Biostatistics (2007), **8**, 2, *pp.* 357–367
doi:10.1093/biostatistics/kxl015
Advance Access publication on July 31, 2006

A statistical method for chromatographic alignment of LC-MS data

PEI WANG*,†, HUA TANG, MATTHEW P. FITZGIBBON, MARTIN MCINTOSH

Fred Hutchinson Cancer Research Center, 1100 Fairveiw Avenue N, M2-B500, PO Box 19204, Seattle, WA, USA
pwang@fhcrc.org

MARC CORAM†

Department of Statistics, University of Chicago, Chicago, IL, USA

HUI ZHANG, EUGENE YI, RUEDI AEBERSOLD

Institute for System Biology, Seattle, WA, USA

SUMMARY

Integrated liquid-chromatography mass-spectrometry (LC-MS) is becoming a widely used approach for quantifying the protein composition of complex samples. The output of the LC-MS system measures the intensity of a peptide with a specific mass-charge ratio and retention time. In the last few years, this technology has been used to compare complex biological samples across multiple conditions. One challenge for comparative proteomic profiling with LC-MS is to match corresponding peptide features from different experiments. In this paper, we propose a new method—Peptide Element Alignment (PETAL) that uses raw spectrum data and detected peak to simultaneously align features from multiple LC-MS experiments. PETAL creates spectrum elements, each of which represents the mass spectrum of a single peptide in a single scan. Peptides detected in different LC-MS data are aligned if they can be represented by the same elements. By considering each peptide separately, PETAL enjoys greater flexibility than time warping methods. While most existing methods process multiple data sets by sequentially aligning each data set to an arbitrarily chosen template data set, PETAL treats all experiments symmetrically and can analyze all experiments simultaneously. We illustrate the performance of PETAL on example data sets.

Keywords: Alignment; LC-MS; Regression; Retention time.

1. INTRODUCTION

An integrated system of liquid-chromatography mass-spectrometry (LC-MS) offers a versatile and high throughput proteomics technology. In such a system, LC efficiently separates a peptide mixture (peptides are short amino acid sequences) based on hydrophobicity; thousands of peptides can then be identified and quantified using MS to address important biology questions (Mann and Aebersold, 2003).

*To whom correspondence should be addressed.
†Equal contributors.

While high precision LC-MS systems are available, bioinformatics tools remain incomplete. LC-MS systems generate massive amounts of data, representing the intensity of peptides with specific mass-charge ratios (mz) and LC column retention times (RT) (see Section 2.1 for more details). Statistical and computational methods are required to detect and quantify the intensity of each feature. A more challenging task is to compare multiple LC-MS profiles, which, for example, can be used to identify discriminating peptides between distinct biological groups. Because the sequence identifications of the peptide are often unavailable at this stage, one relies on RT and mz to match corresponding peptides across different samples. However, the retention time of a specific peptide depends on instrument conditions as well as the underlying composition of the mixture; variation in RT between experiments is often nonnegligible even when all samples are processed by the same LC-MS system. To a lesser extent, mz of a peptide also varies as a result of instrument noise. For these reasons, a prerequisite for quantitative analysis of multiple LC-MS experiments is to align output data with respect to both RT and mz. In Section 2.2, we review two groups of existing methods. The first group align raw spectrum data before peak detection. These methods search for optimal warping functions to map RT of one experiment to that of another. Since the warping function only accounts for "global" variation in RT, these methods may not always align individual peptides. The second group of alignment methods use the detected feature lists, and allow some variation in RT of individual peptides. However, since this method relies on the detected peak and does not take advantage of the raw spectrum information, the alignment decisions are vulnerable to inaccuracy in the peak detection step. In addition, both groups of methods are formulated to work on data sets that are similar to each other, and may produce bias when analyzing different samples, such as cancer and noncancer serum. In order for LC-MS-based analysis to become a routine procedure in biomedical research, a computationally efficient and robust alignment procedure must be developed.

In this paper, we propose a statistical method, called "Peptide Element Alignment" (PETAL), which uses both raw spectrum data and peak detection results to simultaneously align features from multiple LC-MS experiments. PETAL first creates spectrum elements to represent the relative intensity profiles of individual peptides. It then models the variation in retention time and the instrument noise in intensity measurements that produce error in the mz values. Peptides detected in different LC-MS data are aligned if they are represented by the same element. By considering each peptide separately, this method offers greater flexibility than simply matching retention time between profiles. In addition, PETAL treats all experiments symmetrically and avoids the possible biases that may result from choosing one experiment as a template.

The rest of the paper is organized as follows: Section 2 provides a brief description of the LC-MS experiments. The PETAL method is described in Section 3. Section 4 is devoted to real data examples. In Section 5, we make several remarks regarding the strength and weaknesses of our method in comparison to existing methods and discuss the choices of parameters in the model.

2. LC-MS EXPERIMENT

2.1 *Generic LC-MS experiment*

Figure 1 is a cartoon of a typical LC-MS experiment. First, protein mixtures are isolated from biological samples and enzymatically digested into peptides (short amino acid sequences). The peptides are then separated by one or more steps of high-pressure LC, and are eluted into an electro-spray ion source, where they are nebulized in small, highly charged droplets. After evaporation, multiple protonated peptides enter the mass spectrometer, and a mass spectrum of the peptide eluting at each time point is taken (Mann and Aebersold, 2003). A more detailed introduction to LC-MS can be found in Liebler (2002).

The output of an LC-MS experiment can be represented as a two-dimensional image. One dimension represents the elution time (also called retention time and denoted as RT) and the other dimension indicates the mass-charge ratio. Although RT is a continuous variable, the LC-MS system produces mass spectra at

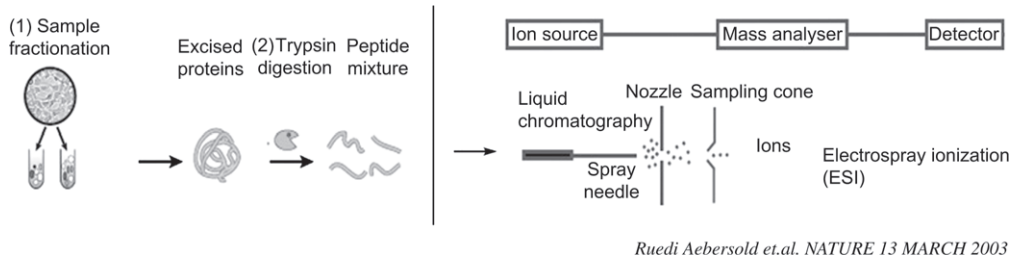

Fig. 1. Outline of one LC-MS experiment. See text for details.

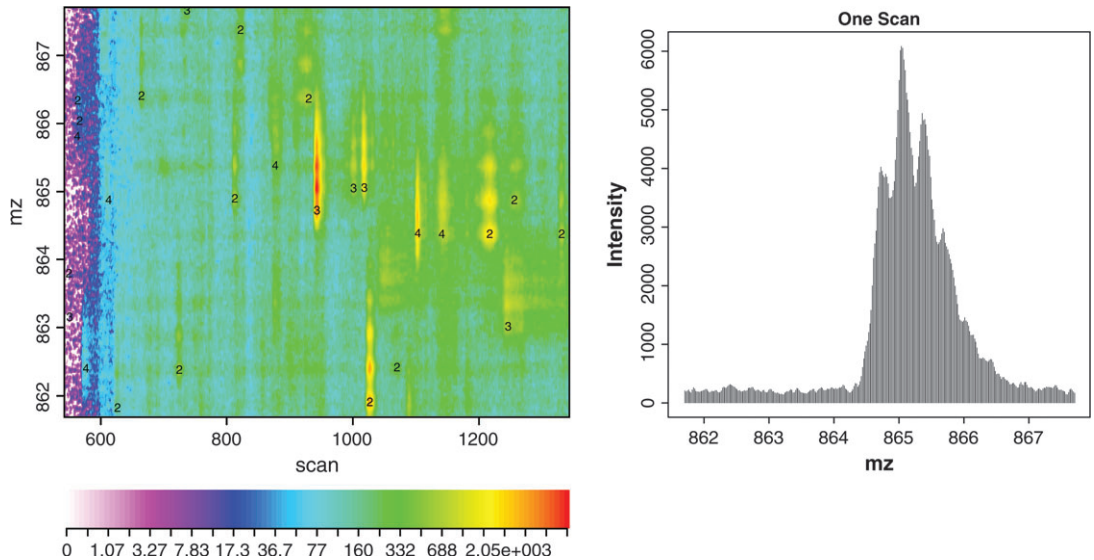

Fig. 2. Output from a LC-MS experiment. Left: Output of one LC-MS experiment in the region mz × RT $\in$ $(861.5, 867.9) \times (540, 1300)$. The horizontal axis represents the retention time and the vertical axis represents the mass-charge ratio. The color at each (mz, RT) indicates peptide intensity (scales are defined in the color bar). Each peak feature identified in previous analysis steps is labelled by number (in black): the vertical coordinate of the number is its monoisotopic mass; the horizontal coordinate of the number is the index of the scan, in which the feature is detected at the highest intensity; the value of the number indicates the estimated charge status. The plot is made with R-package Nimbus (by Marc Coram, available at http://galton.uchicago.edu/∼coram/). Right: Mass spectrum of the scan 943 for mz $\in$ $(861.5, 867.9)$, which corresponds to one column of the left image. The isotopic shape suggests that this peptide has a charge of 3.

a discrete set of RT points, typically a few seconds apart. Thus, it is equivalent to represent RT by scan indices. The mass spectrum at one RT point, i.e. in a single scan, measures the abundance of peptide ions at each mz (each mass-charge ratio point). Figure 2 illustrates part of an LC-MS experiment result.

As shown in Figure 2, the mass spectrum of a peptide feature has a characteristic shape, consisting of multiple peaks equally spaced along mz. This shape, referred to as isotopic pattern (or isotopic distribution), arises as a result of the naturally existing rare isotopes in the sample. The dominant source for isotopic distribution in mass spectrum is carbon-13, which accounts for 1.11% of all naturally occurring carbon atoms. For a peptide with a charge of 1, the molecule with no carbon-13 and the molecule with

one carbon-13 accumulate at locations that are one unit apart on the mass spectrum. More generally, for a peptide of charge k, the gap between two adjacent isotopic peaks is $\frac{1}{k}$. The peak where the analyte consists of only light isotopes is called the mono-isotopic peak, indicating the ordinary mz value of this peptide. Thus, in MS experiments, each peptide can be characterized by its mz value (the position of the mono-isotope) and charge status (the isotope shape). This information can be used for peptide peak detection as well as for subsequent alignment.

2.2 *Existing alignment methods*

The goal of alignment is to match corresponding peptide features in the mz-scan plot (e.g. Figure 2) from different experiments in the presence of retention time variation and experimental noise. Bylund *and others* (2002) proposed a time warping method based on raw spectrum for alignment of LC-MS data, which is a modification of the original correlated optimized warping algorithm (Nielsen *and others*, 1998). After choosing one file as a template, the method warps the time coordinate (RT axis) of another file to give maximal similarity between the two images. This framework was also used by Wang *and others* (2003), who implemented a dynamic time warping algorithm allowing every RT point to be moved. However, compared with the classical one-dimension chromatography profiles, LC-MS data have an added dimension of mass spectral information, which makes the alignment problem more complicated. Different peptides with different mz values may have different retention time shifting between two experiments. In other words, two peptides eluting at the same time in one experiment may not necessarily elute at the same time in another experiment. Therefore, only mapping the retention time coordinates between two LC-MS files is not sufficient to provide alignment for individual peptides.

Instead of using raw spectrum data, Radulovic *and others* (2004) performed alignment based on the (mz, RT) values of detected features. It first divides the mz domain into several intervals and fitted different piece-wise linear time warping functions for each mz interval. After the time warping, a "wobble" function is then applied wherein a peak is allowed to move ($\pm$1–2% of total scan range) in order to match with the nearest adjacent peak in another file. Here, the stratification of mz achieves improved flexibility and accuracy. Since the method relies on only the (mz, RT) values of detected peptide features, it fails to take advantage of other information in the raw image (such as isotope distribution). In addition, the wobble function may produce ambiguous findings when complex mixtures like human serum are processed, where multiple peptides may exist within the $\pm$1–2% window.

Recent software platforms, "msInspect" (Bellew *and others*, 2006), "SpecArray" (Li *and others*, 2005), and "MZmine" (Katajamaa and Orešic, 2005) provide alignment solutions by allowing variation in RT of individual peptides within the detected feature lists. However, since these methods rely on the peak detection result and do not take advantage of the raw spectrum information, the alignment decisions are vulnerable to any inaccuracy estimation in the peak detection step. Moreover, most methods process multiple data sets by sequentially aligning each to an arbitrarily chosen template profile, which may lead to unpredictable errors. Most methods work best on data sets that vary similar to each other. They are likely to produce bias when analyzing samples from different disease classes such as cancer and noncancer tissues.

Other related algorithms are discussed in Listgarten *and others* (2005) including a hierarchical clustering method for aligning MALDI/SELDI spectra (Tibshirani *and others*, 2004), a multi-scale wavelet decomposition approach for aligning MALDI data along the mz axis (Randolph and Yasui, 2004), and a Hidden Markov Model for multiple alignments of time series. Prakash *and others* (2006) recently proposed a novel signal mapping algorithm to perform comparisons directly on the signal level of MS experiments.

To overcome the drawbacks of current methods, we propose a new alignment algorithm, PETAL, for LC-MS data. It uses both the raw spectrum data and the information of the detected peak features for peptide alignment.

3. PETAL FOR LC-MS

In LC-MS profiles, each peptide is characterized by two things: its mass spectrum and its retention time range. The mass spectrum of one scan is a vector recording the intensity measurements along mz for peptides eluting in this scan. For any given peptide, its "element spectrum vector" is defined as the mass spectrum with no experimental noise that contains only one unit abundance of this peptide $\vec{b}$: $\sum_{l=1}^{L} b_l = 1$, where L is the total pixel number along mz and b_l is the intensity value at the lth pixel. One spectrum element vector $\vec{b}$ can be uniquely determined by the mono-isotope position and the charge status (isotopic pattern) of the corresponding peptide. In addition, we denote the theoretical retention time range of one peptide as $\mathrm{RT} = (\mathrm{rt}_{\mathrm{begin}}, \mathrm{rt}_{\mathrm{end}})$. The kth peptide can then be represented as $\mathrm{PEP}_k = \left(\vec{b^k}, \mathrm{RT}^k\right)$.

We define a peptide element library $\{\mathrm{PEP}_k\}_{k=1}^{K}$ as a collection of all possible peptides appearing in the target samples. Given a library of peptide elements, the goal of alignment can be easily achieved by matching the peak features in each profile to this common library. Peak features from different profiles matched to the same peptide element are features representing the same peptide and should be aligned.

We now introduce a loss function and seek the solution of alignment by solving an optimization problem.

3.1 *Loss function*

We first consider the mass spectrum of one scan. Suppose there are H different peptides $\{\mathrm{PEP}_{k_h}\}_{h=1}^{H}$ eluting in this scan, and the measurable abundance (the abundance of peptides that can be measured in LC-MS experiment) of PEP_{k_h} is β_{k_h}. The entire mass spectrum of one scan is the sum of all individual peptide spectra eluting in the scan. The observed mass spectrum $\vec{Y}$ can therefore be represented as a linear combination of the spectrum element vectors of the H peptides: $\vec{Y} = \sum_{h=1}^{H} \beta_{k_h} \cdot \vec{b^{k_h}} + \vec{\epsilon}$, where $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_L)$ is the instrument noise and $\vec{b^{k_h}}$ is the spectrum element vector of PEP_{k_h}.

In reality, we would not know which peptides elute in an observed scan $\vec{Y}$. However, with a peptide element library $\{\mathrm{PEP}_k\}_{k=1}^{K}$, we can estimate the peptide abundances by fitting a L_1 penalized least square regression model $\vec{Y} \sim \sum_{k=1}^{K} \beta_k \vec{b_k}$:

$$\{\beta_k\}_{k=1}^{K} = \mathrm{argmin}_{\{\beta_k\}_k, \beta_k \geqslant 0} \left\| \vec{Y} - \sum_{k=1}^{K} \beta_k \vec{b^k} \right\|^2 + \lambda_1 \sum_{k=1}^{K} w(t; \mathrm{RT}^k) \cdot |\beta_k|, \tag{3.1}$$

where λ_1 is a nonnegative parameter, t is the retention time of scan $\vec{Y}$, and $w(t; \mathrm{RT}^k)$ is a weight function depending on the scan retention time t as well as the theoretical retention time of each peptide RT^k. The L_1 norm penalty controls in the coefficient solution the total number of nonzero coefficients (Tibshirani, 1996). The weight function $w(t; \mathrm{RT}^k)$ gives larger penalty to peptides whose theoretical retention time RT^k is further from the scan retention time t, so that the corresponding predictors (b^k) are selected less often. A simple example for $w(t; \mathrm{RT}^k)$ is

$$w(t, \mathrm{RT}_k) = \begin{cases} 1, & \text{if } t \in [\mathrm{RT}^k_{\mathrm{begin}} - \delta, \mathrm{RT}^k_{\mathrm{end}} + \delta], \\ \infty, & \text{otherwise,} \end{cases} \tag{3.2}$$

where δ is a nonnegative parameter.

In the solution of (3.1), a nonzero estimate of β_k indicates that part of the signal in $\vec{Y}$ matches the kth peptide in the peptide element library. Thus, if we have the proper regression models for two scans, $\vec{Y}$

and $\vec{Y}'$ from two different profiles, the alignment between these two scans can be achieved by comparing the coefficient sets $\{\beta_k\}_k$ and $\{\beta'_k\}_k$.

Suppose there are N profiles and each profile has M_n observed scans. Given a peptide element library $\{\text{PEP}_k\}_{k=1}^K$, we are interested in $\{\beta_{n,k}^m\}_{n,m,k}$ satisfying

$$\{\beta_{n,k}^m\}_{n,m,k} = \text{argmin}_{\{\beta_{n,k}^m:\beta_{n,k}^m>0\}} \sum_{n=1}^{N}\sum_{m=1}^{M_n}\left\{\left\|\vec{Y_n^m} - \sum_{k=1}^{K}\beta_{n,k}^m\,\vec{b^k}\right\|^2 + \lambda_1\sum_{k=1}^{K}\beta_{n,k}^m\cdot w(t_n^m;\text{RT}^k)\right\}, \quad (3.3)$$

where $\vec{Y_n^m}$ and t_n^m are the mass spectrum vector and the retention time of the mth scan in the nth profile.

In most cases, the peptide element library $\{\text{PEP}_k\}_{k=1}^K$ is not available at this point. We therefore also need to identify the peptide element library ($\{\text{PEP}_k\}_{k=1}^K$) that best explains all mass spectrum scans observed in the experiments ($\{\vec{Y_n^m}\}$). Thus, we introduce the overall "loss function":

$$\widetilde{\text{L}}(\{\vec{Y_n^m}, t_n^m\}_{n,m}|\{\{\beta_{n,k}^m\}_{n,m}, \text{PEP}_k\}_k) = \sum_{n=1}^{N}\sum_{m=1}^{M_n}\left[\left\|\vec{Y_n^m} - \sum_{k=1}^{K}\beta_{n,k}^m\,\vec{b^k}\right\|^2 + \lambda_1\sum_{k=1}^{K}\beta_{n,k}^m\cdot w(t_n^m;\text{RT}^k)\right] + \lambda_2 K, \quad (3.4)$$

and search for

$$\{\{\beta_{n,k}^m\}_{n,m}, \text{PEP}_k\}_k = \text{argmin}_{\{\beta_{n,k}^m:\beta_{n,k}^m\geqslant 0;\text{PEP}_k\}}\widetilde{\text{L}}(\{\vec{Y_n^m}, t_n^m\}_{n,m}|\{\{\beta_{n,k}^m\}_{n,m}, \text{PEP}_k\}_k), \quad (3.5)$$

where $\lambda_2 K$ is a penalty term for overall model complexity. The choices of λ_1 and λ_2 are discussed in Section 5.

The main part of the loss function in (3.4) can also be deemed as the negative log joint likelihood of $\{\vec{Y_n^m}\}_{n,m}$ under some reasonable assumptions as discussed in Section A of the supplementary material available at *Biostatistics* online.

Note that besides the random variation in retention time due to individual peptides, there is always some systematic retention time shifting across LC-MS experiments. Thus, we first apply a global transformation to adjust for the systematic trend and then use the adjusted time to calculate $w(t^m;\text{RT}^k)$. The details of the global transformation are described in Section B of the supplementary material available at *Biostatistics* online.

3.2 *Optimization strategy*

From the loss function in (3.5), we can see that if $\{\text{PEP}_k\}_k$ is given, the optimal solution for $\{\beta_{n,k}^m\}$ can be easily calculated with L_1-regression techniques such as "lasso" (Tibshirani, 1996) and "lars" (Efron *and others*, 2003). Thus, our main obstacle is to find the appropriate peptide element library $\{\text{PEP}_k\}_k$.

It is difficult to directly search the whole vector space of element spectra. We therefore approach this problem using two steps. First, we build an initial collection of peptide elements based on all profiles subjected to alignment. This initial collection is expected to represent all peptides appearing in the experiments, but it may also contain redundant or incorrect elements. Then, we search for the subset of the initial collection that minimizes the target loss function. The details of these two steps are described below.

Initial collection. To build the initial set, we include one peptide element for every peak feature detected in the profiles. To do so, we first need to estimate the ideal isotopic shapes. Since, at a given mass, the variation of isotopic shapes resulting from the differences between amino acid sequences is much less than the variation introduced by experimental noise, we assume that the peptides with similar mass values and at the same charge status have the same isotopic pattern. Thus, the "ideal" isotopic shape of a certain charge and mass can be approximated by averaging all feature spectra of the same charge status and similar mass values. With these empirical isotopic shapes, we make spectrum element vectors $\{\vec{b_k}\}$ based on the estimated mono-isotope positions and charge values of detected features. Details are provided in Section C of supplementary material available at *Biostatistics* online.

We denote this initial collection as Ω_0.

Subset selection. There are two major strategies for subset selection, forward-stepwise and backward-stepwise. In this section, we focus on the backward-stepwise strategy, which enjoys a higher computational efficiency than the forward-stepwise strategy (discussed in Section D of the supplementary material available at *Biostatistics* online).

Backward-stepwise begins with the whole collection Ω_0 and removes redundant elements iteratively. Instead of eliminating the redundant elements one by one as is usually done, we propose a more efficient procedure. As mentioned before, each peptide in the experiments may correspond to more than one peptide element in the initial collection contributed by different profiles. If we can cluster peptide elements in some appropriate way, such that elements representing the same peptide are grouped together, we will be able to eliminate the redundancy of multiple clusters simultaneously.

For this purpose, we apply a sparse regression approach called elastic net (Zou and Hastie, 2005), which aims to minimize the loss function $L(\lambda_1, \lambda_2, \beta) = |y - X\beta|^2 + \lambda_2|\beta|^2 + \lambda_1|\beta|$. The ridge penalty term encourages a grouping effect: strongly correlated predictors tend to be in or out of the model together. And the Lasso penalty term enables the algorithm to have a more sparse representation.

The new backward-stepwise procedure is as follows:

1. Take all M feature scans of target profiles and the initial collection of peptide elements Ω_0 ($|\Omega_0| = N_0$).
2. For $j = 1$ to M, do elastic net regression for $\vec{Y^j} \sim \left\{\frac{1}{d(\mathrm{RT}^j, \mathrm{RT}_k)}\ \vec{b_k}\right\}_{k=1}^{N_0}$ with a fixed number of maximum steps. Thus, each element gets M coefficients from M regression models.
3. Cluster elements based on the coefficient vector (length of M), such that elements representing the same peptide are grouped together. Representing each cluster with one element, we get a new set of $N_1 (< N_0)$ elements.
4. Repeat steps 2–3 until $N_k = N_{k-1}$.

We choose not to cluster directly on the original element spectrum vector space because it is not straightforward to define an appropriate distance measurement between elements, taking into consideration of the meaning of isotopic pattern and retention time. However, after we map elements to the coefficient space through regression, we can easily use Euclidean distance for clustering. In addition, the regression procedure enjoys a "selection" effect, such that incorrect basis will not enter the models and will be eliminated from the library directly.

The performance of the algorithm is illustrated with data examples in Section 4.

4. Data Example

PETAL is applied on a data example from a spike-in experiment, and its performance is compared with the performance of two other alignment methods implemented in public available softwares msInspec (Bellew *and others*, 2006) and SpecArray (Li *and others*, 2005). (A more detailed illustration on how PETAL

Table 1. *Alignment results. Column names: FD, feature detection method; Align, alignment method; TN, total number of features in all files after alignment; N3R, number of features appearing in all three replicates of any biological sample (bovine protein, human serum, or bovine + serum mixture); N2R, number of features appearing in at least two replicates of any biological sample; NBF, number of features in the bovine + serum mixture that corresponds to bovine proteins (see text for details)*

Label	FD	Align	TN	N3R	N2R	NBF
MM	msI	msI	18001	3942 (21.9%)	7686 (42.7%)	32
MP	msI	PETAL	12397	5269 (42.5%)	8058 (65.0%)	56
SS	specA	specA	4555	2193 (48.1%)	3088 (67.8%)	12
SP	specA	PETAL	3718	2383 (64.2%)	3510 (94.2%)	58

works to solve the challenges of the alignment problem is shown in Section E of the supplementary material available at *Biostatistics* online, where PETAL is applied on a data example of human serum samples.)

In the spike-in experiment, three different biological samples were analyzed with LC-MS instruments[†]. The three samples were (1) 20 μg of four bovine glycoprotein mix, (2) 80 μl of normal human serum sample, and (3) a mix sample of bovine glycoproteins and 80 μl of human serum in a concentration of 20 μg/ml bovine glycoproteins in human serum. Three LC-MS replica were collected for each biological sample, which resulted in a total of nine LC-MS profiles.

Peak signals corresponding to peptide features in each LC-MS profile were first detected using both msInspect (msI) and specArray (specA). msI returns ∼6000 peptide features for each LC-MS profile and specA returns ∼3000 peptide features. Comparing quality of the feature detection algorithms requires more than comparing features counts for each individual profile. However, since the feature detection step is not the focus of this paper, we will not discuss in further here.

The alignment method of specA makes use of information computed specifically by its own feature detection methods, whereas msI uses only mz, RT, and charge information. Thus, to better characterize the advantages of the different alignment methods, we compare the performance of PETAL and the alignment method in msI using feature lists returned by msI, and we compare the performance of PETAL and the alignment method in specA using feature lists returned by specA.

We assessed the performance of alignment using two criteria: one is the efficiency of recognizing features corresponding to the same peptide and the other is the degree of false-alignment—incorrectly matched features corresponding to different peptides.

First, we use replicate profiles to examine the alignment efficiency. Since the majority of the peptides in a sample should behave the same across replicate LC-MS experiments, we expect to see majority of the features aligned across replicate profiles. In Table 1, column N3R (column N2R) shows the number (percentage) of features aligned across the three replicates (two replicates) of any biological sample by different alignment methods. We can see that PETAL recognized many more matching peptide features across replicate profiles than either msI (8058 vs. 7686) or specA (3510 vs. 3088).

On the other hand, since the same peptide should have similar intensities across LC-MS replica experiments and since none of the alignment methods takes into consideration the intensity information when matching features across different profiles, we can use the correlation of intensities of aligned features between two replicate profiles to assess the alignment quality: the more the false-aligned pairs, the less correlated the intensities of aligned features tend to be. The correlation coefficient of log-intensities of aligned features between each replicate pair is illustrated in Figure 3 (log scale is used to adjust the heavy

[†]The LC-MS system consists of a Bruker Daltonics Micro-TOF mass spectrometer equipped and a home-built nanospray device. Glycopeptides were first isolated from proteins in 80 μl (Zhang, 2005; Zhang *and others*, 2003), and peptides from 5 μl of original serum were used in each MS analysis.

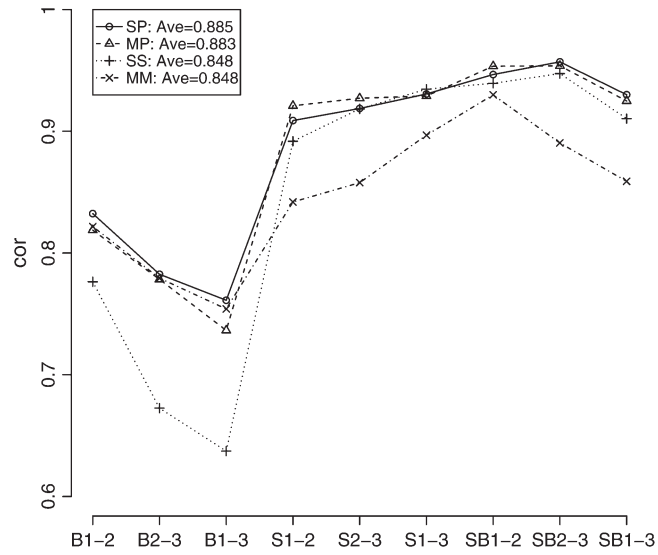

Fig. 3. Correlation coefficients of log-intensities of aligned features. The *x*-axis indicates the sample pair: B1–2, B1–3, and B2–3 are the three replicate pairs of the bovine protein sample; S1–2, S1–3, and S2–3 for the human serum sample; SB1–2, SB1–3, and SB2–3 for the bovine + serum sample. The *y*-axis represents the values of the correlation coefficients. The labels in the legend indicate the different combination of feature detection and alignment methods (see Table 1).

right tail of the intensity distribution). From the figure, we can see that feature pairs aligned by PETAL have more similar intensities than the feature pairs aligned by msI and specA, which suggests that PETAL achieves better alignment quality.

In addition, with the spike-in design of the experiment, it is of interest to investigate the efficiency of detecting the spiked-in bovine peptides in the bovine + serum sample for different methods. Peptide features are deemed as candidate spiked-in bovine peptides if they appear in at least two replicates of the bovine + serum sample, as well as in at least two replicates of the bovine protein sample, but not in any replicates of the serum sample. The numbers of candidate bovine peptides resulting from different methods are listed in column NBF of Table 1. PETAL detects more than 55 candidate bovine peptides on both sets of feature detection results, which is almost twice the number of candidate bovine peptides detected by msI and four times the number of specA. (With the newly developed LTQ-FT instrument, which simultaneously provides intensity measurements and tandem mass spectrum measurements for each target peptide ion, it is possible to further validate those candidate features as bovine peptides by deriving the peptide sequence IDs from the tandem mass spectra through database searching. However, due to the limitation of the facilities, such a data set is not available at this point.)

Overall, we conclude that with the same alignment quality as (if not better than) the other two alignment methods, PETAL achieves the highest alignment efficiency.

5. Discussion

In this paper, we introduce a new alignment method, PETAL, which uses both raw spectrum data and peak detection results to simultaneously align features from multiple LC-MS experiments. By considering each peptide separately, this method offers more flexibility than simply matching retention time between different profiles. It treats all experiments symmetrically and avoids the possible biases that may result

from choosing one experiment as a template. In addition, although PETAL is based on feature lists from the peak detection procedure, the ability to consider spectrum information and jointly learn from multiple profiles enables PETAL to improve the peak detection in return.

The backward-stepwise optimization strategy, whose computational complexity is about $O(N \cdot K)$ (where N is the total number of samples, and K is the total number of peptide features in the study) is more efficient compared with the forward-stepwise strategy, whose computational complexity is about $O(N \cdot K^3)$. For further computational simplicity, we can divide the entire mz domain into multiple mz blocks, and then conduct alignments for individual mz blocks parallel. The L_1 norm penalty parameter λ_1 is controlled by forcing the total number of nonzero coefficients smaller than S_{λ_1} in each regression model. For the data example in Section 4, we used 100 mz blocks with each block averaging 5–10 mz. We choose $S_{\lambda_1} = 3$ in the forward-stepwise strategy and $S_{\lambda_1} = N + 1$ in the backward-stepwise strategy, where N is the total number of samples. The model complexity penalty parameter λ_2 is also controlled differently in the two strategies. For forward-stepwise, controlling λ_2 is equivalent to controlling the stopping constant ϵ_{λ_2} with smaller ϵ_{λ_2} corresponding to larger value of K. For backward-stepwise, λ_2 corresponds to the cutoff criterion in the clustering steps. The number of clusters represents the number of selected elements in the library (K).

PETAL can be easily applied to the scenario where an AMT (Accurate Mass and Time Tag) database is available (Fang *and others*, 2006). In such cases, the peptide element library $\{\text{PEP}_k\}_k$ can be derived directly from the AMT database, and then only the regression coefficients $\{\beta^m_{n,k}\}_{n,m,k}$ need to be estimated. Furthermore, for LC-MS experiments with isotopic labeling, viewing scan spectra as linear combinations of peptide element spectra, as well as the regression techniques discussed in this paper, can be used to accurately estimate the intensity ratio of light versus heavy forms when the mass of the labeling materials does not allow for complete separation of the two forms.

The R-package implementing the PETAL algorithm is available at http://peiwang.fhcrc.org/research-project.html.

ACKNOWLEDGMENTS

We thank M. Igra, M. Bellew, and D. May for assistance on software msInspect; M. Brusniak, O. Vitek, and X. Li for assistance on software specArray; R. Fang for testing the program; and A. E. Detter for helpful suggestions and proof reading of the manuscript. This work was funded by National Cancer Institute (NCI) contract 23XS144A. Martin Mcintosh was supported in part by NCI contract P50 CA83636. Hua Tang, Eugene Yi, and Ruedi Aebersold were supported in part with federal funds from National Heart, Lung, and Blood Institute, National Institutes of Health (NIH) contract N01-HV-28179, NCI, NIH contract N01-CO-12400, and grant R21-CA-114852. *Conflict of Interest:* None declared. Funding to pay the Open Access publication charges for this article was provided by NCI contract 23XS144A.

REFERENCES

BELLEW, M., CORAM, M., IGRA, M., FITZGIBBON, M., RANDOLPH, T., WANG, P., ENG, J., LIN, C., GOODLETT, D., FANG, R. *and others* (July 28, 2006). Informatics method for generating peptide arrays from high resolution lc-ms measurements from complex protein mixtures. *Bioinformatics*. 10.1093/bioinformatics/btl379.

BYLUND, D., DANIELSSON, R., MALMQUIST, G. AND MARKIDES, K. E. (2002). Chromatographic alignment by warping and dynamic programming as a pre-processing tool for parafac modelling of liquid chromatography-mass spectrometry data. *Journal of Chromatography. A* **961**, 237–44.

EFRON, B., JOHNSTRON, I., HASTIE, T. AND TIBSHIRANI, R. (2003). Least angle regression. *Annals of Statistics* **32**, 407–99.

FANG, R., ELIAS, D. A., MONROE, M. E., SHEN, Y., MCINTOSH, M., WANG, P., GODDARD, C. D., CALLISTER, S. J., MOORE, R. J., GORBY, Y. A. *and others* (2006). Differential label-free quantitative proteomic analysis of *Shewanella oneidensis* cultured under aerobic and suboxic conditions by accurate mass and time tag approach. *Molecular & Cellular Proteomics* **5**, 714–25.

KATAJAMAA, M. AND OREŠIC, M. (2005). Processing methods for differential analysis of lc/ms profile data. *BMC Bioinformatics* **6**, 179.

LI, X., YI, E., KEMP, C., ZHANG, H. AND AEBERSOLD, R. (2005). A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Molecular & Cellular Proteomics* **4**, 1328–40.

LIEBLER, D. C. (2002). *Introduction to Proteomics*. Totowa, NJ: Humana Press.

LISTGARTEN, J., NEAL, R. M., ROWEIS, S. T. AND EMILI, A. (2005). Multiple alignment of continuous time series. In: Saul, L. K. *et al.* (editors), *Advances in Neural Information Processing Systems 17 (aka NIPS*2004)*. Cambridge, MA: MIT Press.

MANN, M. AND AEBERSOLD, R. (2003). Mass spectrometry-based proteomics. *Nature* **422**, 198–207.

NIELSEN, N. P., CARSTENSEN, J. M. AND SMEDSGAARD, J. (1998). Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping. *Journal of Chromatography A* **805**, 17–35.

PRAKASH, A., MALLICK, P., WHITEAKER, J., ZHANG, H., PAULOVICH, A., FLORY, M., LEE, H., AEBERSOLD, R. AND SCHWIKOWSKI, B. (2006). Signal maps for mass spectrometry-based comparative proteomics. *Molecular & Cellular Proteomics* **5**, 423–32.

RADULOVIC, D., JELVEH, S., RYU, S., HAMILTON, T. G., FOSS, E., MAO, Y. AND EMILI, A. (2004). Informatics platform for global proteomic profiling and biomarker discovery using liquid-chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics* **3**, 984–97.

RANDOLPH, T. W. AND YASUI, Y. (2004). Multiscale processing of mass spectrometry data. *Biometrics* **62**, 589–97.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* **58**, 267–88.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B., SOLTYS, S., SHI, G., KOONG, A. AND LE, Q. (2004). Sample classification from protein mass spectrometry by peak probability contrasts. *Bioinformatics* **20**, 3034–44.

WANG, W., ZHOU, H., LIN, H., ROY, S., SHALER, T. A., HILL, L. R., NORTON, S., KUMAR, P., ANDERLE, M. AND BECKER, C. (2003). Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Analytical Chemistry* **75**, 4818–26.

ZHANG, H., YI, E. C., LI, X.-J., MALLICK, P., KELLY-SPRATT, K. S., MASSELON, C. D., CAMP, II, D. G., SMITH, R. D., KEMP, C. J. AND AEBERSOLD, R. (2005). High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry. *Molecular & Cellular Proteomics* **4**, 144–55.

ZHANG, H., LI, X., MARTIN, D. AND AEBERSOLD, R. (2003). Identification and quantification of n-linked glycoproteins using hydrazide chemistry, stable isotope labeling and mass spectrometry. *Nature Biotechnology* **21**, 660–6.

ZOU, H. AND HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B* **67**, 301–20.

[*Received December 20, 2005; first revision May 26, 2006; second revision July 11, 2006; accepted for publication July 13, 2006*]