INFORMS Artificial Intelligence and Data Mining Workshop, Pittsburgh, PA, Nov. 2006.

A Bayesian Approach for the Alignment of High-Resolution NMR Spectra

Seoung Bum Kim[†], **Zhou Wang**[‡], **Carlos M. Duran**[†]

[†]Department of Industrial and Manufacturing Systems Engineering
[‡]Department of Electrical Engineering
The University of Texas at Arlington
Arlington, TX 76019 USA

September 1, 2006

Abstract

The rapid progresses in human genome project and biotechnologies result in the sheer volume of datasets associated with in-depth scientific knowledge. Metabolomics is defined as the study for understanding metabolic process in living systems. Metabolomics approaches that used high-resolution nuclear magnetic resonance (NMR) spectroscopy have been used to characterize metabolic variations in response to physiological alternation, disease states, genetic modification, and nutrition intake. An NMR spectrum usually involves tens of thousands of variables and the comparison of multiple spectra lead to a huge number of data points and a situation that poses a great challenge to analytical and computational capabilities. When considering multiple spectra, small variations due to concentration, pH, and temperature, influence the spectral alignment and thus can interfere with direct comparisons between samples. Thus, it is crucial to align spectra before applying any subsequent statistical analyses such as clustering and classification. In this study, we propose a novel algorithm for the NMR spectra alignment within the Bayesian framework, which allows estimating the vertical and horizontal shifts simultaneously in the existence of noise. Effectiveness of our algorithm is demonstrated through the comparison of existing algorithms and experiments with real high-resolution NMR data.

Keywords: metabolomics; nuclear magnetic resonance (NMR); alignment; Bayesian algorithm

1. Introduction

Metabolomics is developing as a major means to investigate dynamic and time-dependent metabolic patterns in association with dietary, environmental, and pathophysiologic stimuli as they occur in integrated biological systems (Nicholson *et al.,* 1999, 2002). A variety of techniques are available for studying metabolomics, of these, high-resolution NMR spectroscopy has advantages in terms of minimal sampling processing, quantitative calibration, minimally invasiveness, and cost (Lindon, 2004). In NMR spectra, the x-axis indicates the chemical shift with units in ppm, and the y-axis indicates the intensity values corresponding to each chemical shift (Fig. 1). A set of chemical shifts constitutes peaks, the specific resonance of chemical species in the sample. An NMR spectrum usually involves tens of thousands of variables and the comparison of multiple spectra lead to a huge number of data points and a situation that poses a great challenge to analytical and computational capabilities. Statistical pattern recognitions, such as unsupervised and supervised methods can reduce such complexity, and thus, facilitate the extraction of implicit pattern and help discriminate spectra according to their biological/experimental conditions (Beckonert *et al.*, 2003; Holmes *et al.*, 2001; Lindon *et al.*, 2001).

NMR spectra require preprocessing steps before applying statistical pattern recognition methods in order to detect subtle variations. Generally, preprocessing steps include phase/base line corrections, alignment, and normalization. In this paper, we focus on spectral alignment to ensure the direct comparison of multiple spectra. Small variations in spectra due to concentration, pH, temperature, and instrumental instabilities influence the spectral alignment and, thus, can interfere with direct comparisons between samples (Lindon, 2004). Therefore, it is crucial to align the spectra prior to any subsequent analyses. Fig. 1 shows an example that demonstrates the necessity of spectral alignment when considering multiple spectra.

A number of methods have been proposed to deal with the alignment of spectra. DTW and COW have received much attention for the alignment of spectral datasets (Pravdova *et al.*, 2002; Tomasi *et al.*, 2004).

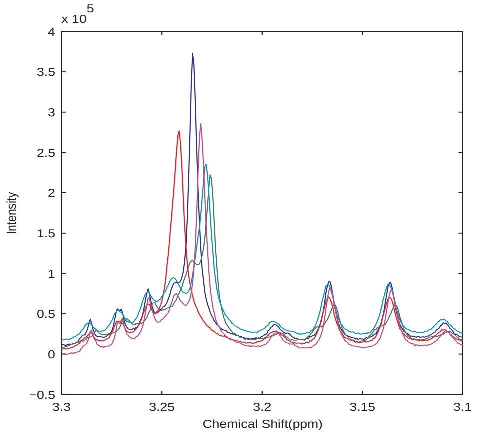

Figure 1: A portion of data for five NMR spectra from human plasma before alignment.

Others include a genetic algorithm-based method (Forshed *et al.*, 2003), partial linear fit (Vogel et al., 1996), and reduced set mapping (Torgrip et al., 2003).

This study presents an algorithm for the alignment of NMR spectra based on the Bayesian statistical modeling framework that enables us to incorporate the noise effect and prior knowledge about the amount of shift in a nature way. Further, the proposed method can simultaneously estimate the spectral shift and the baseline intensity variation. In the next section, we describe the DTW and COW algorithms followed by the proposed algorithm. Finally, effectiveness of the proposed algorithm is demonstrated through the comparison of the dynamic time warping (DTW) and correlation optimized warping (COW) algorithms using real NMR spectra.

2. Dynamic Time Warping & Correlation Optimized Warping

2.1 Dynamic time warping

DTW is originally developed as an alignment method for speech recognition (Sakoe *et al.*, 1971). The DTW algorithm calculates cumulative distance function that measures the similarity of two signals which may vary in time and then find the optimal path using dynamic programming. Consider two spectra, R (length $|R|$) for a reference spectrum and S (length $|S|$) for a sample spectrum that needs to be aligned. A grid plot is constructed with size $|R| \times |S|$. Then a set of warping paths $\mathbf{P}$ of M points is defined as follows:

$$\mathbf{P} = \{[r(m), s(m)], m = 1, 2, \ldots, M\},$$

where r and s denote the index of R and S. We search for an optimal warping path such that a cumulative distance between the two spectra is minimized. The cumulative distance function is calculated based on the following three possible paths:

$$D(i,j) = min \begin{cases} D(i-1,j) + d(i,j) \\ D(i-1,j-1) + d(i,j) \\ D(i,j-1) + d(i,j) \end{cases} ,$$

where $D(i,j)$ is the local cumulative distance of point (i,j) and $d(i,j)$ is the local distance of point (i,j). Fig. 2 provides a graphical illustration of possible paths to compute the $D(i,j)$.

2.2 Correlation optimized warping

The COW algorithm is initially devised for correcting misalignment of chromatogram (Nielsen *et al.*, 1998). Compared with DTW, COW uses the segment of data instead of individual data points to maximize the overall correlation between the reference and the sample spectra. Figure 3 shows the graphical illustration of the COW algorithm. T is a reference spectrum with length L_T. R is a sample (unaligned) spectrum with length L_R and is divided into N sections with the length of m. Each section in R is stretched or shrinked by linear interpolation within the range $(\delta - t, \delta + t)$, so that the aligned spectrum R with the same length as T is obtained. The parameter "t", so called "slack parameter" defines the maximum length increase $(+t)$

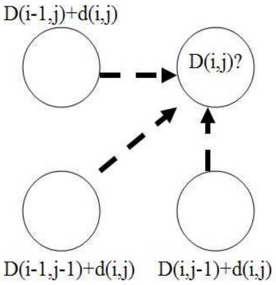

Figure 2: Allowable paths to compute the $D(i,j)$.

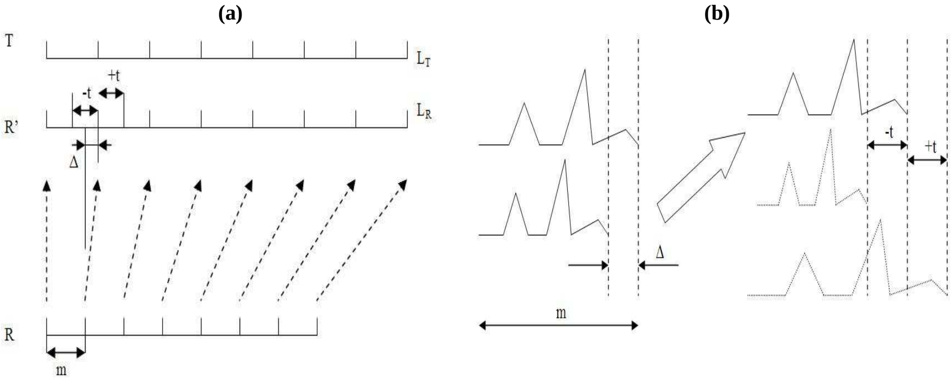

Figure 3: Graphical illustration of COW. Each section in R is stretched or shrinked by linear interpolation within the range $(\delta - t, \delta + t)$ and find an optimal warping path by solving a segment-wise correlation optimization.

or length decrease $(-t)$. δ characterizes the different length between the reference and sample spectra and obtained from $\delta = (L_T/N) - m$.

In order to find the optimal path, two matrices of size $(L_T + 1) \times (N + 1)$ are constructed. Matrix **A** is initialize, assigning zero to the element $(L_T + 1, N + 1)$ and minus infinitive to the rest of the elements. Correlation between the section belonging to the reference profile (ζ_T) and interpolated section of the profile $\zeta_{R'}$is computed as follows:

$$\rho = \frac{(\zeta_T - \overline{\zeta}_T)^T(\zeta_{R'} - \overline{\zeta}_{R'})}{\sigma_{\zeta_T}\sigma\zeta_{R'}}, \tag{1}$$

where $\overline{\zeta}_T$ and σ_ζ are respectively mean and standard deviation of ζ.

Then, for the all possible warpings in the range $[\delta - t, \delta + t]$ the ρs are calculated and added to the value of the benefit function from the previous ending point. Only the highest value (h) is kept in the matrix **A**. The second matrix **B** are constructed, keeping the corresponding values of of h. Once **B** is completely determined, the optimal path is selected by back tracing.

3. Bayesian Statistical Modeling for Spectral Alignment

Let $x(\omega)$ and $y(\omega)$ be two spectral signals to be aligned, where ω is the frequency index of the spectra. We can formulated two spectral signals as follows:

$$y(\omega) = x(\omega + \Delta\omega) + \Delta a, \tag{2}$$

where we call $\Delta\omega$ and Δa the spectral shift and the baseline amplitude variation, respectively. A Taylor series expansion of the right hand side at ω_0 yields

$$y(\omega_0) = x(\omega_0) + \Delta\omega\frac{dx}{d\omega}|_{\omega_0} + \frac{(\Delta\omega)^2}{2!}\frac{d^2x}{d\omega^2}|_{\omega_0} + \cdots + \Delta a. \tag{3}$$

In practice, the amount of the spectral shift and the baseline amplitude variation are typically not fixed, but varies smoothly along the frequency axis. NMR spectral data acquired is discrete along the frequency axis.

Assume that there are N samples within an a local spectral region-of-interest (SROI) from the two spectra. They can be denoted as $\{x(\omega_1),\, x(\omega_2),\, ...,\, x(\omega_N)\}$ and $\{y(\omega_1),\, y(\omega_2),\, ...,\, y(\omega_N)\}$. Also assume that the frequency shift $\Delta\omega$ is small, so that the second and higher order terms can be ignored. We can then write

$$\mathbf{y} = \mathbf{x} + \Delta\omega\, \mathbf{x'} + \Delta a\, \mathbf{1}, \tag{4}$$

where $\mathbf{x} = [x(\omega_1),\, x(\omega_2),\, ...,\, x(\omega_N)]^T$, $\mathbf{y} = [y(\omega_1),\, y(\omega_2),\, ...,\, y(\omega_N)]^T$, $\mathbf{x'} = [\frac{dx}{d\omega}|_{\omega_1},\, \frac{dx}{d\omega}|_{\omega_2},\, ...,\, \frac{dx}{d\omega}|_{\omega_N}]^T$, and $\mathbf{1}$ is an N dimensional column vector with all entries equaling 1. Reorganizing Eq. (4) into a matrix operation format, we obtain

$$\mathbf{Ac} = \Delta\mathbf{x}\,, \tag{5}$$

where $\mathbf{A} = [\mathbf{x'}\;\; \mathbf{1}]$, $\Delta\mathbf{x} = \mathbf{y} - \mathbf{x}$, and $\mathbf{c} = [\Delta\omega\;\; \Delta a]^T$ is a column vector containing the amount of shifts we would like to estimate.

Motivated by the Bayesian approach in optical flow estimation (Simoncelli et al., 1991), to account for the noise effect in a model, we define

$$\epsilon = \mathbf{Ac} - \Delta\mathbf{x}. \tag{6}$$

as a zero-mean Gaussian random vector, in which all entries are independently and identically distributed Gaussian random variables. The covariance matrix of ϵ is thus diagonal and is denoted as $\Lambda_n\mathbf{I}$, where Λ_n is the noise variance and $\mathbf{I}$ is the identity matrix. We can then write the probability density function (PDF) of ϵ for a given $\mathbf{c}$ as

$$p(\epsilon|\mathbf{c}) \propto \exp\left\{-\frac{(\mathbf{Ac} - \Delta\mathbf{x})^T(\mathbf{Ac} - \Delta\mathbf{x})}{2\Lambda_n}\right\}. \tag{7}$$

We obtained the following by applying Bayes' rule:

$$p(\mathbf{c}|\epsilon) \propto p(\epsilon|\mathbf{c})\, p(\mathbf{c})\,. \tag{8}$$

For the prior distribution $p(\mathbf{c})$, we assume c follows Gaussian distribution with mean zero and diagonal covariance matrix $\mathbf{\Lambda}_p$:

$$p(\mathbf{c}) \propto \exp\left\{-\frac{1}{2}\mathbf{c}^T\mathbf{\Lambda}_p^{-1}\mathbf{c}\right\}\,. \tag{9}$$

Finally, the posterior PDF would be

$$\begin{aligned} p(\mathbf{c}|\epsilon) &\propto \exp\left\{-\frac{(\mathbf{Ac} - \Delta\mathbf{x})^T(\mathbf{Ac} - \Delta\mathbf{x})}{2\Lambda_n}\right\}\exp\left\{-\frac{1}{2}\mathbf{c}^T\mathbf{\Lambda}_p^{-1}\mathbf{c}\right\} \\ &= \exp\left\{-\frac{1}{2}\left[\mathbf{c}^T\left(\frac{\mathbf{A}^T\mathbf{A}}{\Lambda_n} + \mathbf{\Lambda}_p^{-1}\right)\mathbf{c}\;\; + \frac{2\,\mathbf{A}^T\Delta\mathbf{x}}{\Lambda_n}\mathbf{c} + \frac{\Delta\mathbf{x}^T\Delta\mathbf{x}}{\Lambda_n}\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\mathbf{c} - \mathbf{m}_c)^T\mathbf{\Lambda}_c^{-1}(\mathbf{c} - \mathbf{m}_c)\right\}\,, \end{aligned} \tag{10}$$

where

$$\mathbf{m}_c = \mathbf{\Lambda}_c\,\frac{\mathbf{A}^T\Delta\mathbf{x}}{\Lambda_n}, \quad \mathbf{\Lambda}_c = \left(\frac{\mathbf{A}^T\mathbf{A}}{\Lambda_n} + \mathbf{\Lambda}_p^{-1}\right)^{-1}. \tag{11}$$

This posterior PDF is obviously Gaussin with the mean vector $\mathbf{m}_c$ and the covariance matrix $\mathbf{\Lambda}_c$. If the squared-error loss function (i.e., $\mathcal{L}[\mathbf{c}, \delta(\mathbf{g})]$) is used, the Bayes' solution $\delta(\mathbf{g})$ is the mean of the $p(\mathbf{c}|\epsilon)$. That is,

$$\begin{aligned} \delta(\mathbf{g}) &= \mathbf{E}[\mathbf{c}|\epsilon] = \mathbf{m_c} \\ &= \left(\frac{\mathbf{A}^T\mathbf{A}}{\Lambda_n} + \mathbf{\Lambda}_p^{-1}\right)^{-1}\frac{\mathbf{A}^T\Delta\mathbf{x}}{\Lambda_n} \\ &= \left(\mathbf{A}^T\mathbf{A} + \Lambda_n\mathbf{\Lambda}_p^{-1}\right)^{-1}\mathbf{A}^T\Delta\mathbf{x}\,. \end{aligned} \tag{12}$$

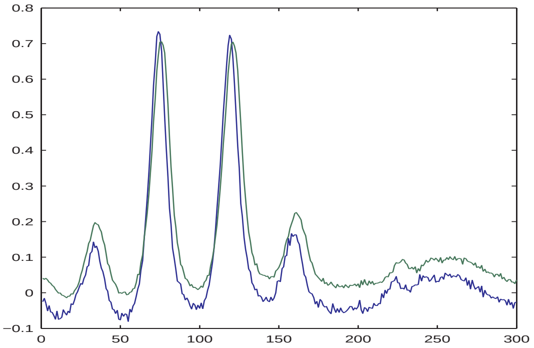

Figure 4: A portion of NMR spectra before alignment.

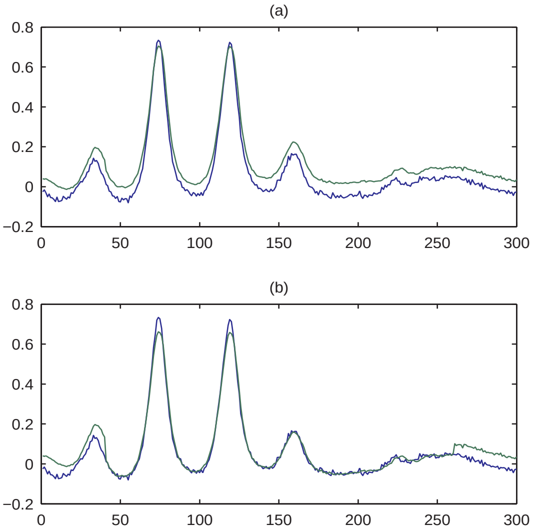

Figure 5: A portion of NMR spectra after correcting (a) only spectral shift, and (b) both spectral shift and amplitude variation.

4. Experiment

The real experimental NMR spectra are acquired from a 600 MHz ^{1}H NMR spectroscopy. Fig. 4 shows a small portion of the spectrum before alignment. Fig. 5 shows the results of alignment after correcting (a) spectral (phase) variation, and (b) baseline (amplitude) variation by using the proposed algorithm. It can be seen that the spectra are well aligned, demonstrating the applicability of the proposed algorithm.

We compared the proposed algorithm with the DTW and COW algorithms. We aligned 158 spectra based on the reference spectrum computed by their median. Each spectrum comprises 8,000 points. It should be noted that DTW and COW run out of memory with the whole spectral region. Thus, we divided the whole spectral region into three subregions in order for the computer to run DTW and COW. The first subregion (1-3,000) contains mostly the large peaks, while the third region (4,001-7,000) contains mostly the small peaks. The second subregion (2,001-5,000) contains both the large and small peaks. The comparison of three methods is reported in Tables 1-3. The root mean square error (RMS), correlation, and execution time are used as performance measures. The results of MS and correlation show that the proposed algorithm performs better than DTW and COW. In terms of execution time, the proposed algorithm outperforms the DTW and COW algorithms.

Table 1: A comparison of three alignment methods (region 1-3,000)

Methods	RMS (Before)	RMS (After)	Corr. (Before)	Corr. (After)	Time (Sec.)
DTW	0.0026375	0.0022386	0.97888	0.98214	4,520
COW	0.0026375	0.0029455	0.97888	0.97462	7,681
Bayesian	0.0026375	0.00058434	0.97888	0.99694	347

Table 2: A comparison of three alignment methods (region 4,001-7,000)

Methods	RMS (Before)	RMS (After)	Corr. (Before)	Corr. (After)	Time (Sec.)
DTW	0.0029509	0.0026358	0.97474	0.97643	4, 492
COW	0.0029509	0.0029933	0.97474	0.97525	7, 676
Bayesian	0.0029509	0.00076368	0.97474	0.99188	346

Table 3: A comparison of three alignment methods (region 2,001-5,000)

Methods	RMS (Before)	RMS (After)	Corr. (Before)	Corr. (After)	Time (Sec.)
DTW	0.0018504	0.0017132	0.9346	0.9478	5, 220
COW	0.0018504	0.0018116	0.9346	0.9522	7, 739
Bayesian	0.0018504	0.00079032	0.9346	0.9547	293

5. Conclusions

A novel algorithm is proposed for the alignment of NMR spectra within the Bayesian statistical modeling framework. The proposed algorithm allows us to simultaneously estimate the spectral shift and baseline variation. Further, Bayesian modeling enables us to incorporate the noise effect and prior knowledge about the spectral shift and baseline variation in a nature way. Our experimental results using real NMR spectra demonstrate the effectiveness of the proposed algorithm compared with the DTW and COW algorithms.

References

[1] Beckonert, O., Bollard, M.E., Ebbels, T.M.D., Keun, H.C., Antti, H.,Holmes, E., Lindon, J.C., Nicholson, J. K. (2003). NMR-based metabonomics toxicity classification:hierarchical cluster analysis and k-nearest-neighbour approaches. *Anal.Chem.Acta* **490** 3-15.

[2] Forshed, J., Schuppe-Koistinen, I., Jacobsson, S.P., (2002). Peak alignment of NMR signals by means of a genetic algorith *Anal.Chem.Acta* **487** 189-199.

[3] Holmes, E., Nicholson, J.K., Tranter, G. (2001) Metabonomic characterization of genetic variations in toxicological metabolic responses using probabilistic neural networks *Chem. Res. Toxicol.* **14** 181-191.

[4] Lindon, J.C. (2004) Metabonomics - techniques and applications, *Business Briefing: Future Drug Discovery* 1-6.

[5] Lindon, J.C., Holmes, E., Nicholson, J.K. (2001) K. Pattern recognition methods and applications in biomedical magnetic resonance, *Progress in Nuclear Magnetic resonance spectroscopy* **39** 1-40.

[6] Nicholson, J. K., Connelly, J.,Lindon, J.C., Holmes, E. (2002). Metabonomics: a platform for studying drug toxicity and gene function. *Nature Review Drug Discovery* **1** 153-161.

[7] Nicholson, J. K.,Lindon, J.C., Holmes, E. (1999). Metabonomics: Understanding the metabolic response of living systems to pathophysiological stimuli via multi-variate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* **29** 1181-1189.

[8] Nielsen, N.P.V., Carstensen, J.M, Smedsgaard, J., (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimized warping *Journal of Chromatography A* **805** 17-35.

[9] Pravdova, V., Walczak, B., Massart, D.L., (2002). A comparison of two algorithms for warping of analytical signals *Anal.Chem.Acta* **456** 77-92.

[10] Sakoe, H., Chiba, S., (1971). *Proceedings of the Internationl Congress of Acoustics*, Budapest, Paper 20 C13.

[11] Simoncelli, E.P., Adelson, E.H., Heeger, D.J., (1991). Probability distributions of optical flow. In IEEE Inter. Conf. Computer Vision & Pattern Recognition, pages 310315, Mauii, Hawaii.

[12] Tomasi, G., Berg, F.V,D, Anderson, C., (2004). Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data *Journal of Chemometrics* **18** 231-241.

[13] Torgrip, R.J.O., Aberg, M, Jacobsson, S.P., (2003). Peak alignment using reducing set mapping *Journal of Chemometrics* **17** 573-582.

[14] Vogels, J.T.W.E., Tas, A.C., Venekamp, J., van der Greef J., J (1996). Partial linear fit: A new NMR spectoscopy preprocessing tool for pattern recognition applications *Journal of Chemometrics* **10** 425-438.