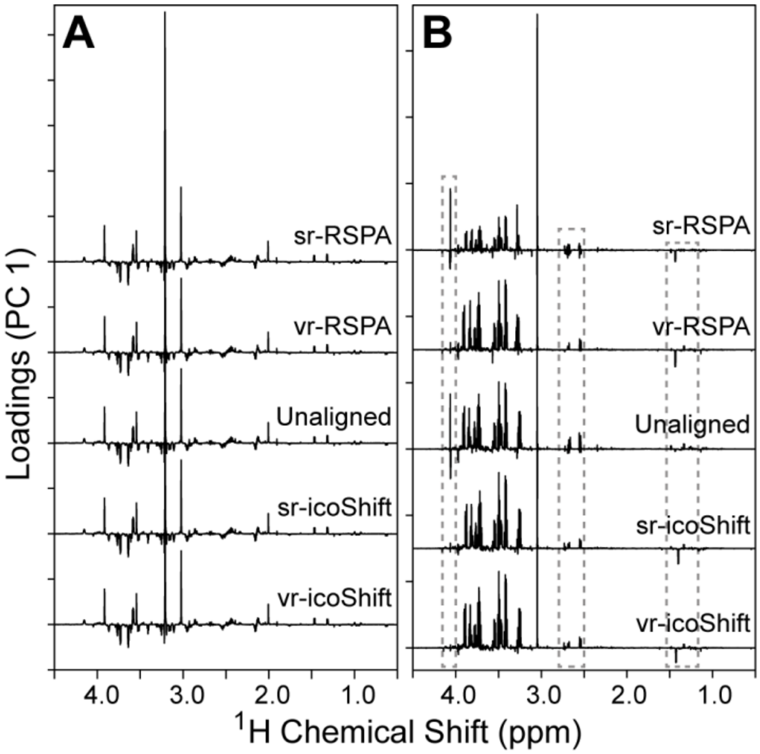

Figure 5.

Principal component (PC) 1 loadings after application of PCA of the cell line data (**A**) and 900 MHz urine data (**B**), pre- and post sr/vr-alignment. The amount of variation described by PC 1 for each data set is given in Table 1. The loadings calculated for the cell line data (A) are nearly identical, regardless of the alignment protocol applied. Significant differences observed in the PC 1 loadings of the urine data (B) are highlighted in the boxed regions. The loadings after application of vr-RSPA are visually quite similar to the sr/vr-icoShift loadings, which is reflected in the similarity of the assessment quality parameters listed in Table 1. The vertical scale is the same for each cell line and urine loading plot.

Table 1

Alignment assessment parameters as a result of applying alignment protocols to cell line and urine data. Equation 1 was used to calculate aq_{global} with an α = 0.05 ppm (α = 0.08 for the urine data at 500 MHz). The amount of variation described by principal component (PC) 1 and 2 after application of PCA is given as a percentage of the total variation. The number of variables with a VIP score greater than 1, calculated from the PCA models, are also given.

Sample	Method	Mean corr	aq_{global}	Simplicity Value	Peak Factor	PC 1 (%)	PC 2 (%)	VIP > 1
Cell Lines	Unaligned	0.806	0.530	0.681	1.00	60.8	18.8	1692
	sr-icoShift	0.809	0.547	0.686	1.00	61.7	18.7	1682
	vr-icoShift	0.809	0.550	0.686	1.00	61.6	18.7	1676
	sr-RSPA	0.807	0.550	0.682	1.00	60.7	18.8	1694
	vr-RSPA	0.808	0.547	0.684	1.00	61.4	18.8	1676
Urine 900 MHz	Unaligned	0.604	0.280	0.446	1.00	29.2	28.7	1528
	sr-icoShift	0.824	0.328	0.648	1.00	39.5	31.2	1847
	vr-icoShift	0.826	0.329	0.650	1.00	40.3	31.2	1788
	sr-RSPA	0.629	0.319	0.467	1.00	31.4	25.0	1558
	vr-RSPA	0.822	0.381	0.648	1.00	40.2	30.9	1727
Urine 500 MHz	Unaligned	0.662	0.348	0.482	1.00	74.9	13.3	1028
	sr-icoShift	0.769	0.392	0.517	0.998	78.0	15.1	1139
	vr-icoShift	0.765	0.393	0.498	0.998	76.9	17.9	1091
	sr-RSPA	0.711	0.373	0.420	1.00	66.3	15.4	1120
	vr-RSPA	0.763	0.405	0.511	1.00	78.3	12.3	1015