Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2008 ; 2008: 4094–4097. doi:10.1109/IEMBS.2008.4650109.

Multi-Class Alignment of LC-MS Data Using Probabilistic-Based Mixture Regression Models

Getachew K. Befekadu,

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, USA

Mahlet G. Tadesse,

Department of Mathematics, Georgetown University, 308 St. Mary's Hall, Washington, DC 20057, USA

Yetrib Hathout, and

Children's National Medical Center, Center for Genetic Medicine, 111 Michigan Avenue, NW, Washington DC 20010, USA

Habtom W. Ressom*

Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC 20057, USA

Abstract

In this paper, a framework of probabilistic-based mixture regression models (PMRM) is presented for multi-class alignment of liquid chromatography-mass spectrometry (LC-MS) data. The proposed framework performs the alignment in both time and measurement spaces of the LC-MS spectra. The expectation maximization (EM) algorithm is used to estimate the joint parameters of spline-based mixture regression models and prior transformation densities. The latter are incorporated to account for variability in time and measurement spaces of the data. As a proof of concept, the proposed method is applied to align a single-class replicate LC-MS spectra generated from proteins of *lysed E.coli* cells. Its performance is compared with the dynamic time warping (DTW) and continuous profile model (CPM) approaches.

I. INTRODUCTION

Liquid chromatography (LC) coupled with high precision mass spectrometers is routinely used in proteomics studies to quantify and determine the abundance of various peptides that represent particular proteins in biological samples [1,2]. This coupling generates large and high dimensional data with peptide intensities at specific mass-charge ratios (m/z) and retention times (RT) in LC run. In studies such as biomarker discovery multi-class liquid chromatography-mass spectrometry (LC-MS) data are compared to identify differentially abundant peptides between distinct biological groups. The detection of differentially abundant peptides from LC-MS data is a challenging task due to the following reasons: i) substantial variation in RT across multiple runs due to the LC instrument conditions and the composition of peptide mixture, and ii) the variation of m/z values of the peptides because of noise in the instrument. Thus, alignment with respect to both RT and m/z is a prerequisite for quantitative comparison of multi-class LC-MS data.

*Corresponding author: E-mail: hwr@georgetown.edu.

Traditionally, alignment algorithms have been used on data points and/or feature vectors of fixed dimension as described previously [3]. However, LC-MS data commonly consist of a variable number of measurements, observed over intervals of varying size with some missing observations. The most common approach for aligning LC-MS data is based on the identification of landmarks or structural points usually associated with maxima, minima, or other critical or inflection points of each spectrum. The whole spectra are then aligned so that the landmarks are synchronized [4]. The drawback of this approach is that it requires landmarks to be identified prior to alignment. On the other hand, methods that rely on optimization of global fitting function provide alternative solution to alignment problems without requiring landmarks. One of the methods that do not utilize landmarks for alignment is dynamic time warping (DTW), which was originally applied in speech recognition [5]. DTW has been applied for aligning chromatographic and LC-MS data [6–8]. However, the above approach is limited for a consensus alignment of all pair-wise combinations of spectra. Another recently introduced method is the continuous profile model (CPM) based on hidden Markov model (HMM) [9]. CPM has been applied for multi-class alignment of continuous time-series data and for detecting differences in LC-MS data [9]. Although CPM is described as a naïve and computationally intensive method, it is clear that the nonlinear relationship between the experimental (physical) and the aligned timescales have been artificially forced to relate during the problem formulation.

In this paper, we propose a framework of probabilistic-based mixed regression models (PMRM) that directly addresses the multi-class alignments of LC-MS data. The proposed method is not confined to landmarks, allows for continuous time alignment, and employs functional curve modeling to deal with issues such as variable sequence length and non-uniformly sampled data. The framework lends itself to an expectation-maximization (EM) algorithm with the following features: i) the explicit use of transformation priors for modeling of the variability in time and measurement spaces of the data, ii) the use of an implicit distance metric for multi-class alignment, iii) the integration of alignment into more general multi-class alignment problem, and (iv) its flexibility to include different prior transformations.

The remainder of this paper is organized as follows. In Section II, we outline the PMRM and describe the generation mechanism of functional curve data. Additionally, this section explains the algorithm for finding the maximum likelihood parameters of the regression models – spline-based mixture regression models – and the prior densities used for modeling the variability in time and measurement spaces of the data. Section III illustrates the applicability of the proposed method by aligning a set of replicate LC-MS spectra and comparing the results with those obtained by DTW and CPM. Section IV summarizes our findings.

II. PROBABILISTIC BASED MIXTURE MODELS FOR MULTICLASS ALIGNMENT PROBLEM

A. Model Representation

We assume that the observed continuous-functional curve data are generated with the following features:

1. An individual is randomly drawn from the population of interest.
2. The individual is assigned to class k with probability α_k , $\sum_{k=1}^K \alpha_k = 1$. These are the prior weights on the K classes.
3. Given an individual i that belongs to class k , there is a density function $p_k(\mathbf{y}_i | \boldsymbol{\theta}_k)$ that generates the observed functional curve data $\mathbf{y}_i$ for this individual.

From the above, it follows that the observed density on the $\mathbf{y}_i$'s will be a mixture model, i.e., a convex combination of component models $p(\mathbf{y}_i/\boldsymbol{\theta}) = \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i/\boldsymbol{\theta}_k)$. Thus, based on observed $\mathbf{y}_i$'s and assumed functional form for the p_k components, we can estimate from the data the most likely values of the parameters $\boldsymbol{\theta}_k$ and the weights α_k . This can be accomplished by using the EM algorithm which is a general procedure for finding the maximum likelihood estimates of the parameters present in mixture models [10–12]. The probabilistic framework then allows one to find the best class-alignment in time and measurement spaces of the functional curve data.

In the following, we consider M spectra $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^M$, where the i th spectrum $\mathbf{y}_i$ has n_i measurements corresponding to the observation points (or time) $\mathbf{x}_i$. To define the mixture regression models, each spectrum $\mathbf{y}_i$ is expressed as a function of some known $\mathbf{x}_i$, i.e., each component model is represented by a conditional density function of the form $p_k(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}_k)$. We assume a standard regression relationship between $\mathbf{y}_i$ and $\mathbf{x}_i$ as:

$$\mathbf{y}_i = \boldsymbol{\Psi}_k(\mathbf{x}_i, \boldsymbol{\beta}_k) + \mathbf{e}_i, \text{ for } i=1, \dots, M \quad (1)$$

where $\mathbf{e}_i$ is a zero-mean Gaussian with variance $\sigma_k^2 \mathbf{I}$ and $\boldsymbol{\Psi}_k(\cdot, \cdot)$'s are deterministic mapping functions of $\mathbf{x}_i$. Here $\boldsymbol{\theta}_k$ includes both the parameters of the mapping model $\boldsymbol{\Psi}_k(\mathbf{x}_i, \boldsymbol{\beta}_k)$, i.e. $\boldsymbol{\beta}_k$, and the noise variances σ_k^2 . Thus, the probability of observing $\mathbf{y}_i$, given $\mathbf{x}_i$ and component model k , is assumed to be a conditional regression model. We define the conditional density of the observed data $p_k(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}_k)$ as a mixture density conditioned on the regression model as well as on $\mathbf{x}_i$ values:

$$p(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}) = \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}_k) \quad (2)$$

where $p_k(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}_k)$'s are the mixture components, α_k 's are the mixing weights, and $\boldsymbol{\theta}_k$ is the set of parameters for the component k . Conditional independence between the functional curve data amounts assuming that the individuals constitute a random sample from a population, and allows the full joint density to be written as:

$$p(\mathbf{Y}/\mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^M \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}_k) \quad (3)$$

Thus, the log-likelihood of the parameters $\boldsymbol{\theta}$ given a dataset D can be obtained directly from Eq. (3) as follows:

$$\ell(\boldsymbol{\theta}/D) = \log p(\mathbf{Y}/\mathbf{X}, \boldsymbol{\theta}) = \sum_{i=1}^M \log \sum_{k=1}^K \alpha_k p_k(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}_k) \quad (4)$$

B. EM Formulation for Mixed Regression Models

The idea here is to pull out the mixture of components from the full joint density, using the observed dataset D as a guide, so that the underlying group behavior can be discovered. A standard approach to deal with the hidden data is to utilize the EM algorithm for consistent estimation. Introducing hidden data, which correspond to the unknown group membership of

the M functional curve data, is more appropriate in the log-likelihood formulation of Eq. (4). With this, let $\mathbf{Z}$ be a matrix of reporting vectors $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ such that $z_{ik} = 1$ for some k and $z_{it} = 0, \forall t \neq k$. Moreover, if $z_{ik} = 1$, then the i th continuous spectrum is generated from the k th mixture component. Thus, the joint density of $\mathbf{Y}$ and $\mathbf{Z}$ given $\mathbf{X}$ can be defined as follows:

$$P(\mathbf{Y}, \mathbf{Z} / \mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^M p(\mathbf{y}_i / \mathbf{z}_i, \mathbf{x}_i, \boldsymbol{\theta}) p(\mathbf{Z}_i) \quad (5)$$

This follows from the conditional independence assumptions on $\mathbf{y}_i$ and that the $\mathbf{Z}_i$'s are assumed to be independent. Thus, the complete-data likelihood follows from Eq. (5) as:

$$\ell(\boldsymbol{\theta} / D, \mathbf{Z}) = \sum_{i=1}^M \sum_{k=1}^K z_{ik} \log \alpha_k + \sum_{i=1}^M \sum_{k=1}^K z_{ik} \log p_k(\mathbf{y}_i / \mathbf{x}_i, \boldsymbol{\theta}_k) \quad (6)$$

To fit the dataset D into this framework for the spline-based regression models, the regression between $\mathbf{y}_i$ and $\mathbf{x}_i$ defined in Eq. (1) can be rewritten as follows:

$$\mathbf{y}_i = \mathbf{B}_i \boldsymbol{\beta}_k + \mathbf{e}_i, \quad \mathbf{e}_i \sim N(0, \sigma_k^2 \mathbf{I}) \quad (7)$$

where $\mathbf{B}_i$ denotes spline basis matrix evaluated at $\mathbf{x}_i$. Then, the regression and the error models derived from the class-specific conditional probability density function for $\mathbf{y}_i$ can be rewritten as follows:

$$p(\mathbf{y}_i / \mathbf{x}_i, \boldsymbol{\theta}_k) = N(\mathbf{y}_i / \mathbf{B}_i \boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}) \quad (8)$$

Thus, Eq. (8) defines the probabilistic model for spline based mixture regression.

C. Multi-Class Alignment of LC-MS Data

The probabilistic framework formulated in the previous subsections allows us to generalize the modeling of arbitrary variability in time and measurement spaces of the data such as shifting and scaling of the LC-MS spectra. We describe in this section the steps that define the principle upon multi-class alignment models are based. These steps provide the data generative model structure, define the transformation prior models, compute the joint probabilities and the log-likelihood function, and define the associated EM algorithm.

The general model definition and the assumed prior models for scaling and shifting parameters are formulated as follows:

$$\mathbf{y}_i = d_{i,k} \mathbf{1} + c_{i,k} \boldsymbol{\Psi}_k ([a_{i,k} \mathbf{x}_i - b_{i,k} \mathbf{1}], \boldsymbol{\beta}_k) + \mathbf{e}_i \quad (9)$$

where $\{a_{i,k}, b_{i,k}\}$ and $\{c_{i,k}, d_{i,k}\}$ are scaling and translation variables for time and measurement spaces specific to each functional class data; and $\mathbf{1}$ is a column vector of ones with an appropriate dimension. The posterior probabilities of these parameters are estimated together with other parameters of the regression models and the error variances from the data.

The model definition given in Eq. (9) and the corresponding mixture regression models allow for arbitrary translation and scaling in time and measurement spaces. The conditional probabilistic density model $p(\mathbf{Y}_i/\mathbf{X}_i, \boldsymbol{\theta}, a_{i,k}, b_{i,k}, c_{i,k}, d_{i,k})$ is conditioned on the global parameter sets $\boldsymbol{\theta} = \{\beta_k, \sigma_k^2\}$ for $k = 1, \dots, K$. The transformation priors $\{a_{i,k}, b_{i,k}\}$ and $\{c_{i,k}, d_{i,k}\}$ are associated to data using the model given in Eq. (9). Thus, together with prior models, the problem of deriving the resulting joint probabilities can be directly handled. By considering $\{a_{i,k}, b_{i,k}\}$ and $\{c_{i,k}, d_{i,k}\}$ as class specific random variables with their associated prior probability density functions, one can proceed with the mixture regression models for further multi-class alignment analysis. With the assumption of independency, the joint prior probabilistic model for the time scaling and translation $p(a_{i,k}, b_{i,k}) = p(a_{i,k}) p(b_{i,k})$ should encode the idea that the most likely translation is the zero time translation and should discount the likelihood of large translations.

A zero-mean Gaussian prior is a good fit for this, i.e., $p(b_{i,k}) = N(b_{i,k}/0, s_k^2)$, where s_k^2 is the variance. Moreover, for the time-scaling prior $a_{i,k}$, it is desirable that a value of one be the most likely value (i.e., no scaling at all) and therefore the corresponding Gaussian prior model is given as $p(a_{i,k}) = N(a_{i,k}/1, r_k^2)$. The variances r_k^2 and s_k^2 will be learned from the data within the ensuing EM algorithm. Similarly, the joint prior probabilistic model for the affine transformation in measurement space with independence assumption is given as $p(c_{i,k}, d_{i,k}) = p(c_{i,k}) p(d_{i,k})$. Here, we follow similar argument as that of the affine time transformation described above. In other words, we assume that $p(c_{i,k}) = N(c_{i,k}/1, u_k^2)$ and $p(d_{i,k}) = N(d_{i,k}/0, v_k^2)$ are the prior probabilistic models and specify the most likely transformations. After obtaining both the class-specific conditional probability density function for $\mathbf{y}_i$, i.e. $p(\mathbf{y}_i/\mathbf{x}_i, \boldsymbol{\theta}, a_{i,k}, b_{i,k}, c_{i,k}, d_{i,k})$, and the complete-data likelihood function $\ell(\boldsymbol{\theta}/D, \mathbf{Z})$ (where $\boldsymbol{\theta} = \{\beta_k, \sigma_k^2, r_k^2, s_k^2, u_k^2, v_k^2\}$ for $k = 1, \dots, K$), the EM algorithm executes sequentially the following two steps:

- i. *E-Step*: take the expectation of $\ell(\boldsymbol{\theta}/D, \mathbf{Z})$ with respect to $p(\mathbf{Z}/\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}^{t-1})$, where $\boldsymbol{\theta}^{t-1}$ is the current set of parameters.
- ii. *M-Step*: maximize the expected value of the computed log-likelihood from *E-Step* over the parameters $\boldsymbol{\theta}$ to yield new parameters $\boldsymbol{\theta}^t$.

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

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

where $h_{ik} = E\{z_{ik}\} \propto p(\mathbf{y}_i/z_{ik} = 1, \mathbf{x}_i, \boldsymbol{\theta}^{t-1})p(z_{ik} = 1)$ and it corresponds to the posterior probability where the $\mathbf{y}_i$'s are generated by component k . Therefore, the *E*- and *M*-steps of the EM algorithm proceed iteratively to estimate the following parameters:

$\boldsymbol{\theta} = \{\beta_k, \sigma_k^2, r_k^2, s_k^2, u_k^2, v_k^2\}$. The following are the steps involved in estimating the parameters of the spline-based mixture regression and the prior transformation models using the EM algorithm:

1. Initialize randomly the membership probabilities $h_{ik} = E\{z_{ik}\}$ and compute the parameters estimate for $\boldsymbol{\theta}^0 = \{\widehat{\beta}_k, \widehat{\sigma}_k^2, \widehat{r}_k^2, \widehat{s}_k^2, \widehat{u}_k^2, \widehat{v}_k^2\}^0$ using these particular membership probabilities.
2. Compute the new membership probabilities $h_{ik} = E\{z_{ik}\}$ using the current set of parameters $\boldsymbol{\theta}^{t-1} = \{\widehat{\beta}_k, \widehat{\sigma}_k^2, \widehat{r}_k^2, \widehat{s}_k^2, \widehat{u}_k^2, \widehat{v}_k^2\}^{t-1}$.

3. Estimate the new set of parameters $\Theta^l = [\hat{\beta}_k, \hat{\sigma}_k^2, \hat{r}_k^2, \hat{s}_k^2, \hat{u}_k^2, \hat{v}_k^2]^T$ using the new membership probabilities from Step 2.
4. Repeat Steps 2 and 3 until a stopping criterion is met (e.g. maximum number of iteration or convergence rate).

III. EVALUATION AND COMPARISON

To evaluate the framework proposed in this paper, we used LC-MS spectra data generated from proteins of *lysed E.coli* cells previously described in [9]. Each raw LC-MS spectrum consists of three dimensions, namely the m/z location, RT, and ion intensity. By calculating the total ion current (TIC) for each RT point across m/z values, we obtained 11 two-dimensional spectra (see Fig. 1A). From this figure, we can see that the replicate spectra show significant shifts along the experimental time as well as distortions in the measurement space. We applied our proposed framework that involves the probabilistic spline-based (i.e. a linear combination of B-spline bases) mixture regression models to align these spectra. The prior scaling and translation variables in time and measurement spaces, described in the generalized model of Eq. (9), were included in the maximum likelihood estimation-based EM algorithm. Fig. 1B depicts the aligned spectra using the PMRM. We calculated the coefficient of variation (CV) of the log-transformed TIC values for the raw spectra and the ones aligned using PMRM, DTW and CPM. Although each alignment method yielded smaller CV compared to the raw spectra, we see that DTW and CPM spectra stretched the original (experimental) time range by almost 1.5 and 2 folds, respectively, as illustrated in Fig. 1C. The stretch in using DTW is due to the insertion and deletion steps with respect to the time axis that causes the aligned spectra to have varying time range across spectra. The stretch by CPM is attributed to the assumed latent timescale, i.e., the latent timescale is assumed to be at least two times wider than the physical timescale. On the other hand, the warping path obtained by PMRM is determined from the underlying physical experiment. Thus, its time range tends to remain close to the time-scale of the physical experiment.

IV. CONCLUSION

In this paper, we present probabilistic-based mixture regression models (PMRM) for aligning a single-class replicate LC-MS data. We utilize the well known maximum likelihood-based EM algorithm for estimating the generative models of the data as well as the prior scaling and translation parameters. The latter effectively accounts for the variability in time and measurement spaces of the data. We compared the performance of PMRM with dynamic time warping (DTW) and continuous profile model (CPM) in terms of aligning replicate LC-MS spectra. Although the three methods exhibited comparable performance in terms of reducing the CV of the raw spectra, PMRM achieved better performance in preserving the underlying physical experimental time-frame. Whereas DTW and CPM stretched the original time-scale by approximately 1.5 and 2 folds, respectively, PMRM maintained the original time-scale of the raw spectra. Future work will focus on extending PMRM for multi-dimensional alignment (i.e., to align both m/z and RT drifts) and to apply the method for multi-class alignment (e.g., to align LC-MS data obtained from samples that represent cases and controls).

Acknowledgments

This work was supported in part by the National Cancer Institute (NCI) R21 CA130837 Grant, NCI R03 CA119313 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. Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol. Cell Proteomics* 2004;3(4):367–378. [PubMed: 14990683]
3. Ramsay, JO.; Silverman, BW. *Applied functional data analysis: methods and case studies*. New York: Springer-Verlag; 2002.
4. Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Mol. Cell Proteomics* 2004;3(10):984–997. [PubMed: 15269249]
5. Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing* 1978;26(1):43–49.
6. Bylund D, Danielsson R, Malmquist G, Markides KE. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography-mass spectrometry data. *Journal of Chromatogr A* 2002;961(2):237–244.
7. Prince JT, Marcotte EM. Chromatographic alignment of ESILC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem* 2006;78(17):6140–6152. [PubMed: 16944896]
8. Tomasi G, van den Berg F, Andersson C. Correlation Optimized Warping and Dynamic Time Warping as Preprocessing Methods for Chromatographic Data. *Journal of Chemometrics and Intelligent Laboratory Systems* 2004;18:231–241.
9. Listgarten J, Neal RM, Roweis ST, Emily A. Multiple Alignment of Continuous Time Series. *Neural Information Processing Systems* 2005;17:817–824.
10. Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 1977;39(1):1–38.
11. Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. *Neural Comput* 1994;6(2):181–214.
12. Redner RA, Walker HF. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review* 1984;26(2):195–239.

Fig. 1.

Alignment results for 11 replicate LC-MS spectra of proteins extracted from lysed *E. coli* cells and CV calculated for each set of replicated spectra. A: Raw spectra before alignment. B: Spectra aligned by PMRM. C: Mean spectrum for each set of spectra (raw and spectra aligned by DTW, CPM and PMRM).