Published in final edited form as:

IEEE/ACM Trans Comput Biol Bioinform. 2011 ; 8(5): 1417–1424. doi:10.1109/TCBB.2010.88.

Probabilistic Mixture Regression Models for Alignment of LC-MS Data

Getachew K. Befekadu,

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC 20057. gkb8@georgetown.edu

Mahlet G. Tadesse,

Department of Mathematics, Georgetown University, Washington, DC 20057. mgt26@georgetown.edu

Tsung-Heng Tsai, and

Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Arlington, VA 22203. thtsai@vt.edu

Habtom W. Ressom

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC 20057. hwr@georgetown.edu

Abstract

A novel framework of a probabilistic mixture regression model (PMRM) is presented for alignment of liquid chromatography-mass spectrometry (LC-MS) data with respect to both retention time (RT) and mass-to-charge ratio (m/z). The expectation maximization algorithm is used to estimate the joint parameters of spline-based mixture regression models and prior transformation density models. The latter accounts for the variability in RT points, m/z values, and peak intensities. The applicability of PMRM for alignment of LC-MS data is demonstrated through three datasets. The performance of PMRM is compared with other alignment approaches including dynamic time warping, correlation optimized warping, and continuous profile model in terms of coefficient variation of replicate LC-MS runs and accuracy in detecting differentially abundant peptides/proteins.

Keywords

liquid chromatography; mass spectrometry; mixed-regression model; expectation-maximization

1 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 consisting 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 due to occasional drift in the calibration of the mass spectrometry instrument, and 3) variation in peak intensities due to spray conditions (in most cases this is proportional to concentration of peptides in the sample). Thus,

alignment with respect to both RT and m/z and normalization of peak intensities are a prerequisite for quantitative comparison of multiple LC-MS runs.

Various alignment methods have been described in the literature including dynamic time warping (DTW) [2], correlation optimized warping (COW) [2], vectorized peaks [3], (semi-) supervised alignment using nonlinear regression methods [4], statistical alignment [5], and clustering [6]. Most of these algorithms are either limited to a consensus pair-wise combination of LC-MS runs or use a reference (template) for alignment. 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 runs 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 [7], [8]. Although CPM is described as a straightforward and computationally intensive method, the method has limitations. For example, it creates superfluous signal gaps, leading to non-uniform trace points across multiple LC-MS runs. Also, it is susceptible to fall into local minimum solutions due to the sub-optimal problem formulation. Another limitation of CPM is its poor performance with time complexity scales, requiring substantial computation time in modeling high resolution data. Thus, CPM is more suitable for alignment of low resolution LC-MS data generated from less complex fractionations.

Recently, Morris et al. developed a Bayesian method for analysis of matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) proteomic data [9]. Their motivation extends from earlier work on Bayesian implementation of the wavelet-based functional mixed effects model [10]. The approach is similar to the spline-based functional mixed effects model introduced by Guo [11], which involves generalized mixed model equations 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 a different research context, Gaffney and Smyth proposed a framework for simultaneous clustering and alignment of time series data [12]. This work motivated us to integrate multi-group LC-MS data alignment in a mixture model for better identification of differentially abundant peptides.

In this paper, we propose a probabilistic mixture regression model (PMRM) for alignment of LC-MS data. Important features of this framework include: 1) it is not confined to landmark/feature alignment, 2) it allows continuous time alignment, 3) it utilizes a probabilistic modeling to deal with potentially irregular data, and 4) it is amenable to model both low and high resolution LC-MS data. Moreover, the framework lends itself to the expectation-maximization (EM) algorithm with the following features: 1) the explicit use of transformation priors for modeling the ion abundance (peak intensity) variability in both RT and m/z dimensions of the data, 2) the use of a probabilistic metric that allows estimation of the distance among multiple LC-MS data instead of computing pair-wise distances, and 3) the ability to extend the method for alignment of LC-MS data involving multiple groups.

The remainder of the paper is organized as follows. In Section 2, we outline the probabilistic mixture regression model for alignment of LC-MS data. This section also describes the estimation algorithm for finding the maximum likelihood parameters of the mixture regression models and the prior transformation density models used for modeling the variability in RT, m/z , and measurement space (peak intensities) of the LC-MS data. Section 3 briefly describes the datasets used in this paper. Results and discussion illustrating the applicability of our approach for analysis of LC-MS data are given in Section 4. We finally summarize our findings in Section 5.

2 Methods

2.1 Probabilistic Mixture Regression Models

We propose a probabilistic mixture regression model (PMRM) for alignment of LC-MS data. A particular advantage of PMRM is its ability to model non-Gaussian multi-modal density functions using simpler component density functions that can be defined on non-vector data such as LC-MS data.

We consider an observation dataset D consisting of M LC-MS runs $\{\mathbf{X}, \mathbf{Y}\} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^M$. Each observation $\mathbf{y}_i$, corresponding to its observation (RT) point $\mathbf{x}_i$, is independently drawn from a population of K groups, where the prior probability of the k -th group is α_k and $\sum_{k=1}^K \alpha_k = 1$. For each group, the observation is governed by the probability density function $p_k(\mathbf{y}_k | \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ are parameters of $p_k(\cdot)$.

From the above, it follows that the observed density on the $\mathbf{y}_i$'s is a mixture model, i.e., a convex combination of component models $p(\mathbf{y}_i | \theta) = \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i | \mathbf{x}_i, \theta_k)$, where $\theta = \{\theta_k\}_{k=1}^K$ is the complete parameter set for the mixture model. For the group specific probability density function, we model each observation $\mathbf{y}_i$ as a function of its observation point $\mathbf{x}_i$. The mapping function between $\mathbf{y}_i$ and $\mathbf{x}_i$ can be written as:

$$\mathbf{y}_i = \psi_k(\mathbf{x}_i, \boldsymbol{\beta}_k) + \mathbf{e}_i, \quad (1)$$

where $\mathbf{e}_i$ is a zero-mean Gaussian with variance $\sigma_k^2 \mathbf{I}$, i.e., $\mathbf{e}_i \sim \mathcal{N}(0, \sigma_k^2 \mathbf{I})$ and $\psi_k(\cdot, \cdot)$'s are splines-based regression functions of $\mathbf{x}_i$ with regression coefficients $\boldsymbol{\beta}_k$. Splines, which are capable of polynomial representation, are incorporated for flexible fitting of local variation of translation and scaling. We use cubic B-splines for efficient computation. With decomposition of $\mathbf{x}_i$ into a spline-basis matrix $\mathbf{B}_i = [B_1(\mathbf{x}_i) B_2(\mathbf{x}_i) \dots B_L(\mathbf{x}_i)]$ (the number of bases L is determined by the number of control points and the number of knots), the mapping function can be rewritten as:

$$\mathbf{y}_i = \mathbf{B}_i \boldsymbol{\beta}_k + \mathbf{e}_i, \quad (2)$$

where the dimension of $\boldsymbol{\beta}_k$ equals to the number of bases. Based on (2) and the error assumption, the group specific conditional probability density function for $\mathbf{y}_i$ can then be represented as:

$$p_k(\mathbf{y}_i | \mathbf{x}_i, \theta_k) = \mathcal{N}(\mathbf{B}_i \boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}). \quad (3)$$

Here, $\boldsymbol{\theta}_k$ includes both the regression coefficients $\boldsymbol{\beta}_k$ and the noise variances σ_k^2 . The probability of observing $\mathbf{y}_i$ conditioned on mixture model parameters $\boldsymbol{\theta}_k$'s can be written as:

$$p(\mathbf{y}_i | \mathbf{x}_i, \theta) = \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i | \mathbf{x}_i, \theta_k). \quad (4)$$

Conditional independence between the functional data representing multiple groups allows the full joint density to be written as:

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{i=1}^M \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i|\mathbf{x}_i, \theta_k), \quad (5)$$

and the log-likelihood of θ for the full observation $\mathbf{Y}$ can be obtained directly from (5) as follows:

$$\ell(\theta|\mathbf{D}) = \log p(\mathbf{Y}|\mathbf{X}, \theta) = \sum_{i=1}^M \log \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i|\mathbf{x}_i, \theta_k). \quad (6)$$

2.2 Modeling Variability along the RT Dimension and the Measurement Dimension

In the previous section, we introduced a probabilistic framework that incorporates B-splines into mixture regression model. This framework can be used to align multiple total ion chromatograms (TICs) involving two dimensions, i.e. total ion count values vs. RT points. In this section, we describe the data generative model with consideration of scaling and translation variations. With definition of prior transformation, the joint probabilities and the log-likelihood function can then be computed with the EM algorithm.

With a minor abuse of notation, (7) presents the general model formulation and the assumed prior transformation models for scaling and translation parameters:

$$\mathbf{y}_i = d_i \cdot \mathbf{1} + c_i \psi_k([\mathbf{a}_i \mathbf{x}_i - b_i \cdot \mathbf{1}], \beta_k) + \mathbf{e}_i, \quad (7)$$

where $\{a_i, b_i\}$ and $\{c_i, d_i\}$ are scaling and translation variables for RT and measurement spaces, respectively; and $\mathbf{1}$ is a unit column vector with an appropriate dimension. The definition given in (7) characterizes global scaling and translation variations in RT and measurement spaces. The posterior probabilities of these parameters are estimated together with the parameters of mixture model. By considering $\{a_i, b_i\}$ and $\{c_i, d_i\}$ as random variables with their associated prior probability density functions, one can then proceed with the mixture regression models for alignment of data representing multiple groups. Together with prior transformation models, the problem of deriving the resulting joint probabilities can therefore be directly handled.

With the assumption of independence between scaling and translation effects, the joint prior probabilistic model for the RT scaling and translation $p(a_i, b_i)$ can be factorized as $p(a_i) \times p(b_i)$. To present the idea that the translation variation is unbiased, a zero-mean Gaussian

prior is selected for RT translation, i.e., $b_i \sim \mathcal{N}(0, s_k^2)$ where s_k^2 is the variance. For the RT scaling prior a_i , since a value of one is the most likely value (i.e., no scaling at all), the

corresponding prior is given as $a_i \sim \mathcal{N}(1, r_k^2) \cdot I\{a_i > 0\}$. The variances r_k^2 and s_k^2 will be estimated within the ensuing EM algorithm. Similarly, the joint prior probabilistic model for the affine transformation on measurement space with independence assumption is given as $p(c_i, d_i) = p(c_i) \times p(d_i)$. With similar assumption of unbiased variation, we model the priors of scaling

and translation variations on measurement space as $c_i \sim \mathcal{N}(1, u_k^2) \cdot I\{c_i > 0\}$ and $d_i \sim \mathcal{N}(0, v_k^2)$. In the latter of this paper, the parameters for the mixture regression models are augmented to

include the scaling and translation variations, i.e., $\theta_k = \{\beta_k, \sigma_k^2, r_k^2, s_k^2, u_k^2, v_k^2\}$. After obtaining both the group-specific conditional probability density function for $\mathbf{y}_i$, i.e., $p_k(\mathbf{y}_i|\mathbf{x}_i, \theta_k)$, and

the likelihood function $\ell(\theta|\mathbf{D}, \mathbf{Z})$ (where $\theta = \{\beta_k, \sigma_k^2, r_k^2, s_k^2, u_k^2, v_k^2\}_{k=1}^K$ is the complete parameter set for the model in (7)), we perform the EM algorithm to estimate the most likely values for

parameters $\theta = \{\theta_k\}_{k=1}^K$, which allows us to find the best group alignment in time and measurement spaces.

2.3 Expectation-Maximization Estimation for the Mixture Regression Models

In order to infer the underlying group behavior, the group specific density components involved with unknown group membership need to be factorized from the joint density. A standard approach is to incorporate the hidden variables and utilize the EM algorithm [13]-[15] for consistent estimation. Introducing the hidden variables, which correspond to the unknown group membership of the M functional data, is appropriate in the log-likelihood formulation of (4). We introduce $\mathbf{Z}$ as a matrix of indicator vectors $\mathbf{z}_i = (z_{i1}, z_{i2}, \dots, z_{iK})$ where $z_{ik} = 1$ for some k and $z_{it} = 0, \forall t \neq k$. If $z_{ik} = 1$, it is interpreted that the i -th run is generated from the k -th mixture component. Thus, the joint density of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$ can be rewritten as follows:

$$p(\mathbf{Y}, \mathbf{Z} | \mathbf{X}, \theta) = \prod_{i=1}^M p(\mathbf{y}_i | \mathbf{z}_i, \mathbf{x}_i, \theta) \times p(\mathbf{z}_i). \quad (8)$$

This follows from the conditional independence assumption on $\mathbf{Y}$ and $\mathbf{z}_i$'s. Thus, the complete-data log-likelihood follows from (8) is represented as:

$$\ell(\theta | D, \mathbf{Z}) = \log p(\mathbf{Y}, \mathbf{Z} | \mathbf{X}, \theta) = \sum_{i=1}^M \sum_{k=1}^K z_{ik} \log \alpha_k + \sum_{i=1}^M \sum_{k=1}^K z_{ik} \log p(\mathbf{y}_i | \mathbf{x}_i, \theta_k). \quad (9)$$

After obtaining both the group-specific conditional probability density function for $\mathbf{y}_i$, i.e., $p_k(\mathbf{y}_i | \mathbf{x}_i, \theta_k, a_{i,k}, b_{i,k}, c_{i,k}, d_{i,k})$, and the complete data likelihood function $\ell(\theta | D, \mathbf{Z})$, the EM algorithm sequentially executes the following two steps:

Expectation-Step (E-Step): take the expectation of $\ell(\theta | D, \mathbf{Z})$ with respect to $p(\mathbf{Z} | \mathbf{Y}, \mathbf{X}, \theta^{t-1})$, where θ^{t-1} is the current set of parameters, and

Maximization-Step (M-Step): maximize the expected value of the computed log-likelihood from E-Step over the parameters θ to yield new parameters θ^t .

For the particular form of $\mathbf{Z}$ chosen here, the expectation of $\ell(\theta | D, \mathbf{Z})$ is

$$E\{\ell(\theta | D, \mathbf{Z})\} = \sum_{i=1}^M \sum_{k=1}^K h_{ik} \log \alpha_k + \sum_{i=1}^M \sum_{k=1}^K h_{ik} \log p_k(\mathbf{y}_i | \mathbf{x}_i, \theta_k), \quad (10)$$

where $h_{ik} = E\{z_{ik}\} \propto p_k(\mathbf{y}_i | \mathbf{x}_i, \theta^{t-1}) \times p(z_{ik} = 1)$ and it corresponds to the posterior probability of the i -th run being generated from the k -th group. Therefore, the E- and M-steps of the EM estimation algorithm proceed iteratively to estimate the parameters θ . The following are the steps involved in estimating the parameters for the mixture regression model using the EM algorithm:

Step 1: Initialize randomly the membership h_{ik} and estimate the parameter set θ^0 using the assigned membership.

Step 2: Compute the new membership expectation $h_{ik} = E\{z_{ik}\}$ using the current parameter set θ^{t-1} .

Step 3: Estimate the new parameter set θ^t using the membership expectation from Step 2.

Step 4: Repeat Steps 2 and 3 until a stopping criterion is met (e.g., maximum number of iteration or convergence rate).

2.4 Alignment of Three Dimensional LC-MS Data

Thus far, we have implicitly assumed that $\mathbf{y}_i$ represents the total ion chromatogram. To take advantage of the entire LC-MS data for differential protein expression studies, we extend our method for three-dimensional alignment (i.e., aligning with respect to RT points and m/z bins; and performing implicit normalization of peak intensities). We denote a three-dimensional mapping of an LC-MS run as $\tilde{\mathbf{y}}_i$ which consists of Q columns where the q -th column $\tilde{\mathbf{y}}_i^{(q)}$ contains two-dimensional data for the i -th observation. In other words, each $\tilde{\mathbf{y}}_i^{(q)}$ corresponds to an extracted total ion chromatogram stacked into a vector. The three-dimensional nature of each LC-MS run can be incorporated into the alignment model by defining a separate regression model for each column. Thus, the general regression model based on the assumed priors for scaling and translation parameters for each of the three dimensions (RT, m/z and peak intensities of peptides) can be presented as follows:

$$\begin{aligned} \tilde{\mathbf{y}}_i^{(q)} = & d_i^{(q)} \cdot 1 + c_i^{(q)} \psi_k^{(q)} \left([a_i \mathbf{x}_i - b_i] \mathbf{1}, \tilde{\beta}_k^{(q)} \right) + \mathbf{e}_i^{(q)} \\ & a_i \sim N(1, \tilde{\tau}_k^2) \cdot I\{a_i > 0\}, \quad b_i \sim N(0, \tilde{\sigma}_k^2), \\ & c_i^{(q)} \sim N(1, \tilde{u}_k^{2(q)}) \cdot I\{c_i^{(q)} > 0\}, \quad d_i^{(q)} \sim N(0, \tilde{v}_k^{2(q)}), \\ & \mathbf{e}_i^{(q)} \sim N(0, \tilde{\sigma}_k^{2(q)} \mathbf{I}) \\ & i=1, 2, \dots, M, \quad q=1, 2, \dots, Q, \quad \text{and} \quad k=1, 2, \dots, K \end{aligned} \quad (11)$$

where $\tilde{\beta}_k^{(q)}$ consists of the regression coefficients for the q -th column (i.e., the regression coefficients for the q -th column of $\tilde{\mathbf{y}}_i^{(q)}$; $\{a_i, b_i\}$ and $\{c_i^{(q)}, d_i^{(q)}\}$ are scaling and translation variables for RT points and measurement space (peak intensities), respectively. Note that the scaling and translation parameters for measurement space vary from column to column.

In the following, we illustrate the derivation using only the translation in measurement space, $\{d_i^{(q)}\}_{q=1}^Q$, for three-dimensional data $\tilde{\mathbf{y}}_i$ with the model specified in (11):

$$p(\tilde{\mathbf{y}}, d_i^{(1)}, \dots, d_i^{(Q)}) = \prod_{q=1}^Q N(\tilde{\mathbf{y}}_i^{(q)} | \tilde{\mathbf{B}}_i^{(q)} \tilde{\beta}_k^{(q)} + d_i^{(q)} \cdot 1, \tilde{\sigma}_k^{2(q)} \cdot \mathbf{I}) \times N(d_i^{(q)} | 0, \tilde{v}_k^{(q)}). \quad (12)$$

The joint density can be factored due to the following necessary conditions: 1) conditional independence is assumed between the columns of $\tilde{\mathbf{y}}_i$, and 2) each column is assumed to have its own set of translation parameters. Thus, the marginal density $p(\tilde{\mathbf{y}}_i)$ is given as

$$\prod_{q=1}^Q p(\tilde{\mathbf{y}}_i^{(q)}) \text{ and the resulting log-likelihood of } \mathbf{Y} = \{\tilde{\mathbf{y}}_i\}_{i=1}^M \text{ takes the form}$$

$$\log p(\mathbf{Y}) = \sum_{i=1}^M \log p(\tilde{\mathbf{y}}_i) = \sum_{i=1}^M \sum_{q=1}^Q \log \int d_i^{(q)} p(\tilde{\mathbf{y}}_i^{(q)} | d_i^{(q)}) \times p(d_i^{(q)}) dd_i^{(q)}. \quad (13)$$

Equation (13) includes all the $M \times Q$ chromatograms in the log-likelihood function and analytically integrating (13) gives the log-likelihood as:

$$\log p(\mathbf{Y}) = \sum_{i=1}^M \sum_{q=1}^Q \log N(\tilde{\mathbf{y}}_i^{(q)} | \tilde{\mathbf{B}}_i^{(q)} \tilde{\beta}_k^{(q)}, \tilde{v}_k^{(q)} \cdot 1 + \tilde{\sigma}_k^{2(q)} \cdot \mathbf{I}). \quad (14)$$

Thus, for a complete alignment problem in the measurement space including both scaling and translation parameters, the individual columns require $2Q$ separate scaling and translation parameters to be incorporated in this framework.

We address RT point alignment with a different perspective where the dynamic behavior for each column is assumed the same. Thus, the RT transformation parameters are shared over the Q columns for each group data. For the RT translation case, each chromatogram/trace in $\tilde{\mathbf{y}}_i$ shares a single translation parameter b_i from its corresponding group behavior and the conditional density of $\tilde{\mathbf{y}}_i$ becomes

$$p(\tilde{\mathbf{y}}_i|b_i) = \prod_{q=1}^Q p(\tilde{\mathbf{y}}_i^{(q)}|b_i) = \prod_{q=1}^Q N(\tilde{\mathbf{y}}_i^{(q)} | \tilde{\mathbf{B}}_i^{(q)} \tilde{\beta}_k^{(q)}, \tilde{\sigma}_k^{2(q)} \cdot \mathbf{I}), \quad (15)$$

in which there is only one b_i for all q in each group. The corresponding marginal density $p(\tilde{\mathbf{y}}_i)$ cannot be factored as that on measurement space translation due to the dependence among dimensions through the translation parameters b_i . Therefore the log-likelihood of $\mathbf{Y}$ is given as follows:

$$\log p(\mathbf{Y}) = \sum_{i=1}^M \log p(\tilde{\mathbf{y}}_i) = \sum_{i=1}^M \log \int_{b_i} \left[p(b_i) \prod_{q=1}^Q p(\tilde{\mathbf{y}}_i^{(q)}|b_i) \right] db_i. \quad (16)$$

The product over the column is now inside of the integration operation and an approximate log-likelihood can be computed using the Monte-Carlo numerical integration technique [16]. Thus, the approximation can be computed as follows:

$$\log p(\mathbf{Y}) = \sum_{i=1}^M \log \left[\sum_{r=1}^R \prod_{q=1}^Q p(\tilde{\mathbf{y}}_i^{(q)}|b_i^{(r)}) \right] - M \log R, \quad (17)$$

where $b_i^{(r)} \sim \mathcal{N}(0, \tilde{s}_k^2)$ for $r = 1, 2, \dots, R$ and R is the sample size in the Monte-Carlo numerical integration.

The derivation for the scaling variables $c_i^{(q)}$ on measurement space and a_i for RT points will follow directly from the previous derivations except for handling the various individual q subscripts on the column dependent parameters. Moreover, the EM algorithm outlined in Section 2.3 is then modified at each iteration to update the following parameter set

$$\theta' = \left\{ \tilde{\beta}_k^{(q)}, \tilde{\sigma}_k^{2(q)}, \tilde{r}_k^2, \tilde{s}_k^2, \tilde{u}_k^{2(q)}, \tilde{v}_k^{2(q)} \right\}^t \text{ for } q = 1, 2, \dots, Q \text{ and } k = 1, 2, \dots, K.$$

3 LC-MS Data

3.1 Single-class Listgarten et al. LC-MS Dataset

The single-class Listgarten et al. LC-MS dataset was obtained from <http://www.cs.toronto.edu/~jenn/LCMS>. It consists of 11 replicate LC-MS runs generated from proteins of *lysed E.coli* cells using a capillary-scale LC coupled on-line to an ion trap mass spectrometer. Each run is represented by a 400×2400 data matrix corresponding to 400 RT points (~55 min.) and 2400 m/z bins spanning between 400 and 1600 Da. Detailed experimental information can be found in [17].

3.2 Two-class Listgarten et al. LC-MS Dataset

The two-class Listgarten et al. dataset (downloaded from <http://www.cs.toronto.edu/~jenn/LCMS>) was obtained from human serum samples that

represent two classes. The dataset consists of 14 LC-MS runs generated from two groups of human serum samples. Seven runs were derived from a sample, in which the human serum was mixed with three known peptides: Angiotensin II, Angiotensin I, and Substance P with molecular weights of 1047.2, 1297.51, and 1348.66 Da, respectively. The other seven runs were from a sample consisting of human serum only. The LC-MS runs were generated using high performance liquid chromatography (HPLC) pump interfaced with a Thermo Finnigan LCQ quadrupole ion trap tandem mass spectrometer. Each run is represented by a 500×2400 data matrix corresponding to 500 RT points (~55 min.) and 2400 m/z bins between 400 and 1600 Da.

3.3 Mueller et al. LC-MS Dataset

The Mueller et al. LC-MS dataset was obtained from http://prottools.ethz.ch/muellelu/web/Latin_Square_Data.php. It consists of 18 LC-MS runs 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 from serum using solid-phase Nglycocalapture. The LC-MS runs were obtained using the Fourier transformed-linear trap quadrupole (FT-LTQ) mass spectrometer. Detailed experimental information can be found in [6]. The 18 LC-MS runs were grouped into six classes based on the concentration of the spike-in proteins. Specifically, Latin square dilution with different concentrations of the six standard proteins was used. Each run is represented by a 2000×750 data matrix corresponding to 2000 RT points (~55 min.) and 750 m/z bins between 300 and 1600 Da.

4 Results and Discussion

We applied PMRM to align the single-class dataset based on total ion chromatograms. The remaining two datasets were analyzed on the basis of three-dimensional data involving RT points, m/z bins, and peak intensities of peptides.

4.1 Single-class Listgarten et al. LC-MS Dataset

The single-class Listgarten et al. dataset was converted into 11 total ion chromatograms by calculating the total ion count at each RT point (i.e., at each RT point the sum of all ions across m/z values was calculated). Fig. 1 depicts TICs of the 11 replicate LC-MS runs. From the figure we observe significant shifts along the experimental RT points as well as distortions in the measurement space (TIC values). The TICs aligned by the PMRM approach are shown in Fig. 2. The mean TICs for the original (before alignment) and those aligned by DTW, COW, CPM, and PMRM are depicted in Fig. 3. From this figure, we see that CPM produces superfluous signal gaps, thus stretching the experimental RT points by almost 2 folds. Although the stretched RT points can be rescaled to the original RT points, the difficulty arises in sorting out the inserted gaps during peptide sequence identification to accurately match the sequence information with the experimental RT points.

Table 1 presents the coefficient of variations (CVs) for the original data and those aligned by DTW, COW, CPM, and PMRM. As shown in the table, both CPM and PMRM yield better results than DTW and COW.

4.2 Two-class Listgarten et al. LC-MS Dataset

Alignment—Each LC-MS run in the two-class Listgarten et al. dataset was reduced from a 500×2400 to a 500×8 data matrix by partitioning the m/z columns into eight equal m/z bins. This reduction was necessary to handle the computational complexity. We applied the PMRM to the reduced matrix and obtained a set of aligned data matrices, i.e., for each LC-

MS run ($i=1,2,\dots,14$), a data matrix $\mathbf{M}^i$ was obtained, where $\mathbf{M}_{\tau,q}^i$ is the ion abundance after alignment at the τ -th aligned RT point ($\tau=1,2,\dots,500$) and the q -th m/z bin ($q=1,2,\dots,8$). For visualization purpose, we calculated the TICs from the reduced original and aligned matrices. Figs 4 and 5 depict the TICs of the original and aligned LC-MS runs, respectively. Table 1 presents the CVs for the original and those aligned by DTW, COW, CPM, and PMRM. As illustrated in the table, both CPM and PMRM have yielded better results than DTW and COW.

Although CV is a good measure for assessing the variability among replicate LC-MS traces, knowledge of such information does not necessarily confirm or directly link with actual interpretation of differences between the LC-MS runs. In the following, we use statistical measures for detected differences between the two classes to assess the performance of various alignment methods.

Difference detection—We used the aligned data matrices for difference detection with a t -statistic similar to the one in [17]. Specifically, we calculated the t -statistic at each pair of RT point and m/z bin among different groups. The group mean $\mathbf{C}_{\tau,q}^k$ and variance $\mathbf{V}_{\tau,q}^k$ of group k at the τ -th aligned RT point and q -th m/z bin, are calculated as follows:

$$\mathbf{C}_{\tau,q}^k = \frac{1}{N_k} \sum_{\{i/i \in \text{Group } k\}} \mathbf{M}_{\tau,q}^i, \quad (18)$$

and

$$\mathbf{V}_{\tau,q}^k = \frac{1}{(N_k - 1)} \sum_{\{i/i \in \text{Group } k\}} (\mathbf{C}_{\tau,q}^i - \mathbf{M}_{\tau,q}^i)^2, \quad (19)$$

where N_k is the number of samples in group k . The pooled standard deviation, $S_{\tau,q}$ (group k versus the other groups) is calculated as:

$$S_{\tau,q} = \sqrt{\mathbf{V}_{\tau,q}^k / N_k + \mathbf{V}_{\tau,q}^{l/l \neq k} / N_{\{l/l \neq k\}}}. \quad (20)$$

Based on (18)-(20), we can calculate the t -statistic among the case and control conditions as:

$$t_{\tau,q} = (\mathbf{C}_{\tau,q}^{\text{Case}} - \mathbf{C}_{\tau,q}^{\text{Control}}) / S_{\tau,q}. \quad (21)$$

The t -statistic provides quantitative information to detect differences among the LC-MS data. With a list of experimentally extracted ground truth peptide m/z values provided in [17], we used the following steps to compute the precision-recall curve: 1) set a threshold T_0 and collect all m/z values with a t -statistic larger in magnitude than T_0 , 2) find all unique m/z values from this collection, 3) obtain the number of true positives by counting how many of these m/z values appear in the ground truth list within the specified tolerance, and 4) calculate the number of false positives m/z values, which did not match the ground truth peaks within the specified tolerance. By decreasing the threshold T_0 , we repeated the steps 1) - 4) to trace the precision-recall curves. In an ideal difference detection, the precision value would keep 1 as the recall value goes upwards. Fig. 6 shows the computed precision-recall curves for the two-class Listgarten et al. dataset. PMRM presents better performance than CPM.

Evaluations including sensitivity, specificity, accuracy, and precision were also compared. For three different significant levels ($p < 0.05$, $p < 0.01$, and $p < 0.001$), we performed a hypothesis test to determine whether the ion abundances differ between the two classes. Table 2 presents the computed sensitivity, specificity, accuracy, and precision in detecting differentially abundant peptides using the LC-MS data aligned by DTW, COW, CPM and PMRM. The table also provides an F-measure, which is a weighted harmonic mean measure combining the precision and recall, calculated as follows with $\beta = 1$ [18]:

$$F_{\text{measure}} = \frac{(\beta^2 + 1) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}.$$

As shown in Table 2, the LC-MS runs aligned by PMRM have yielded overall better performance than those aligned by DTW, COW and CPM. For $p < 0.001$ significant level, CPM did not yield any differentially abundance ions. Possible reasons for this could be due to over smoothing and attenuation of peaks by CPM.

Moreover, we performed a comparison by controlling a false discovery rate (FDR) at 10% for a 95% confidence level in detecting differentially abundant peptides from the LC-MS runs aligned by DTW, COW, CPM and PMRM. Table 3 presents the computed sensitivity, specificity, accuracy, precision and F-measure in detecting differentially abundant peptides. From this table, we see that both CPM and PMRM provide better sensitivity and F-measure results than DTW and COW.

4.3 Mueller et al. LC-MS Dataset

Alignment—The Mueller et al. LC-MS dataset was reduced from a 2000×750 to 2000×8 data matrix by partitioning the m/z columns into eight equal m/z bins. After applying PMRM to this reduced matrix, we obtained a set of aligned data matrices, i.e., for each LC-MS run ($i=1,2,\dots,18$), a data matrix $\mathbf{M}^i$ was obtained, where $\mathbf{M}_{\tau,q}^i$ is the ion abundance obtained after alignment at the τ -th aligned RT points ($\tau=1,2,\dots,2000$) and the q -th m/z bin ($q=1,2,\dots,8$). Figs. 7 and 8 depict TICs of the original and aligned LC-MS data, respectively, for RT points between 800 and 2000.

Difference detection—The aligned data matrices were used for difference detection. Since the original LC-MS dataset consisted of the diluted series of six proteins with different concentration in each sample, we determined the peptides with statistically significant difference in their expression level based on an F-statistic. To accomplish this, we first generated a list of peptides assuming no mis-cleavage from trypsin digestion of six non-human proteins (myoglobin, carbonic anhydrase, cytochrome c, lysozyme, alcohol dehydrogenase, and aldolase A) using the MS-Digest software from the University of California, San Francisco (<http://prospector.ucsf.edu>). Then, we used the permutation method to examine each RT point and m/z bin. We compared the statistically significantly different RT- m/z pairs $p < 0.01$ against the list of peptides that represent the six proteins to calculate sensitivity, specificity, accuracy, precision, and F-measure. Table 4 presents the results for the LC-MS runs aligned by DTW, COW, CPM and PMRM. The table illustrates that PMRM yields overall better results than DTW, COW, and CPM. Table 5 presents the computed sensitivity, specificity, accuracy, precision and F-measure by controlling the FDR at 10% for a 95% confidence level in detecting differentially abundant peptides from the LC-MS runs aligned by DTW, COW, CPM and PMRM. From the table, we see that both CPM and PMRM provide better sensitivity and F-measure results than DTW and COW.

4.4 Reproducibility Study

We examine the reproducibility behavior of the PMRM in aligning LC-MS data. To accomplish this, it is necessary that the identifiability property holds with respect to the mixture models given by (4) (or (5) and the general model definition given by (7)). However, the general statement on the identifiability property requires minimizing distributions using Kullback-Leibler divergence from the true mixing distribution [19], [20]. Since the identifiability property of mixing distributions is of interest in its own right, we focus here on examining the consistency of the PMRM parameters by running the algorithm multiple times for the same LC-MS dataset. Table 6 and 7 present the 95%, 99% and 99.9% confidence intervals for the mean of alignment field (i.e., the expected posteriors for scaling parameters a_i and translation parameters b_j) obtained in aligning the single-class Listgarten et al. LC-MS dataset. The estimated values show fairly narrow intervals, which indicates consistent estimation of the parameters across multiple runs. This analysis also provides similar mean posterior estimates for the other alignment field parameters. The algorithm took approximately nine minutes for the algorithm to complete one run on a PC with an Intel Core 2 Duo 64 bit 2.66 GHz CPU and 8 GB RAM.

5 Conclusion

This paper proposes a probabilistic mixture regression model (PMRM) for alignment of multiple LC-MS data. We utilize the well known maximum likelihood-based EM algorithm for estimating the mixture regression models and the prior transformation (scaling and translation parameters) models of the LC-MS data. The latter accounts for the variability in RT points, m/z values, and peak intensities. The proposed framework allows alignment with respect to both RT points and m/z bins; and it implicitly performs normalization of peak intensities in multiple LC-MS runs. The performance of the approach is assessed through three datasets: 1) replicate LC-MS runs generated from proteins of *lysed E.coli* cells, 2) LC-MS runs representing two classes of human serum samples (with and without spike-in peptides), and 3) LC-MS runs representing six classes, where six proteins are spiked at different concentrations into a complex sample background of human peptides. Through these datasets, it is demonstrated that PMRM achieves good coefficient of variation among replicate LC-MS runs. In addition, through spike-in peptides, the paper demonstrates that PMRM leads to more accurate identification of differentially abundant peptides than DTW and COW, and comparable or slightly better performance than CPM. Future work will focus on extending PMRM for alignment and normalization of multiple LC-MS runs, to account for variability among samples from distinct biological groups, and to model the heterogeneity within group.

Acknowledgments

This work was supported in part by the National Science Foundation Grant (IIS-0812246), National Cancer Institute (NCI) R21 CA130837 Grant, NCI Early Detection Research Network Associate Membership Grant, and the Prevent Cancer Foundation Grant awarded to HWR.

References

1. Aebersold R, Mann M. Mass Spectrometry-Based Proteomics. *Nature*. 2003; 422(6928):198–207. [PubMed: 12634793]
2. Tomasi G, van den Berg F, Andersson C. Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data. *J. of Chemometrics*. 2004; 18:231–241.

3. Hastings CA, Norton SM, Roy S. New Algorithms for Processing and Peak Detection in Liquid Chromatography/Mass Spectrometry Data. *Rapid Commun Mass Spectrom*. 2002; 16(5):462–7. [PubMed: 11857732]
4. Fischer B, Grossmann J, Roth V, Gruissem W, Baginsky S, Buhmann JM. Semi-Supervised LC/MS Alignment for Differential Proteomics. *Bioinformatics*. 2006; 22(14):e132–40. [PubMed: 16873463]
5. Wang P, Tang H, Fitzgibbon MP, McIntosh M, Coram M, Zhang H, Yi E, Aebersold R. A Statistical Method for Chromatographic Alignment of LC-MS Data. *Biostatistics*. 2007; 8(2):357–67. [PubMed: 16880200]
6. Mueller LN, Rinner O, Schmidt A, Letarte S, Bodenmiller B, Brusniak MY, Vitek O, Aebersold R, Müller M. SuperHirn - A Novel Tool for High Resolution LC-MS-Based Peptide/Protein Profiling. *Proteomics*. 2007; 7(19):3470–80. [PubMed: 17726677]
7. Listgarten J, Neal RM, Roweis ST, Wong P, Emili A. Difference Detection in LC-MS Data for Protein Biomarker Discovery. *Bioinformatics*. 2007; 23(2):e198–204. [PubMed: 17237092]
8. Listgarten J, Neal RM, Roweis ST, Emili A. Multiple Alignment of Continuous Time Series. *Advances in Neural Information Processing Systems*. 2005; 17:817–824.
9. Morris JS, Brown PJ, Herrick RC, Baggerly KA, Coombes KR. Bayesian Analysis of Mass Spectrometry Proteomic Data Using Wavelet-Based Functional Mixed Models. *Biometrics*. 2008; 64(2):479–89. [PubMed: 17888041]
10. Morris JS, Carroll RJ. Wavelet-Based Functional Mixed Models. *J. Royal Statistical Soc. B*. 2006; 68(2):179–199.
11. Guo W. Functional Data Analysis in Longitudinal Settings Using Smoothing Splines. *Stat Methods Med Res*. 2004; 13(1):49–62. [PubMed: 14746440]
12. Gaffney S, Smyth P. Joint Probabilistic Curve Clustering and Alignment. *Advances in Neural Information Processing Systems*. 2005; 17:473–480.
13. Dempster AP, Laird NM, Rubin DB. “Maximum Likelihood from Incomplete Data via the EM Algorithm” (with discussion). *J. Royal Statistical Soc. B*. 1977; 39(1):1–38.
14. Jordan MI, Jacobs RA. Hierarchical Mixtures of Experts and the EM Algorithm. *Neural Comput*. 1994; 6(2):181–214.
15. Redner RA, Walker HF. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*. 1984; 26(2):195–239.
16. Lange, K. *Numerical Analysis for Statisticians*. Springer; New York: 1999.
17. Listgarten, J. *Analysis of Sibling Time Series Data: Alignment and Difference Detection*. Univ. of Toronto; 2006. Ph.D. thesis
18. van Rijsbergen, CJ. *Information Retrieval*. 2nd ed.. Butterworths; Boston: 1979.
19. Henna J. On Estimating of the Number of Constituents of a Finite Mixture of Continuous Distributions. *Annals of the Institute of Statistical Mathematics*. 1985; 37(1):235–240.
20. McLachlan, GJ.; Basford, KE. *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker; New York: 1988.

Fig. 1.
TICs of the original single-class Listgarten et al. dataset.

Fig. 2.
TICs of the single-class Listgarten et al. dataset after alignment by PMRM.

Fig. 3. Mean TICs for single-class Listgarten et al. datasets (original and those aligned by DTW, COW, CPM and PMRM).

Fig. 4.
TICs for the two-class Listgarten et al. dataset before alignment.

Fig. 5.
TICs for the two-class Listgarten et al. dataset after alignment by PMRM.

Fig. 6.
Precision-recall curves in detecting differentially abundant peptides in the two-class Listgarten et al. dataset by CPM, and PMRM.

Fig. 7.
TICs for the Mueller et al. LC-MS dataset before alignment.

Fig. 8.
TICs for the Mueller et al. LC-MS dataset after alignment by PMRM.

TABLE 1

Coefficient of variations for the single-class and two-class Listgarten et al. datasets before alignment (original) and after alignment by DTW, COW, CPM, and PMRM. The result of the method with the best performance in each category is shown in boldface.

	Single-class Listgarten et al.	Two-class Listgarten et al.	
		Class 1	Class 2
Original	82%	57%	58%
DTW	70%	56%	57%
COW	80%	56%	57%
CPM	57%	44%	44%
PMRM	65%	47%	47%

TABLE 2

Sensitivity, specificity, accuracy, precision and F-measure in detecting differentially abundant peptides in the two-class listgarten et al. dataset by DTW, COW, CPM, and PMRM (for $p < 0.05$, $p < 0.01$ and $p < 0.001$). The result of the method with the best performance in each category is shown in boldface.

	P-values	Sensitivity	Specificity	Accuracy	Precision	F-measure
DTW		60%	55%	58%	60%	60%
COW	0.05	64%	57%	60%	62%	63%
CPM		59%	83%	70%	73%	64%
PMRM		65%	79%	73%	73%	69%
DTW		63%	63%	61%	64%	63%
COW	0.01	63%	61%	62%	77%	69%
CPM		45%	90%	68%	73%	57%
PMRM		62%	81%	72%	73%	67%
DTW		63%	64%	64%	61%	62%
COW	0.001*	67%	66%	66%	63%	65%
CPM		-	-	-	-	-
PMRM		59%	83%	71%	73%	64%

*The condition $p < 0.001$ is too stringent for CPM.

TABLE 3

Sensitivity, specificity, accuracy, precision and F-measure in detecting differentially abundant peptides in the two-class Listgarten et al. dataset by DTW, COW, CPM, and PMRM for fdr=10% at 95% confidence level. The result of the method with the best performance in each category is shown in boldface.

	Sensitivity	Specificity	Accuracy	Precision	F-measure
DTW	69%	91%	76%	88%	77%
COW	75%	87%	80%	89%	81%
CPM	81%	87%	84%	90%	85%
PMRM	81%	87%	83%	90%	85%

TABLE 4

Sensitivity, specificity, accuracy, precision and F-measure in detecting differentially abundant peptides in the Mueller et al. dataset by DTW, COW, CPM, and PMRM for $p < 0.01$. The result of the method with the best performance in each category is shown in boldface.

	Sensitivity	Specificity	Accuracy	Precision	F-measure
DTW	76%	70%	72%	70%	73%
COW	83%	75%	79%	76%	79%
CPM	90%	75%	83%	84%	80%
PMRM	91%	78%	85%	83%	87%

TABLE 5

Sensitivity, specificity, accuracy, precision and F-measure in detecting differentially abundant peptides in the Mueller et al. dataset by DTW, COW, CPM, and PMRM for $fdr = 10\%$ at 95% confidence level. The result of the method with the best performance in each category is shown in boldface.

	Sensitivity	Specificity	Accuracy	Precision	F-measure
DTW	80%	87%	83%	90%	85%
COW	82%	88%	84%	90%	86%
CPM	90%	87%	87%	90%	90%
PMRM	90%	84%	86%	90%	90%

TABLE 6

The expected values of the scaling parameters a_i .

$E\{a_i\}$	Confidence Intervals		
	95%	99%	99.9%
1.066	[1.059, 1.073]	[1.056, 1.076]	[1.054, 1.078]
1.092	[1.084, 1.099]	[1.082, 1.102]	[1.079, 1.105]
1.101	[1.094, 1.108]	[1.091, 1.111]	[1.088, 1.113]
1.072	[1.065, 1.080]	[1.063, 1.083]	[1.060, 1.085]
1.107	[1.100, 1.115]	[1.098, 1.118]	[1.095, 1.121]
1.130	[1.123, 1.138]	[1.120, 1.140]	[1.117, 1.143]
1.100	[1.093, 1.108]	[1.090, 1.110]	[1.087, 1.113]
1.095	[1.088, 1.103]	[1.085, 1.105]	[1.083, 1.108]
1.062	[1.056, 1.070]	[1.053, 1.072]	[1.051, 1.075]
1.028	[1.021, 1.035]	[1.018, 1.037]	[1.016, 1.039]
1.081	[1.074, 1.088]	[1.072, 1.091]	[1.069, 1.093]

TABLE 7The expected values of the translation parameters b_i .

$E\{b_i\}$	Confidence Intervals		
	95%	99%	99.9%
-8.546	[-9.066, -8.025]	[-9.232, -7.860]	[-9.422, -7.669]
-8.486	[-9.060, -7.913]	[-9.242, -7.731]	[-9.451, -7.521]
-3.454	[-3.950, -2.959]	[-4.107, -2.801]	[-4.288, -2.620]
-18.998	[-19.541, -18.456]	[-19.713, -18.284]	[-19.912, -18.085]
-6.532	[-7.077, -5.986]	[-7.250, -5.813]	[-7.450, -5.614]
12.305	[11.695, 12.914]	[11.502, 13.108]	[11.279, 13.330]
-6.316	[-6.871, -5.761]	[-7.047, -5.585]	[-7.250, -5.382]
-9.057	[-9.595, -8.520]	[-9.765, -8.350]	[-9.962, -8.153]
-20.278	[-20.818, -19.738]	[-20.989, -19.567]	[-21.187, -19.370]
0.442	[-0.034, 0.918]	[-0.185, 1.069]	[-0.359, 1.243]
-0.007	[-0.258, 0.243]	[-0.337, 0.322]	[-0.429, 0.414]