2011 IEEE International Conference on Bioinformatics and Biomedicine

Bayesian Alignment Model for LC-MS Data

Tsung-Heng Tsai*†, Mahlet G. Tadesse‡, Yue Wang† and Habtom W. Ressom*§
*Lombardi Comprehensive Cancer Center, Georgetown University, Washington DC, USA
†Department of Electrical and Computer Engineering, Virginia Tech, Arlington, VA, USA
‡Department of Mathematics and Statistics, Georgetown University, Washington DC, USA
§Corresponding Author: hwr@georgetown.edu

***Abstract*—A Bayesian alignment model (BAM) is proposed for alignment of liquid chromatography-mass spectrometry (LC-MS) data. BAM is composed of two important components: prototype function and mapping function. Estimation of both functions is crucial for the alignment result. We use Markov chain Monte Carlo (MCMC) methods for inference of model parameters. To address the trapping effect in local modes, we propose a block Metropolis-Hastings algorithm that leads to better mixing behavior in updating the mapping function coefficients. We applied BAM to both simulated and real LC-MS datasets, and compared its performance with the Bayesian hierarchical curve registration model (BHCR). Performance evaluation on both simulated and real datasets shows satisfactory results in terms of correlation coefficients and ratio of overlapping peak areas.**

***Keywords*-alignment; Bayesian inference; block Metropolis-Hastings algorithm; liquid chromatography-mass spectrometry (LC-MS); Markov chain Monte Carlo (MCMC).**

I. Introduction

Recent advances in liquid chromatography-mass spectrometry (LC-MS) technology have led to more effective approaches for measuring changes in peptide/protein abundances in biological samples [1]. Label-free LC-MS methods [2] have been used for extraction of quantitative information and for detection of differentially abundant peptides/proteins. However, difference detection using label-free LC-MS methods requires that various preprocessing steps are appropriately handled [3]. Retention time alignment is one of the most important steps in the pipeline. Since LC processes result in substantial variation in retention time across multiple LC-MS runs, without appropriate correction, the subsequent analysis can yield misleading results. Therefore, retention time alignment is a prerequisite for the quantitative analysis of LC-MS data and is the focus of this paper.

Each LC-MS run generates data consisting of thousands of ion intensities distinguished in specific retention time (RT) and mass-to-charge ratio (m/z) values. Based on the type of input data, alignment approaches can be roughly classified into two categories [4]: 1) profile-based approaches and 2) feature-based approaches. The profile-based approaches make use of the unprocessed LC-MS data to estimate the variability along retention time and adjust the LC-MS runs accordingly. It is assumed that there exists a pattern representing multiple LC-MS runs from samples in the same group and the profile variability is relatively small compared to distortions caused by misalignment. The feature-based approaches, on the other hand, distinguish relevant signals (*features*, usually referred to as peaks) from irrelevant parts in the first step, and rely on the identified features for the alignment task.

Most of the existing alignment approaches are based on *hard* selection of reference (profile or feature), which does not allow any adjustment of the reference during alignment. Considering the lack of consistency in terms of the presence across all samples, important reference information may be missed. Another issue of the current alignment approaches is the lack of uncertainty assessment that is desired for better justification or decision making in subsequent analysis. From the perspective of study design, strategies utilizing spike-in information and/or MS/MS identification results may lead to better alignment results, e.g., in [5], [6]. Such hybrid approaches could bring great benefits if complementary information from multiple sources is combined properly. However, without uncertainty assessment, it is difficult to proceed with information fusion.

In this paper, we present a Bayesian approach for profile-based alignment. The alignment approach searches for 1) a prototype function that characterizes the consistent pattern and 2) a set of mapping functions that characterize the relationship between the prototype function and observed data. The goal is to estimate the prototype/mapping functions that produce data close to the observations. It is essential to define a similarity metric based on which the prototype function and the mapping functions can be estimated accordingly. An earlier work by Telesca and Inoue [7] presents a Bayesian hierarchical model for curve registration (BHCR) and provides a Markov chain Monte Carlo (MCMC) method for parameter inference. For LC-MS data alignment, due to the complexity of data consisting of many chromatographic peaks, more rigorous ways for MCMC methods should be considered. We observe that the element-wise Metropolis-Hastings algorithm utilized in BHCR is prone to overfitting. To resolve the problem, based on the hierarchy in [7], we propose a block Metropolis-Hastings algorithm using a mixture of block moves [8] for more flexible and effective updates. For performance evaluation, we applied BAM to

978-0-7695-4574-5/11 $26.00 © 2011 IEEE
DOI 10.1109/BIBM.2011.81

both simulated and real LC-MS datasets and compared the results with BHCR.

The remainder of this paper is organized as follows. Section II introduces the methodology of the proposed Bayesian alignment model (BAM). The hierarchy, parameter inference, and the block Metropolis-Hastings algorithm are described in this section. Section III demonstrates applications of BAM to both simulated and real LC-MS data. Performances of BAM and BHCR are also compared in this section. Finally, we conclude this paper with a summary and possible extensions for future works in Section IV.

II. Model Formulation

A Bayesian alignment model (BAM) is proposed for LC-MS data alignment. We address the alignment problem within a Bayesian framework. The inference is drawn by Markov chain Monte Carlo methods which estimate the posterior distribution of model parameters.

A. Bayesian Alignment Model

BAM is a generative model that performs retention time alignment based on multiple total ion chromatograms (TICs) of LC-MS runs. The observed TICs from the same group,

$$y_i(t),\ i = 1, \ldots, N \text{ and } t = t_1, \ldots, t_T,$$

are assumed to share a similar profile characterized by the prototype function $m(t)$. We use a piecewise linear function to model the nonlinear variability along retention time. For the i-th TIC at retention time t, the intensity value is referred to as the prototype function indexed by the mapping function $u_i(t)$, i.e., $m\left(u_i(t)\right)$.

By incorporating the variability of intensity using affine transformation, each TIC is modeled as:

$$y_i(t) = c_i + a_i \cdot m\left(u_i(t)\right) + \epsilon_i,\ i = 1, 2, \ldots, N, \qquad (1)$$

where a_i and c_i are the scaling and translation variables of intensity, and the error ϵ_i is independent and identically distributed on a normal distribution $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\epsilon^2)$. These parameters characterize the individual variability of each TIC, and conjugate normal prior distributions are chosen, i.e., $c_i \sim \mathcal{N}(c_0, \sigma_c^2)$, and $a_i \sim \mathcal{N}(a_0, \sigma_a^2)$.

The prototype function is modeled with B-spline regression:

$$\boldsymbol{m} = \boldsymbol{B}_m \boldsymbol{\psi}, \qquad (2)$$

where $\boldsymbol{m} = (m(t_1), \ldots, m(t_T))^\top \in \mathbb{R}^{T\times 1}$, $\boldsymbol{B}_m \in \mathbb{R}^{T\times L}$, and $\boldsymbol{\psi} \in \mathbb{R}^{L\times 1}$. The number of spline basis functions L depends on the specification of knots and order of splines. The regression coefficients for the prototype function, $\boldsymbol{\psi}$, are specified by the first-order random walk: $\psi_l \sim \mathcal{N}(\psi_{l-1}, \sigma_\psi^2)$, where $\psi_0 = 0$. Given the value of σ_ψ^2 that controls the smoothness of $\boldsymbol{\psi}$, it can be shown that $\boldsymbol{\psi}$ follows a multivariate normal distribution $\boldsymbol{\psi}|\sigma_\psi^2 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\psi)$, where the inverse of the covariance matrix is a triple-diagonal matrix.

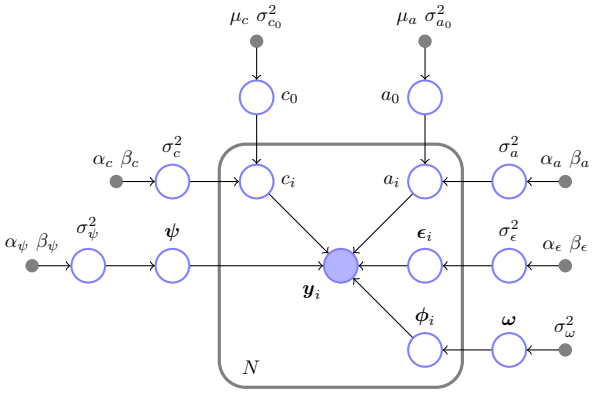


Figure 1. Directed acyclic graph of BAM

The mapping function $u_i(t)$ is a piecewise linear function characterized by knots $\boldsymbol{\tau} = (\tau_0, \tau_1, \ldots, \tau_{K+1})$ and their corresponding mapping indices $\boldsymbol{\phi}_i = (\phi_{i,0}, \phi_{i,1}, \ldots, \phi_{i,K+1})$, where $\tau_0 = t_1$ and $\tau_{K+1} = t_T$. The mapping function is defined in terms of $\boldsymbol{\tau}$ and $\boldsymbol{\phi}_i$,

$$u_i(t) = \begin{cases} \phi_{i,j} & \text{for } t = \tau_j \\ \frac{\tau_{j+1}-t}{\tau_{j+1}-\tau_j}\phi_{i,j} + \frac{t-\tau_j}{\tau_{j+1}-\tau_j}\phi_{i,j+1} & \text{for } \tau_j < t < \tau_{j+1} \end{cases} \qquad (3)$$

To keep the elution order of LC process, the monotonicity constraint needs to be satisfied, i.e., $\phi_{i,0} < \cdots < \phi_{i,K+1}$. The prior of $\boldsymbol{\phi}_i$ is specified via a slope value $\omega_{i,j} = (\phi_{i,j} - \phi_{i,j-1})/(\tau_j - \tau_{j-1})$. The slope value is assumed to follow a normal distribution with mean $\omega_{i,j-1}$ and variance σ_ω^2 truncated below by 0 to ensure monotonicity of $\boldsymbol{\phi}_i$.

Finally, we specify the priors for the other model parameters to complete the hierarchy: $c_0 \sim \mathcal{N}(\mu_c, \sigma_{c_0}^2)$, $\sigma_c^2 \sim \mathcal{IG}(\alpha_c, \beta_c)$, $a_0 \sim \mathcal{N}(\mu_a, \sigma_{a_0}^2)$, $\sigma_a^2 \sim \mathcal{IG}(\alpha_a, \beta_a)$, $\sigma_\epsilon^2 \sim \mathcal{IG}(\alpha_\epsilon, \beta_\epsilon)$, $\sigma_\psi^2 \sim \mathcal{IG}(\alpha_\psi, \beta_\psi)$. These priors are chosen to be conjugate to the likelihood function. Figure 1 presents the directed acyclic graph of BAM where the model parameters are denoted by open circles, the hyperparameters are denoted by solid dots, and the observations are denoted by filled circles.

B. Posterior Inference

Based on the generative model introduced in Section II-A, the alignment problem is translated to an inference task: given the TICs $\boldsymbol{y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$, estimate the model parameters $\left\{\boldsymbol{a}, \boldsymbol{c}, \boldsymbol{\psi}, \boldsymbol{\phi}, a_0, c_0, \sigma_a^2, \sigma_c^2, \sigma_\epsilon^2, \sigma_\psi^2\right\}$. Once the inference is complete, the alignment can be carried out by applying an inverse mapping function to each TIC, i.e., $\hat{y}_i(t) = y_i\left(\hat{u}_i^{-1}(t)\right)$.

The parameter inference is drawn using Markov chain Monte Carlo (MCMC) methods. For the parameters whose full conditionals are with closed-form, we use the Gibbs sampler to update their values iteratively. The remaining parameters, which are the mapping function coefficients

$\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_N\}$, are updated sequentially using the Metropolis-Hastings algorithm with a uniform proposal density that reflects the constraints on the boundaries.

C. Block Metropolis-Hastings Algorithm

When the misalignment involves translation shift along retention time, the element-wise Metropolis-Hastings move for $\boldsymbol{\phi}_i$ may appear cumbersome since a series of successive proposals in the same direction needs to be accepted sequentially. In addition, the monotonicity constraint inhibits the flexibility of the proposal for each of the mapping function coefficients. To address this issue, we consider the block move [8] to allow a batch of successive coefficients to be adjusted. Rather than updating each coefficient $\phi_{i,j}$ sequentially, we propose to group $\boldsymbol{\phi}_i$ into several non-overlapping blocks at the beginning and then update each block accordingly. Each block consists of successive coefficients along retention time. We introduce the indicator variable $b_j \in \{0, 1\}$, $j = 1, \ldots, K$ to determine the block boundary where $b_j = 1$ denotes τ_j is at the boundary of one block and $b_0 = b_{K+1} = 1$. The indicator variable follows a Bernoulli distribution: $p(b_j) = r_B^{b_j}(1 - r_B)^{1-b_j}$.

Based on the boundary configuration, coefficients within the same block are attempted to move in the same direction. We consider a mixture of transitions where r_B is randomly selected from $\{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$ at each iteration. The configuration of blocks is therefore variable within a MCMC chain. Two proposal step sizes (3 and 10) are considered to allow both small and large adjustments.

III. Results

We applied BAM to both simulated data and real LC-MS data. To demonstrate the effectiveness of the block Metropolis-Hastings algorithm, we compared the performance of BAM to the Bayesian hierarchical curve registration (BHCR) model proposed by Telesca and Inoue [7].

A. Simulated Dataset

We generated a profile pattern composed of three Gaussian shape peaks of the same standard deviation 1.25 but distinct mean values 15, 25 and 35. The mapping function u_i was generated through the even-numbered order statistics as described in [9], within the range $t = 1, \ldots, 50$, to model the retention time variability. Ten replicate runs were generated by the formulation:

$$y_i(t) = m\left(u_i(t)\right) + \epsilon_i, \;\; i = 1, 2, \ldots, 10,$$

where no retention time variation was applied to the first replicate run, i.e., $u_1(t) = t$. Random noises were produced based on signal to noise ratio values: 20, 25 and 30 dB.

The first replicate run is used as reference for assessing alignment performance. To assess the variability among replicate runs, we calculated the mean of correlation coefficients between the first replicate run and other replicate runs. In addition, to ensure the peak patterns were not significantly distorted during alignment, we calculated the ratio of overlapping peak areas based on the measurement

$$\frac{y_i(\hat{u}_i^{-1}(t))/y_1(\hat{u}_1^{-1}(t))}{y_i(u_i^{-1}(t))/y_1(u_1^{-1}(t))}$$

within the peak range, which reflects the degree of peak preservation compared to the perfect alignment result.

Table I summarizes the performance measurements before alignment and after alignment by BHCR and BAM. The difference between the two alignment methods is due to the capability to address the multimodal problem where BHCR is prone to getting stuck at local modes. In the next section, we demonstrate the trapping effect on a real LC-MS dataset. Misalignment resulting from the trapping effect can lead to wrong conclusion when proceeding with difference detection.

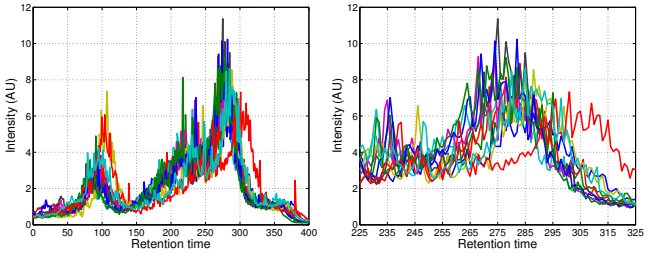
Table I
CORRELATION COEFFICIENTS AND RATIO OF OVERLAPPING PEAK AREAS FOR THE SIMULATED AND REAL LC-MS DATA, BEFORE ALIGNMENT (ORIGINAL) AND AFTER ALIGNMENT BY BHCR AND BAM. MEANS (STANDARD DEVIATIONS) ARE REPORTED FOR THE SIMULATED DATA BASED ON $1,000$ REALIZATIONS.

			After alignment	
		Original	BHCR	BAM
No noise	Corr	0.47 (0.096)	0.90 (0.083)	0.92 (0.066)
	Ratio	0.72 (0.076)	0.93 (0.068)	0.95 (0.059)
SNR 30	Corr	0.44 (0.094)	0.83 (0.090)	0.85 (0.079)
	Ratio	0.72 (0.078)	0.93 (0.069)	0.95 (0.064)
SNR 25	Corr	0.39 (0.087)	0.73 (0.099)	0.76 (0.078)
	Ratio	0.72 (0.079)	0.93 (0.071)	0.95 (0.061)
SNR 20	Corr	0.28 (0.074)	0.51 (0.112)	0.56 (0.092)
	Ratio	0.72 (0.084)	0.90 (0.088)	0.94 (0.068)
Real data	Corr	0.81	0.96	0.96

B. Real LC-MS Dataset

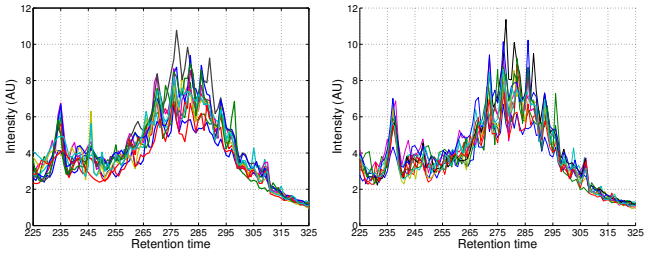
We applied BHCR and BAM to Listgarten's LC-MS dataset [10] that consists of 11 replicate LC-MS runs of proteins extracted from lysed *Escherichia coli* cells using a capillary-scale LC coupled to an ion trap mass spectrometer. Figure 2 depicts TICs of the 11 replicate LC-MS runs where significant shifts are along the RT points. We highlight the multimodal problem at RT range $225 - 325$ where many chromatographic peaks are observed. The alignment result by BHCR is shown in Fig. 3a. It is noted that two peaks at RT ~ 245 were not correctly aligned to the majority of peaks at RT ~ 235. Instead, they were mistakenly aligned to other tiny peaks around their original retention times. As mentioned in Section II-C, the element-wise Metropolis-Hastings move utilized by BHCR is prone to getting stuck at local modes. It is particularly hard to get away from the trap if the parameter values of interest are far from the current values.

We applied BAM to the same dataset (Fig. 3b). The inference was based on $15,000$ MCMC iterations obtained after



(a) Whole retention time range (b) Retention time $225 - 325$

Figure 2. TICs of the original LC-MS data.



(a) Alignment by BHCR (b) Alignment by BAM

Figure 3. TICs of the aligned LC-MS data at RT range $225 - 325$.

discarding the initial $20,000$ iterations as burn-in. Rather than localization around some local mode, the trapping effect is overcome by BAM that is effective to correct the significant variations at the beginning of the Markov chain (usually within 100 iterations based on our observation).

IV. Conclusion

In this paper, we propose a Bayesian alignment model (BAM) for LC-MS data analysis and use MCMC methods for parameter inference. Due to mathematical intractability and monotonicity constraint of the mapping function, designing an effective sampling scheme is crucial for better mixing of the MCMC runs. We propose a block Metropolis-Hastings algorithm that enables flexible transition and avoids the tendency to get trapped in some local mode. Evaluation on both simulated and real datasets shows satisfactory results in terms of correlation coefficients and ratio of overlapping peak areas as well as visual assessment.

In the future, there are a number of studies that deserve further investigation. Current setting of BAM is based on the profile of TICs. Identifying representative extracted ion chromatograms (EICs) provides more comprehensive information with respect to retention time variability of LC-MS runs, which can potentially improve the alignment effectiveness. Rather than naïve inclusion, we are interested in utilization of informative metric to select chromatograms for achieving better alignment results. The main assumption of the profile-based alignment models is that a consistent pattern is representative of multiple LC-MS runs from the same group. This assumption somewhat disregards the heterogeneity across subgroups that is a common phenomenon in biological samples. Extensions in consideration of heterogeneity are desired where appropriate components for alignment can be utilized accordingly. More importantly, for difference detection, the extension may enable simultaneous alignment of samples from multiple groups, which ensures processing coherence and data comparability.

Acknowledgment

This work was supported by the National Science Foundation Grant IIS-0812246 and the National Cancer Institute Grant R01CA143420.

References

[1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[2] D. Radulovic, S. Jelveh, S. Ryu, T. Hamilton, E. Foss, Y. Mao, and A. Emili, "Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry," *Molecular and Cellular Proteomics*, vol. 3, no. 10, pp. 984–997, 2004.

[3] Y. V. Karpievitch, A. D. Polpitiya, G. A. Anderson, R. D. Smith, and A. R. Dabney, "Liquid chromatography mass spectrometry-based proteomics: Biological and technological aspects," *The Annals of Applied Statistics*, vol. 4, no. 4, pp. 1797–1823, 2010.

[4] M. Vandenbogaert, S. Li-Thiao-T, H.-M. Kaltenbach, R. Zhang, T. Aittokallio, and B. Schwikowski, "Alignment of LC-MS images, with applications to biomarker discovery and protein identification," *Proteomics*, vol. 8, pp. 650–672, 2008.

[5] J. D. Jaffe, D. R. Mani, K. C. Leptos, G. M. Church, M. A. Gillette, and S. A. Carr, "PEPPeR, a platform for experimental proteomic pattern recognition," *Molecular and Cellular Proteomics*, vol. 5, no. 10, pp. 1927–1941, 2006.

[6] N. Jaitly, M. E. Monroe, V. A. Petyuk, T. R. W. Clauss, J. N. Adkins, and R. D. Smith, "Robust algorithm for alignment of liquid chromatographymass spectrometry analyses in an accurate mass and time tag data analysis pipeline," *Analytical Chemistry*, vol. 78, no. 21, pp. 7397–7409, 2006.

[7] D. Telesca and L. Y. T. Inoue, "Bayesian hierarchical curve registration," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 328–339, 2008.

[8] G. O. Roberts and S. K. Sahu, "Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, vol. 59, pp. 291–317, 1997.

[9] P. J. Green, "Reversible-jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, vol. 82, no. 4, p. 711732, 1995.

[10] J. Listgarten, R. M. Neal, S. T. Roweis, P. Wong, and A. Emili, "Difference detection in LC-MS data for protein biomarker discovery," *Bioinformatics*, vol. 23, no. 2, pp. e198–e204, 2007.