31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Bayesian Based Functional Mixed-Effects Model for Analysis of LC-MS Data

Getachew K Befekadu, Mahlet G Tadesse, and Habtom W Ressom*, *Senior Member, IEEE*

***Abstract*—A Bayesian multilevel functional mixed-effects model with group specific random-effects is presented for analysis of liquid chromatography-mass spectrometry (LC-MS) data. The proposed framework allows alignment of LC-MS spectra with respect to both retention time (RT) and mass-to-charge ratio (*m/z*). Affine transformations are incorporated within the model to account for any variability along the RT and *m/z* dimensions. Simultaneous posterior inference of all unknown parameters is accomplished via Markov chain Monte Carlo method using the Gibbs sampling algorithm. The proposed approach is computationally tractable and allows incorporating prior knowledge in the inference process. We demonstrate the applicability of our approach for alignment of LC-MS spectra based on total ion count profiles derived from two LC-MS datasets.**

I. Introduction

In proteomic studies, liquid chromatography coupled with mass spectrometry (LC-MS) is a common platform to identify and determine the abundance of various peptides that characterize particular proteins in biological samples [1]. Each LC-MS run generates data comprised of thousands of peak intensities for peptides with specific retention time (RT) and mass-to-charge ratio (*m/z*) values. In differential protein expression studies, multiple LC-MS runs are compared to identify differentially abundant peptides between distinct biological groups. This is a challenging task because of the following reasons: (1) substantial variation in RT across multiple runs due to the LC instrument conditions and the variable complexity of peptide mixtures, (2) variation in *m/z* values of the peptides due to occasional drift in the calibration of the mass spectrometry instrument, and (3) variation in peak intensities due to spray conditions. Thus, efficient and robust alignment algorithms are needed for qualitative comparison of multiple LC-MS runs.

This work was supported in part by the National Science Foundation Grant IIS-0812246, the National Cancer Institute (NCI) R21CA130837 Grant, NCI R03CA119313 Grant, NCI Early Detection Research Network Associate Membership Grant, and the Prevent Cancer Foundation Grant awarded to HWR.

G. K. Befekadu is with the Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, USA.

M. G. Tadesse is with the Department of Mathematics, Georgetown University, 308 St. Mary's Hall, Washington, DC 20057, USA.

H. W. Ressom is with the Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, USA.

*Corresponding author: hwr@georgetwon.edu.

Various alignment methods have been described in literature including dynamic time warping (DTW) [2], correlation optimized warping (COW) [2], vectorized peaks [3], statistical alignment [4], and clustering [5]. Most of these algorithms are either limited to a consensus pair-wise combination of spectra for alignment or may use reference (template) spectra to find matching among datasets. These limitations may lead to sub-optimal results compared to global alignment techniques. Methods that rely on optimization of global fitting functions provide an alternative solution to alignment of multiple LC-MS spectra representing distinct biological groups. For example, a recently introduced method called continuous profile model (CPM) has been applied for alignment of continuous time-series data and for detection of differences in multiple LC-MS data [6]. Although CPM is described as a naïve and computationally intensive method, the method has some limitations, such as the susceptibility to fall into local minimum solutions due to the sub-optimal problem formulation. Also, the method creates superfluous signal gaps, leading to nonuniform trace points across multiple LC-MS spectra. Another notable limitation of CPM algorithm is its poor performance with time complexity scales, requiring substantial computation time in modeling high resolution data. Thus, CPM is more suitable for low resolution of LC-MS data generated from less complex fractionations. Recently, Morris et al. developed a Bayesian-based method for analysis of matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) proteomics data [7]. Their motivation extends from earlier work on Bayesian implementation of the wavelet-based functional mixed effects models introduced by Morris and Carroll [8]. The approach is similar to the spline-based functional mixed effects models introduced by Guo [9], which involves a generalized mixed models equation to handle potentially irregular data. The method specifically deals with the identification of differentially expressed spectral regions across different experimental conditions assuming the alignment issue has already been taken care of.

In this paper, we introduce a Bayesian multilevel functional mixed effects model with group-specific random effects. The method provides the capability to account for population homogeneous behavior (i.e., fixed systematic changes across the entire LC-MS spectra representing distinct biological groups) while allowing for modeling heterogeneity within a group (i.e., random effects). Also, this

paradigm allows us to incorporate additional hierarchies such as affine transformation within the model to account for any variability along the RT and *m/z* dimensions, while handling implicitly the normalization of peak intensities of peptides from multiple LC-MS spectra. The method is amenable to model both low and high resolution mass spectra, since it does not introduce superfluous signal gaps across multiple LC-MS spectra. We demonstrate this through two LC-MS datasets obtained from: (1) proteins of *lysed E.coli* cells, and (2) six groups of tryptic digests non-human proteins with different concentrations spiked into a complex sample background of human peptides.

The remainder of this paper is organized as follows. In Section II, we outline the Bayesian hierarchical model (BHM) that describes the data modeling mechanism, based on the functional mixed-effects model, for alignment of LC-MS spectra. This section explains the Markov chain Monte Carlo (MCMC) method using the Gibbs sampling algorithm for simultaneous posterior inference of all unknown parameters. Results and discussions demonstrating the applicability of the proposed method for alignment of LC-MS spectra are given in Section III. Finally, our findings are summarized in Section IV.

II. METHODS

A. Bayesian Hierarchical Model (BHM)

We propose a functional mixed-effects model to align LC-MS spectra from multiple LC-MS runs. The idea behind this approach is two-fold: (1) to model the fixed effects as a realization of partially diffused integrated Gaussian processes which account for population homogeneous behaviors (i.e., fixed systematic changes in the LC-spectra across biological groups), and (2) to model the random effects as random realizations from the same partially integrated Gaussian processes with proper variances which, in turn, allow the modeling of heterogeneity within biological groups. The estimation procedure is implemented by taking advantage of the connection between B-splines (at the design points) and mixed effects models. Let the proposed functional mixed-effects model be represented mathematically as follows:

$$\underbrace{\mathbf{y}_i \mid \{z_{ij}=j\}}_{\mathbf{n}_i\times 1} \equiv \underbrace{\mathbf{B}_{1i}}_{\mathbf{n}_i\times p}\underbrace{\boldsymbol{\gamma}_j}_{p\times 1}+\underbrace{\mathbf{B}_{2i}}_{n_i\times q}\underbrace{\boldsymbol{\eta}_{ij}}_{q\times 1}+\underbrace{\boldsymbol{\varepsilon}_{ij}}_{n_i\times 1} \qquad (1)$$
$$\boldsymbol{\eta}_{ij}\sim N_q(\mu_j,\boldsymbol{\Sigma}_j) \text{ and } \boldsymbol{\varepsilon}_{ij}\sim N_{n_i}(0,\sigma^2\mathbf{I})$$

where $i = 1, 2,\ldots s_j$ denote sample size in each group j; $j = 1, 2,\ldots, m$ are indices for the mixture components that identify group membership with $m << M$, where M is the total number of observation spectra ($M=\sum_{j=1}^{m} s_j$); n_i denotes the data length for each spectrum $\mathbf{y}_i$; $\mathbf{B}_{1i}$ and $\mathbf{B}_{2i}$ are B-spline basis matrices associated with fixed and random effects for the i-th sample, respectively; $\boldsymbol{\gamma}_j$ accounts for fixed systematic changes in group j; while $\boldsymbol{\eta}_{ij}$ accounts for the random variation; $\boldsymbol{\varepsilon}_{ij}$'s correspond to measurement errors, where $\boldsymbol{\eta}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ are assumed to be independent. Alignment is performed using the information from the matrices $\mathbf{B}_{1i}$ and $\mathbf{B}_{2i}$ as well as from fixed systematic changes $\boldsymbol{\gamma}_j$ and random variation $\boldsymbol{\eta}_{ij}$.

Let $\boldsymbol{\Theta}$ be a vector consisting of all the unknown parameters in Eq. (1) and the priors. Let $\mathbf{Y}=(\mathbf{y}_1^T,\mathbf{y}_2^T,\ldots,\mathbf{y}_M^T)^T$ represent a set of LC-MS spectra. Then, according to the Bayes' theorem

$$p(\boldsymbol{\Theta}\mid\mathbf{Y})=\frac{p(\mathbf{Y}\mid\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{p(\mathbf{Y})}\propto L(\boldsymbol{\Theta}\mid\mathbf{Y})\times p(\boldsymbol{\Theta}) \qquad (2)$$

Using the functional mixed-effects modeling of Eq. (1), the likelihood function assuming that the group information is known and that the samples are independent is given by

$$L(\boldsymbol{\Theta}\mid\mathbf{Y},\mathbf{Z})=\prod_{j=1}^{m}\prod_{i=1}^{s_j}N_n(\mathbf{y}_i,\mathbf{Z}_i;\mathbf{B}_{1i}\boldsymbol{\gamma}_j+\mathbf{B}_{2i}\boldsymbol{\eta}_{ij},\sigma^2\mathbf{I}) \qquad (3)$$

where $\mathbf{Z}$ denotes a matrix of indicator vectors $\mathbf{Z}_i=(z_{i1},z_{i2},\ldots,z_{im})$, such that $z_{i1}=1$ for some j, and $z_{it}=0,\ \forall t\neq j$. Hence, the joint posterior has the general form

$$p(\boldsymbol{\Theta}\mid\mathbf{Y})\propto\prod_{j=1}^{m}\prod_{i=1}^{s_j}N_n(\mathbf{y}_i,\mathbf{Z}_i;\mathbf{B}_{1i}\boldsymbol{\gamma}_j+\mathbf{B}_{2i}\boldsymbol{\eta}_{ij},\sigma^2\mathbf{I})\times p(\boldsymbol{\Theta}) \qquad (4)$$

B. Prior Distributions and Implementation

The first step in fitting BHM is to specify all prior distributions. A list of the hierarchical priors assigned to the parameters of the model is given below. The list represents the standard choice of priors for mixture models:

$$\begin{aligned}&\boldsymbol{\gamma}_j\sim N_p(w,\mathbf{W})\\ &\boldsymbol{\mu}_j\sim N_q(0,\mathbf{V}),\ \ \Sigma_j^{-1}\mid\mathbf{R}\sim W_q(\rho,(\rho\mathbf{R})^{-1})\\ &\mathbf{R}\sim W_q(r,(r\mathbf{R}_0)^{-1}),\ \ \sigma^{-2}\sim\Gamma(g,h)\end{aligned} \qquad (5)$$

where $W_{\cdot}(,)$, $N_{\cdot}(,)$ and $\Gamma(,)$ signify the Wishart, multivariate normal and gamma distributions, respectively. In specifying the prior distribution $p(\boldsymbol{\Theta})$, a hierarchical structure with independence assumption is considered. Combining this structural information with prior beliefs, we obtain the following joint posterior for the unknown parameters:

$$\begin{aligned}p(\boldsymbol{\Theta}\mid\mathbf{Y})\propto\prod_{j=1}^{m}\{&\prod_{i=1}^{s_j}N_n(\mathbf{y}_i,\mathbf{Z}_i;\mathbf{B}_{1i}\boldsymbol{\gamma}_j+\mathbf{B}_{2i}\boldsymbol{\eta}_{ij},\sigma^2\mathbf{I})\times N_p(\boldsymbol{\gamma}_j;w,\mathbf{W})\times\\ &N_q(\boldsymbol{\eta}_{ij};\boldsymbol{\mu}_{z_{ij}},\Sigma_{z_{ij}})\times N_q(\boldsymbol{\mu}_j;0,\mathbf{V})\times W_q(\Sigma_j^{-1};\rho,(\rho\mathbf{R})^{-1})\}\times\\ &W_q(\mathbf{R};r,(r\mathbf{R}_0)^{-1})\times\Gamma(\sigma^{-2};g,h)\end{aligned} \qquad (6)$$

Using all prior and hyperprior distributions in Eq. (5), the full conditional distributions for the parameters are as follows:

$$p(\boldsymbol{\eta}_{ij}\mid\text{rest})\propto N_q([\sigma^{-2}\mathbf{B}_{2i}^T\mathbf{B}_{2i}+\Sigma_j^{-1}]^{-1}(\sigma^{-2}\mathbf{B}_{2i}^T(\mathbf{y}_i-\mathbf{B}_{2i}\boldsymbol{\gamma}_j)+\Sigma_j^{-1}\boldsymbol{\mu}_j),[\sigma^{-2}\mathbf{B}_{2i}^T\mathbf{B}_{2i}+\Sigma_j^{-1}]^{-1})$$

$$p(\boldsymbol{\gamma}_j\mid\text{rest})\propto N_p([\sigma^{-2}\textstyle\sum_{i:z_{ij}=1}\mathbf{B}_{1i}^T\mathbf{B}_{1i}+\mathbf{W}^{-1}]^{-1}(\sigma^{-2}\sum_{i:z_{ij}=1}\mathbf{B}_{1i}^T(\bar{\mathbf{y}}_i-\mathbf{B}_{2i}\bar{\boldsymbol{\eta}}_i)+\mathbf{W}^{-1}w),[\sigma^{-2}\textstyle\sum_{i:z_{ij}=1}\mathbf{B}_{1i}^T\mathbf{B}_{1i}+\mathbf{W}^{-1}]^{-1})$$

$$p(\boldsymbol{\mu}_j\mid\text{rest})\propto N_q(s_j\Sigma_j^{-1}+\mathbf{V}^{-1})^{-1}s_j\Sigma_j^{-1}\bar{\boldsymbol{\eta}}_j,[s_j\Sigma_j^{-1}+\mathbf{V}^{-1}]^{-1}) \quad \text{where } \bar{\boldsymbol{\eta}}_j=\textstyle\sum_{i:z_{ij}=1}\boldsymbol{\eta}_{ij}$$

$$p(\Sigma_j^{-1}\mid\text{rest})\propto W_q(s_j+\rho,[r\mathbf{R}+\textstyle\sum_{i:z_{ij}=1}(\boldsymbol{\eta}_{ij}-\boldsymbol{\mu}_j)(\boldsymbol{\eta}_{ij}-\boldsymbol{\mu}_j)^T]^{-1})$$

$$p(\mathbf{R}\mid\text{rest})\propto W_q(r+m\rho,[r\mathbf{R}_0+\rho\textstyle\sum_{j=1}^{m}\Sigma_j^{-1}]^{-1})$$

$$p(\sigma^{-2}\mid\text{rest})\propto\Gamma(\frac{\sum_{j=1}^{m}\sum_{i=1}^{s_j}n_i}{2}+g,[\frac{1}{h}+\frac{1}{2}\textstyle\sum_{j=1}^{m}\sum_{i:z_{ij}=1}(\mathbf{y}_i-\mathbf{B}_{1i}\boldsymbol{\gamma}_j-\mathbf{B}_{2i}\boldsymbol{\eta}_{ij})(\mathbf{y}_i-\mathbf{B}_{1i}\boldsymbol{\gamma}_j-\mathbf{B}_{2i}\boldsymbol{\eta}_{ij})^T]^{-1})$$

C. Gibbs Sampling Algorithm

Consider the Bayesian model of Eq. (4). Let the number of groups m be fixed and $\boldsymbol{\Theta}$ denote all of the unknown parameters in the model, i.e.,

$$\boldsymbol{\Theta} = \left(\left\{ \left\{ \boldsymbol{\eta}_{ij} \right\}_{i=1}^{s_j}, \boldsymbol{\gamma}_j, \boldsymbol{\mu}_j, \Sigma_j \right\}_{j=1}^{m}, \mathbf{R}, \sigma^2 \right)$$

Then, using $\boldsymbol{\Theta}^{(0)}$ as starting value, the Gibbs sampling algorithm [10, 11] proceeds as follows for t = 1,2, ..., N iterations:

Draw $\boldsymbol{\eta}_{ij}^{(t+1)}$ from $p(\boldsymbol{\eta}_{ij} \mid \mathbf{Y}, \boldsymbol{\gamma}_j^{(t)}, \ldots, \sigma^{2(t)})$ for i = 1,2,...,s_j and j =1,2,...,m

Draw $\boldsymbol{\gamma}_j^{(t+1)}$ from $p(\boldsymbol{\gamma}_j \mid \mathbf{Y}, \boldsymbol{\eta}_{ij}^{(t+1)}, \ldots, \sigma^{2(t)})$ for j = 1,2,..., m

Draw $\boldsymbol{\mu}_j^{(t+1)}$ from $p(\boldsymbol{\eta}_{ij} \mid \mathbf{Y}, \boldsymbol{\eta}_{ij}^{(t+1)}, \boldsymbol{\gamma}_j^{(t+1)}, \ldots, \sigma^{2(t)})$ for j = 1,2,..., m

$\vdots$

Draw $\sigma^{2(t+1)}$ from $p(\sigma^2 \mid \mathbf{Y}, \boldsymbol{\eta}_{ij}^{(t+1)}, \boldsymbol{\gamma}_j^{(t+1)}, \boldsymbol{\mu}_j^{(t+1)}, \Sigma_j^{(t+1)}, \mathbf{R}^{(t+1)})$

Note that the computations for the conditional probabilities are highly simplified due to the conjugacy of the prior distributions and their conditional independence.

D. Modeling of Variability along the RT or m/z Dimensions among Different LC-MS Datasets

The BHM presented in Eq. (1) can be easily extended to incorporate detail modeling. It is important to introduce priors that appropriately apportion the variability among the replicates and separating out the differing locations or scales along the RT or *m/z* dimensions. This provides a distinct interpretation of the LC-MS data. The alignment model and the associated parameters should allow each replicate sample to have its own affine warping transformation in RT or *m/z* dimensions. Let each spectrum $\mathbf{y}_i(\mathbf{x}_{ij})$ be replaced by $\mathbf{y}_i^{(d)}(a_i \mathbf{x}_{ij} - b_i)$, where a_i and b_i are the scaling and shifting parameters for the i-th replicate LC-MS spectrum along the dimension d=1,2, ..., D corresponding to the *m/z* dimensions of each sample spectra. To treat alignment in the hierarchical structure, we include the priors with suitable hyperparametrs in Eq. (6). With the assumption of independence, the joint prior model for the time scaling and translation $p(a_i, b_i) = p(a_i) \times p(b_i)$ should encode the idea that the most likely translation is the affine warping translation and should also discount large and unlikely translations. A normal prior distribution is a good fit for this, i.e., $b_i \sim N(\mu_b, \sigma_b^2)$ where μ_b and σ_b^2 are the mean and variance of the hyperparameters. Moreover, we assume a normal distribution $a_i \sim N(\mu_a, \sigma_a^2) I\{a_i > 1\}$ for the time-scaling prior a_i, since its most likely values for the mean and variance could be captured from the data. Therefore, the parameters μ_a, μ_b, σ_a^2 and σ_b^2 are estimated from the data within the ensuing MCMC algorithm. The corresponding directed acyclic graph (DAG) in Fig. 1 shows the dependences of all hierarchical parameters in the model.

Combining the affine warping transformation with our prior beliefs for a_i and b_i, the posterior distribution for the unknown parameters is modified as follows:

$$\begin{aligned} p(\boldsymbol{\Theta}/\mathbf{Y}) \propto & \prod_{d=1}^{D} \prod_{j=1}^{m} \{ \prod_{i=1}^{s_j} N_n(\mathbf{y}_i^{(d)}, \mathbf{Z}_i; \mathbf{B}_{1i}^{(d)} \boldsymbol{\gamma}_j^{(d)} + \mathbf{B}_{2i}^{(d)} \boldsymbol{\eta}_{ij}^{(d)}, \sigma^{2(d)} \mathbf{I}) \times \\ & N_p(\boldsymbol{\gamma}_j^{(d)}; w^{(d)}, \mathbf{W}^{(d)}) \times N_q(\boldsymbol{\eta}_{ij}^{(d)}; \boldsymbol{\mu}_{z_{ij}}^{(d)}, \Sigma_{z_{ij}}^{(d)}) \times N(\mu_b; \sigma_b^2) \times \\ & N(\mu_a; \sigma_a^2) I\{a_i > 1\} \times N_q(\boldsymbol{\mu}_j^{(d)}; 0, \mathbf{V}^{(d)}) \times W_q(\Sigma_j^{-1(d)}; \rho^{(d)}, (\rho^{(d)} \mathbf{R}^{(d)})^{-1}) \} \times \\ & W_q(\mathbf{R}^{(d)}; r, (r\mathbf{R}_0)^{-1}) \times \Gamma(\sigma^{-2(d)}; g, h) \end{aligned} \tag{7}$$

where $\boldsymbol{\Theta}$ denotes all unknown parameters in this new model.

$$\boldsymbol{\Theta} = \left(\left\{ \left\{ \left\{ \boldsymbol{\eta}_{ij}^{(d)} \right\}_{i=1}^{s_j}, \boldsymbol{\gamma}_j^{(d)}, \boldsymbol{\mu}_j^{(d)}, \Sigma_j^{(d)}, \mathbf{R}^{(d)}, \sigma^{2(d)} \right\}_{d=1}^{D}, \mu_a, \mu_b, \sigma_a^2, \sigma_b^2 \right\}_{j=1}^{m} \right) \tag{8}$$

The B-spline basis matrices associated with $\mathbf{y}_i^{(d)}(a_i \mathbf{x}_{ij} - b_i)$ need to be updated at each iteration based on the estimates of RT transformation parameters a_i and b_i. Moreover, these parameters need to be shared over the D dimensions for each group data since the dynamic behavior for each dimension occurs over the same time scale. The Gibbs sampling algorithm for the modified BHM of Eq. (7) proceeds in the same manner as that of Section C. The algorithm continues to draw the other parameters in the order outlined below:

$\boldsymbol{\eta}_{ij}^{(d)(t+1)}, \boldsymbol{\gamma}_j^{(d)(t+1)}, \boldsymbol{\mu}_j^{(d)(t+1)}, \Sigma_j^{(d)(t+1)}, \mathbf{R}^{(d)(t+1)}, \sigma^{2(d)(t+1)}, \mu_a^{(t+1)}, \mu_b^{(t+1)},$

$\sigma_a^{2(t+1)}, \sigma_b^{2(t+1)}$, for $i = 1, 2, \ldots, s_j$, $j = 1, 2, \ldots, m$ and $d = 1, 2, \ldots, D$

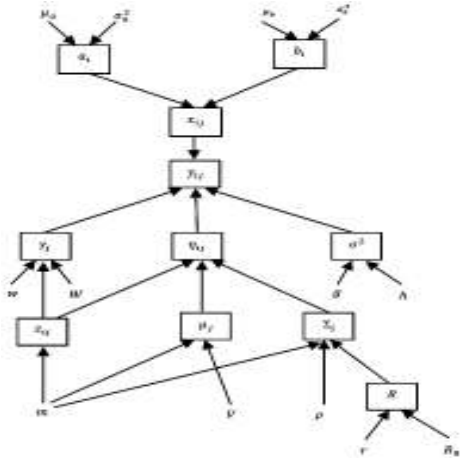

Fig. 1. DAG of the Bayesian hierarchical model

III. RESULTS AND DISCUSSIONS

We used BHM to align 11 replicate LC-MS spectra obtained from http://www.cs.toronto.edu/~jenn/LCMS. The spectra are generated from proteins of *lysed E.coli* cells by capillary-scale LC coupled on-line to an ion trap mass spectrometer (see Listgarten et al. [6] for details). Each spectrum was represented by two dimensions after calculating the total ion count (TIC) profiles for each RT point across the *m/z* values from the original 400×2400 data matrix corresponding to 400 RT points (~55 min.) and 2400 *m/z* bins spanning between 400 and 1600 Dalton (Da). Fig. 2 depicts these 11 two-dimensional replicate spectra. From this figure, we can see that the spectra show significant shifts along RT as well as distortions in the ion abundance measurement space. We applied our BHM method for alignment of LC-MS spectra with respect to RT. Fig. 3 depicts the aligned spectra. BHM reduced the coefficient of variation (CV) of the original TIC profiles from 82% to 66%. The CV of the spectra aligned by DTW, COW and CPM were 70%, 80% and 57%, respectively.

The second dataset was obtained from http://prottools.ethz.ch/muellelu/web/Latin_Square_Data.php It consists of 18 LC-MS spectra generated from tryptic digests of six standard non-human proteins (myoglobin, carbonic anhydrase, cytochrome c, lysozyme, alcohol dehydrogenase, and aldolase A) spiked with different concentrations into a complex sample background of human peptides and isolated by solid-phase Nglycocapture from serum. The LC-MS spectra generation for these samples was performed using the Fourier transformed-linear trap quadrupole (FT-LTQ) mass spectrometer (see Mueller et al. [5] for details). The 18 spectra represent six groups based on the concentration of the proteins. We processed the raw spectra and obtained for each spectrum a 2000x1300 data

matrix corresponding to 2000 RT points (~55 min.) and 1300 m/z bins between 300 and 1600 Da. We calculated the TIC for each RT point across the *m/z* values and obtained 18 two-dimensional TIC profiles (for each of the six groups). Figs. 4 and 5 depict TIC plots of the original and aligned LC-MS spectra, respectively. Fig. 6 shows the corresponding heat maps for the original and aligned LC-MS spectra. BHM reduced the average CV of the original TIC profile across the six groups from 18% to 13%. Both DTW and COW yielded a CV of 17%, while CPM resulted in a CV of 13%.

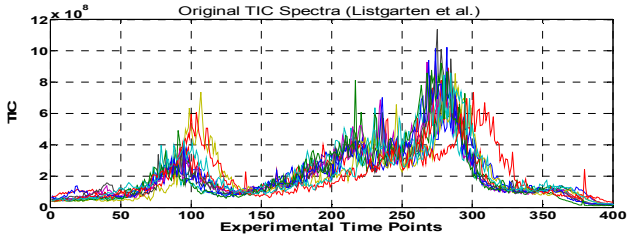

Fig. 2. Plots of TIC profiles before alignment

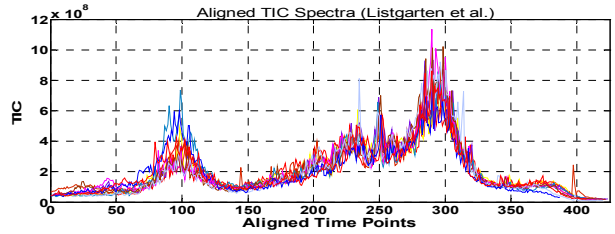

Fig. 3. TIC profiles after alignment by BHM.

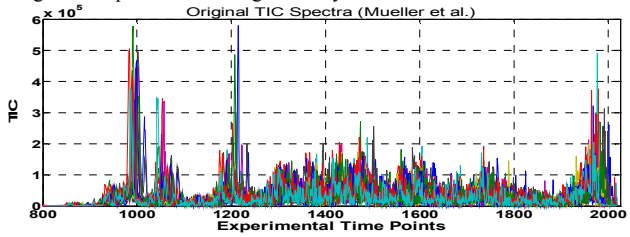

Fig. 4. TIC profiles before alignment.

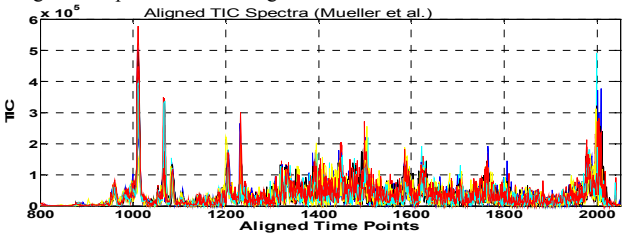

Fig. 5. TIC profiles after alignment by BHM.

IV. Conclusion

This paper utilizes a Bayesian hierarchical model for alignment of LC-MS spectra. Specifically, it presents a fully Bayesian mixed-effects model that effectively accounts for population homogeneous behavior across biological groups (i.e., fixed systematic changes) and for heterogeneity within groups (random effects). Bayesian inference of unknown parameters is carried out via MCMC method using the Gibbs sampling technique with conjugate priors. The proposed approach not only allows alignment with respect to RT and *m/z* dimensions, it also implicitly normalizes the peak intensities of peptides. The performance of the approach is assessed through two LC-MS datasets: replicate spectra generated from proteins of *lysed E.coli* cells and spectra representing six groups, where six proteins are spiked at different concentrations into a complex sample background of human peptides. Through these datasets, it is demonstrated that BHM achieves good performance in reducing coefficient of variation of replicate TIC profiles, while preserving the original experimental retention time (i.e., without introducing superfluous signal gaps across multiple LC-MS spectra). A limitation of BHM is that it requires considerable amount of computation time in aligning LC-MS data with respect to both RT and *m/z* dimensions. Future work will focus on addressing this limitation through optimization of the algorithm.

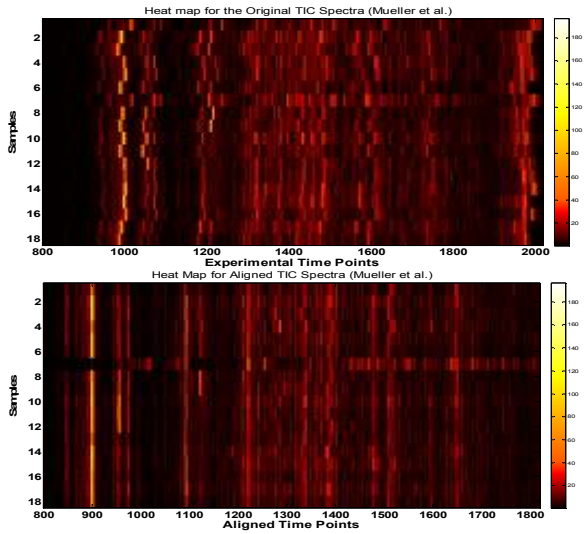

Fig. 6. Heat maps of the TIC profiles for Mueller et al. dataset.

References

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

[2] G. Tomasi, F. van den Berg, and C. Andersson, "Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data," *Journal of Chemometrics and Intelligent Laboratory Systems*, vol. 18, pp. 231-241, 2004.

[3] C. A. Hastings, S. M. Norton, and S. Roy, "New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data," *Rapid Commun Mass Spectrom*, vol. 16, No. 5, pp. 462-7, 2002.

[4] P. Wang, H. Tang, M. P. Fitzgibbon, et al., "A statistical method for chromatographic alignment of LC-MS data," *Biostatistics*, vol. 8, No. 2, pp. 357-67, 2007.

[5] L. N. Mueller, O. Rinner, A. Schmidt, et al., "SuperHirn - a novel tool for high resolution LC-MS-based peptide/protein profiling," *Proteomics*, vol. 7, No. 19, pp. 3470-80, 2007.

[6] J. Listgarten, "Analysis of sibling time series data: alignment and difference detection," in *Department of Computer Science*, vol. Ph.D. . Toronto: University of Toronto, 2006

[7] J. S. Morris, P. J. Brown, R. C. Herrick, et al., "Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models," *Biometrics*, vol. 64, No. 2, pp. 479-89, 2008.

[8] J. S. Morris and R. J. Carroll, "Wavelet-based functional mixed models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, pp. 179-199, 2006.

[9] W. Guo, "Functional data analysis in longitudinal settings using smoothing splines," *Statistical Methods in Medical Research*, vol. 13, No. 1, pp. 49-62, 2004.

[10] G. Casella and E. I. George, "Explaining the Gibbs Sampler," *The American Statistician*, vol. 46, No. 3, pp. 167-174, 1992.

[11] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," in *Readings in uncertain reasoning*: Morgan Kaufmann Publishers Inc., pp. 452-472, 1990.