BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 17 2014, pages 2464–2470
doi:10.1093/bioinformatics/btu299

Data and text mining

Advance Access publication May 9, 2014

PeakLink: a new peptide peak linking method in LC-MS/MS using wavelet and SVM

Mehrab Ghanat Bari*, Xuepo Ma and Jianqiu Zhang

Department of Electrical and Computer Engineering, The University of Texas at San Antonio, San Antonio, TX 78246, USA

Associate Editor: Igor Jurisica

ABSTRACT

Motivation: In liquid chromatography–mass spectrometry/tandem mass spectrometry (LC-MS/MS), it is necessary to link tandem MS-identified peptide peaks so that protein expression changes between the two runs can be tracked. However, only a small number of peptides can be identified and linked by tandem MS in two runs, and it becomes necessary to link peptide peaks with tandem identification in one run to their corresponding ones in another run without identification. In the past, peptide peaks are linked based on similarities in retention time (*rt*), mass or peak shape after *rt* alignment, which corrects mean *rt* shifts between runs. However, the accuracy in linking is still limited especially for complex samples collected from different conditions. Consequently, large-scale proteomics studies that require comparison of protein expression profiles of hundreds of patients can not be carried out effectively.

Method: In this article, we consider the problem of linking peptides from a pair of LC-MS/MS runs and propose a new method, PeakLink (*PL*), which uses information in both the time and frequency domain as inputs to a non-linear support vector machine (SVM) classifier. The *PL* algorithm first uses a threshold on an *rt* likelihood ratio score to remove candidate corresponding peaks with excessively large elution time shifts, then *PL* calculates the correlation between a pair of candidate peaks after reducing noise through wavelet transformation. After converting *rt* and peak shape correlation to statistical scores, an SVM classifier is trained and applied for differentiating corresponding and non-corresponding peptide peaks.

Results: *PL* is tested in multiple challenging cases, in which LC-MS/MS samples are collected from different disease states, different instruments and different laboratories. Testing results show significant improvement in linking accuracy compared with other algorithms.

Availability and implementation: M files for the PL alignment method are available at http://compgenomics.utsa.edu/zgroup/PeakLink

Contact: Michelle.Zhang@utsa.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 22, 2013; revised on April 3, 2014; accepted on April 23, 2014

1 INTRODUCTION

Liquid chromatography–mass spectrometry/tandem mass spectrometry (LC-MS/MS) is a well-known tool for analyzing complex protein samples both quantitatively and qualitatively (Silva *et al.*, 2006). An LC-MS/MS experiment begins with the enzymatic digestion of proteins into peptides, which are then injected to and separated in an elution column in the LC step. The retention time (*rt*) it takes for a peptide to pass through the LC column is determined by the physiochemical properties of the peptide. Subsequently, peptides are ionized and separated by their mass over charge (*m*/*z*) ratios. At a given charge state, a peptide registers its ^{12}C monoisotopic 2D peak at (*rt*, *m*/*z*), which is considered as the representing feature of the peptide (Lange *et al.*, 2008). In addition, peptide peak shape can also be treated as a feature (Cui *et al.*, 2011). Note that in this article, properties of peptide peaks are defined as *features*.

After the LC-MS step, tandem MS may perform further fragmentation of some peptides that it picks up. The resulted spectrum can be submitted to search engines such as MASCOT (Eng *et al.*, 1994), SEQUEST (Perkins *et al.*, 1999) or X!tandem (Craig and Beavis, 2004) for peptide identification. Compared with older peptide identification technologies such as the accurate mass and time tag (Pasa-Toli *et al.*, 2004) approach, tandem MS search provides more confident peptide sequence and *rt* information, based on which, peptide LC-MS peaks can be located for confident quantification. However, only a small portion of peptide precursors are sampled in tandem MS, and only a small number of commonly identified proteins/peptides can be compared across multiple LC-MS/MS runs (Zhang *et al.*, 2009). This creates the need to link peptide peaks in runs with identification to their corresponding peaks in runs without tandem MS identification. In this article, *feature linking* or *peak linking*, are used interchangeably to describe the process of linking peptide peaks based on peak features.

There are many algorithms, such as msInspect (Bellew *et al.*, 2006), MZmine 2 (Pluskal *et al.*, 2010), MultiAlign (LaMarche *et al.*, 2013), SuperHirn (Mueller *et al.*, 2008), *ChromAlign* (Neilson *et al.*, 2011), OpenMS (Sturm *et al.*, 2008), SIMA (Voss *et al.*, 2010) and most recently SMFM/SMFM-g (Lin *et al.*, 2013), that consider the problem of linking corresponding peptide peaks in LC-MS/MS runs. Most of these algorithms are warping-based correspondence algorithm (WCA), in which, the elution time shifts between two LC-MS/MS runs are corrected by finding a warping (alignment) function, and peaks are linked based on their closeness in (*rt*, *m*/*z*).

However, few of these algorithms use MS/MS information for generating warping functions (Smith *et al.* 2013) directly, which could provide good performance with little computing resource. For example, Polynomial-4 (P-4) is a simple WCA that uses commonly identified peptides as anchors, whose elution times

*To whom correspondence should be addressed.

© The Author 2014. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

in two LC-MS/MS runs are fitted to generate a simple P-4 warping function. P-4 has not been published as an algorithm separately. It has been described in (Lin *et al.*, 2013), which shows that P-4 performs similarly to SMFM/SMFM-g and MZMine 2 and significantly superior to msInspect and MultiAlign in four testing cases. In our own testing, P-4 is shown to have a performance that fluctuates around the reported accuracy of SMFM/SMFM-g in 70 rounds of testing. This validates the results reported in (Lin *et al.*, 2013) and confirms that P-4 is an easily implementable algorithm with comparable performance to the best of WCAs in the literature.

Besides WCAs, a unique statistical corresponding feature identification algorithm (SCFIA) is proposed in (Cui *et al.*, 2011). SCFIA differs from WCA algorithms in two aspects. Firstly, after correcting the mean elution time shifts using a warping function, SCFIA uses peak shape correlation between candidate corresponding peaks as an additional feature for matching; secondly, it treats peak shape correlation and elution time shifts between candidate corresponding peaks as random variables, and statistical models are used for making maximum likelihood decisions. The statistical models are trained based on common tandem MS-identified peptides between two LC-MS/MS runs. We categorize SCFIA as a WCA with matching.

Although WCAs can perform well for technical replicates and LC-MS/MS data collected from simple organisms, their performance deteriorates significantly when samples are collected on complex organisms or from different laboratories, as shown in our own tests and in (Lin *et al.*, 2013). The linking accuracy could be as low as 70%. This seriously limits the capability of comparing protein expression profiles across multiple samples required in biological and clinical applications.

Although SCFIA improves the performance significantly over P-4 [which is called Gwarping in (Cui *et al.*, 2011)], and consequently, other WCAs, there is still significant room for improvement beyond SCFIA. SCFIA calculates peak shape correlation directly based on peak shapes, which are affected by both signal and noise. To address this issue, we propose to use wavelet decomposition to reduce noise and calculate correlation scores after de-noising. Also, SCFIA assumes that *rt* shifts and peak shape correlations are independent statistically, which is not a verified assumption. To address these issues, we propose to treat peak shape correlations and *rt* shifts as correlated features and use the support vector machine (SVM) to perform classification. Our proposed algorithm is called PeakLink (*PL*).

PL first uses a non-linear warping function generated based on tandem MS information to correct mean *rt* shifts between LC-MS/MS runs. Then, it uses a threshold on an *rt* shift likelihood ratio score to eliminate candidate peak pairs with big time shifts. In the next step, wavelet transform is applied to remaining candidate corresponding peaks, and the correlation between low-frequency coefficients of wavelet transformation is calculated. *PL* processing could terminate at this point by choosing the candidate pair with the highest correlation. Alternatively, *PL* continues the processing by converting *rt* shifts and peak shape correlations to probability scores and setting these scores as inputs to an SVM classifier (Meyer *et al.*, 2003), so that corresponding features can be identified with high accuracy.

Our requirement on tandem MS identification is not a limitation because tandem MS is widely available. Even without it, as long as reliable peptide identification is available, *PL* can be applied.

To evaluate the performance of *PL*, two groups of publicly available datasets are selected. Group 1 contains super-SILAC datasets (Geiger *et al.*, 2010) collected from breast cancer and normal tissues. Group 2 contains two label-free datasets of yeast proteins from two different laboratories (Nagaraj *et al.*, 2012; Swaney *et al.*, 2008). In Group 1, the data were collected from different tissues with large *rt* shift variations. If P-4 is used, the alignment accuracy is only 72.88% on average. In contrast, SCFIA and *PL* can align the dataset with 82.81 and 90.05% accuracy on average, respectively. In Group 2, *PL* achieves 92% accuracy on average, whereas P-4 and SMFM/SMFM-g are reported to have 79 and 82% accuracy, respectively, in (Lin *et al.*, 2013). These results show that *PL* consistently improves peak linking accuracy in the most challenging cases and could have wide applications in biological and clinical research based on LC-MS/MS.

Note that we consider the problem of peak linking between two LC-MS/MS runs in this article. Although it is possible to extend the algorithm to multiple runs, special considerations are needed for selecting reference runs as pointed out in (Smith *et al.*, 2013).

2 METHODS

Suppose each LC-MS/MS run registers an LC-MS map (M) with a list of tandem MS-identified peptides. Given $P^i_{M_1}$, a peptide peak located in M_1 through tandem MS identification, our goal is to find its corresponding peaks, $P^i_{M_2}, P^i_{M_3}, \ldots$ in other LC-MS maps $[M_2, M_3, \ldots]$, in which the peptide has not been identified. The proposed algorithm will first perform LC-MS data preprocessing, which includes tandem MS identification, training data selection and mean *rt* shift correction. Subsequently, *PL* Level 1 (PL_1) and *PL* Level 2 (PL_2) processing are applied.

Note that in some cases, when *rt* variation is small and the sample has fewer peptides, there is little ambiguity in linking peptides based on *rt* and peak shape correlation. The additional SVM step in PL_2 will not improve the performance. In such cases, it is more advantageous to terminate the algorithm after PL_1 and select the candidate pair with the highest peak shape correlation. We propose to estimate the performance of PL_1 and PL_2 and select the appropriate level of processing online. The flow diagram of *PL* is shown in Figure 1, which includes nine steps. Steps 1–3 are preprocessing steps, Steps 4, 5 and 6 belong to PL_1, and the rest of the steps belong to PL_2.

2.1 Preprocessing

Given a pair of LC-MS maps, the goal of preprocessing is to obtain a list of peptides that are simultaneously identified by tandem MS in both runs, based on which we can obtain a set of training and testing peptide peak pairs. Subsequently, based on the training set, a warping function will be estimated to correct mean time shifts. The testing set will be used for the online evaluation of linking accuracy to determine if PL_1 or PL_2 should be used. The preprocessing steps are described as the following:

1. **Tandem MS identification and LC-MS feature extraction.** Many tandem MS search engines exist nowadays. We select the widely available *Andromeda* search engine embedded in MaxQuant (Cox *et al.*, 2011) version 1.4.1.2. MaxQuant uses raw data as input for tandem MS identification, and it provides a list of identified peptides. In MaxQuant, we set the false discovery rate to 0.01,

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

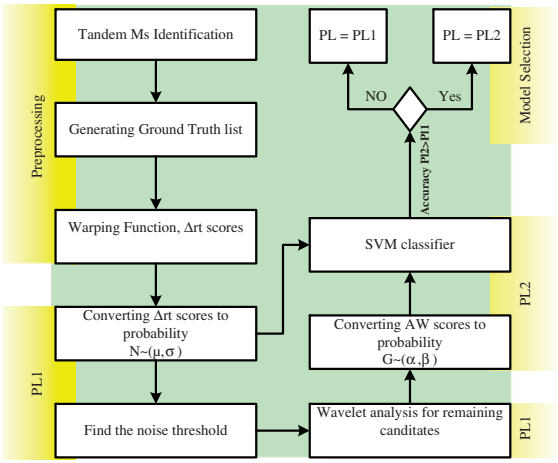

Fig. 1. Overview of *PL* workflow

minimum length of peptides to 6 and the maximum number of missed cleavage to 2. The International Protein Index for Human (ipi.HUMAN.v3.83.fasta) and yeast (yeast_orf_trans_all_05-Jan-2010.fasta) are selected as sequence databases.

To locate the LC peaks registered by tandem MS-identified peptides in M_1 and M_2, we follow the same procedure as that shown in (Cui *et al.*, 2011). After feature extraction, peptides are annotated by their peak shape, *rt* and *m*/*z* information in the LC-MS runs that they are identified by tandem MS.

2. **Ground truth generation.** Through reliable tandem MS identification, we will have a list of commonly identified peptides between M_1 and M_2, which can be used as the ground truth list for corresponding feature identification. Note that this ground truth list contains some false positive findings because of imperfections in tandem MS identification.

   We further divide this ground truth list into a training and a testing set. The training set is used to build statistical models and generate a warping function, and the testing set is used for PL_1/PL_2 selection and performance analysis.

   The training set is further divided into corresponding and non-corresponding subsets. Suppose P_{M_1} is an existing peptide peak in M_1, and there are n potential candidates $\{P^1_{M_2}, P^2_{M_2}, ..., P^i_{M_2}, ..., P^n_{M_2}\}$ in M_2, where we know $P^i_{M_2}$ is its corresponding peak through tandem MS. Consequently, the pair $(P_{M_1}, P^i_{M_2})$, is considered as corresponding, and $\{(P_{M_1}, P^1_{M_2}), (P^1_{M_1}, P^2_{M_2}), ..., (P^1_{M_1}, P^n_{M_2})\}$ are considered as non-corresponding.

   For all peptides in the testing set with peaks in M_1, *PL* will predict their matching peaks in M_2. The linking accuracy is estimated as the ratio between correctly identified corresponding pairs over the total number of commonly identified peptides in the testing set. The accuracy is estimated for both PL_1 and PL_2, and the one with better accuracy will be selected.

3. **Applying a warping function to correct the mean *rt* shift**. Given identified peptide peaks in M_1, their corresponding peaks in M_2 will be shifted in *rt* because of experimental variations. The main purpose of this step is to correct the mean time shift. We generate the warping function by fitting a polynomial of degree 4 to the pair, $\{(t^i_{M_1}, t^i_{M_2})\}_{i=1:n}$, where n is the total number of corresponding peaks in the training set, $t^i_{M_1}$ is the *rt* of the *i-th* training peptide in M_1 and $t^i_{M_2}$ is its corresponding *rt* in M_2. The five coefficients of the warping function, $f(x)$, are calculated using the Matlab function *polyfit(.)*. Note that this warping function is equivalent to the P-4 function in (Lin *et al.*, 2013) and Gwarping function in (Cui *et al.*, 2011). Using $f(x)$, the *rt* shift between a pair of candidate peaks is defined as

$$\Delta rt_i = f(t^1_{M_1}) - t^i_{M_2}, i = 1:n \quad (1)$$

   Here, $t^1_{M_1}$ is the *rt* of a peptide in M_1, and $t^i_{M_2}$ is the *rt* of the *i-th* candidate corresponding peak of the peptide in M_2. It is expected that true corresponding features should have smaller Δrt scores.

2.2 PL_1 processing

4. **Fit a normal distribution on Δrt.** After preprocessing, Δrt has been calculated for all peptides in the training set. We find that Δrt follows a normal distribution for both corresponding and non-corresponding features, as observed in (Cui *et al.*, 2011). Consequently, we can estimate the mean and variance of elution time shifts of corresponding features, (μ_c, σ_c), and (μ_{nc}, σ_{nc}) for non-corresponding features based on the training set. Once these parameters are estimated, then, given the Δrt of a candidate corresponding pair, we can estimate its likelihood probabilities of being corresponding and non-corresponding as $P(\Delta rt|\mu_c, \sigma_c)$ and $P(\Delta rt|\mu_{nc}, \sigma_{nc})$, respectively. We define the following likelihood ratio as $\Delta rtLR$

$$\Delta rtLR = \frac{P(\Delta rt|\mu_c, \sigma_c)}{P(\Delta rt|\mu_{nc}, \sigma_{nc})} \quad (2)$$

   Typically, $\Delta rtLR$ scores should be >1 for corresponding features.

5. **Outlier rejection based on $\Delta rtLR$ scores.** Generally, $\Delta rtLR$ is much larger for corresponding than for non-corresponding peaks. However, there always exist outliers with small $\Delta rtLR$s, which could be attributed to false-positive findings in tandem MS identification. We set the minimum $\Delta rtLR$ score of the top (98%) corresponding features as the threshold on $\Delta rtLR$ and reject the rest as outliers. Note that this process does not require user intervention.

   After rejecting candidate pairs with small $\Delta rtLR$ scores in PL_1, we want to pick corresponding peak pairs from the remaining candidates. In the past, SCFIA (Cui *et al.*, 2011) has successfully improved linking accuracy by using peak shape correlation, which is calculated based on noisy LC-MS peak shapes. To further improve performance, we propose to use wavelet decomposition to reduce the influence of noise.

6. **Wavelet decomposition.** Wavelet transform localizes a signal both in position and scale more quickly than Fourier transform (Neelamani *et al.*, 2004; Vonesch *et al.*, 2007). All wavelet transforms are implemented as sequences of decompositions. Daubechies family wavelets (db1–db20) are mostly used in signal processing, where the number after 'db' shows the order. Multilevel 1D decomposition of a Haar wavelet, which is considered as the first member of the Daubechies family (db1), is shown in Supplementary Figure S1.

The necessary number of decomposition levels depends on the kind of wavelet transform and the property of signals. In general, for a signal of length 2^K, the number of levels will be in the range of 1–K. Suppose $M \in [1, K]$ decompositions were performed, then $[L_M, H_M, H_{M-1}, ..., H_2, H_1]$ will be the wavelet representation of the signal S. In this study, we use Daubechies orthogonal wavelet (db12), and the decomposition level is set to 6.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

By applying wavelet transform to remaining candidates, LC peaks after PL_1, each peak will be represented by wavelet coefficients $[L_6, H_6, H_5, ..., H_2, H_1]$. We focus on L_6, the low-frequency subsignal that holds the main information of the peak. High-frequency coefficients, $H_6..., H_2, H_1$, are discarded.

Before wavelet decomposition, the LC-MS peaks of candidate pairs need resampling because LC-MS/MS devices may not record enough samples for wavelet decomposition. We need 2^6 samples to perform db12 Level 6 decomposition. The Matlab function, *interp1(.)*, is used for resampling, and after which two peaks from the candidate pair are aligned by shifting the shorter peak along the longer one until the highest correlation is reached. Then, the tails of the longer peak are trimmed off so that both peaks have equal lengths. Now wavelet decomposition can be applied, and we define AW as the absolute value of peak shape correlation between seven low-frequency wavelet coefficients of L_6:

$$AW_i = |corr(L_6^{M_1}(1:7), L_6^{M_2}(1:7))|_i, i = 1:n \quad (3)$$

where L_6^{M1} is the wavelet coefficient of an LC peak in M_1, $L_6^{M_2}$ is the wavelet coefficient of the *i-th* candidate in M_2 and n is the number of candidate pairs. A larger AW indicates a larger correlation, and PL_1 chooses the peak candidate pair with the largest AW score as the corresponding pair after applying the $\Delta rtLR$ score threshold. Compared with SCFIA, PL_1 further improves the performance by using wavelet transformation to reduce the influence of noise.

We select the number of low-frequency coefficients by examining the separation of peak shape correlation histograms of corresponding and non-corresponding peaks. The separation is examined using the receiver-operating characteristic (ROC) curves. Our test shows that the first seven coefficients of L_6 give the best separation. In Supplementary Figure S2, we compare the ROC curves before and after wavelet decomposition. To obtain the ROC curves, we apply a varying threshold on peak shape correlations (AWs). At a given threshold *th*, suppose the total number of corresponding peak pairs with peak shape correlation above the threshold is TP in the training set, the total number non-corresponding peptide pairs with peak shape correlation above the threshold is FP in the training set, the total number corresponding peptide pairs is C and the total number of non-corresponding pairs is NC. Then the false-positive rate is calculated as FP/NC, and the true-positive rate is calculated as TP/C. From Supplementary Figure S2, we can see that the separation increases significantly after wavelet decomposition.

2.3 PL_2: SVM classification

In PL_2 processing, our goal is to classify candidate pairs of peptide peaks from a pair of LC-MS maps as corresponding or non-corresponding. SCFIA assumes that the *rt* shift and peak shape correlation of a candidate pair are statistically independent. Although such an assumption allows the derivation of a simple algorithm, it does not hold in general. To address this issue, we propose to treat peak shape correlations and *rt* shifts as correlated features, and SVM is used to perform classification.

7. **Fit a gamma distribution on peak shape correlation.** Although we can directly combine AW and Δrt for classification, their scale is different, and this causes numerical problems in classification, which leads to poorer performance. Consequently, we need to convert AW to probability scores as in the case of Δrt. AW scores are expected to be close to 1 for corresponding peaks and 0 for non-corresponding ones. In practice, AW has a gamma-like distribution, $(x:\alpha,\beta)$, where α and β are the model parameters and need to be estimated like in Step 1.

   We estimate two sets of parameters (α_c, β_c) for the corresponding training set and $(\alpha_{nc}, \beta_{nc})$ for the non-corresponding set. Then, AW scores are converted to probability scores $P(AW|\alpha_c, \beta_c)$ and $P(AW|\alpha_{nc}, \beta_{nc})$. These probability scores are short noted as P^c_{AW} and P^{nc}_{AW}. Given a candidate peak pair, we have two probability scores based on Δrt and two probability scores based on AW.

8. **SVM classification**. After training the statistical models for Δrt and AW, we can calculate the likelihood ratios $\{P^c_{\Delta rt}/P^{nc}_{\Delta rt}, P^c_{AW}/P^{nc}_{AW}\}$ for any candidate feature pair. Based on that, we want to determine whether the pairs are corresponding. We propose to use SVM, a powerful classifier suitable for complex problems in bioinformatics (Zhenqiu *et al*., 2010). By using linear or non-linear kernel functions, SVM can map the input data to high-dimensional space, in which it can specify accurate decision boundaries to classify inputs more accurately.

   Suppose a candidate pair is formed by the *i-th* true peak in M_1 and its *j-th* candidate corresponding peak in M_2, then a pair of features $\{P^c_{\Delta rt}/P^{nc}_{\Delta rt}, P^c_{AW}/P^{nc}_{AW}\}_{i,j}$ can be used as inputs to the SVM classifier. We calculate the two features for all peptides in the training set and train the SVM classifier. We set 1 as the class label for corresponding pairs and 0 for non-corresponding ones. The polynomial kernel function is selected, and the box constraint is set to 3. The decision boundary calculated by the SVM classifier based on a label-free yeast dataset is shown in Supplementary Figure S3.

9. **Choose PL_1 or PL_2 for alignment.** We pick the corresponding peaks based on maximum AW scores in PL_1, and in PL_2, we combine $\Delta rtLR$ and $AWLR = P^c_{AW}/P^{nc}_{AW}$ scores and use SVM to perform a classification. Although PL_2 is expected to perform better than PL_1, there are cases in which there is no ambiguity by linking corresponding peaks according to AW after applying the threshold on $\Delta rtLR$, and PL_2 does not offer performance improvement. PL_2 could offer worse performance than PL_1, because of artifacts in classification. Consequently, we propose to evaluate the performance online to choose the best classifier. The performance is evaluated based on the testing set from the ground truth list, which is available through tandem MS identification in most LC-MS/MS runs. Then based on the evaluated linking accuracy, either PL_1 or PL_2 will be applied for linking all target peptides.

3 RESULTS

3.1 Evaluation datasets and methods

To evaluate our method, two groups of datasets are selected. Group 1 data are gathered from breast cancer research using super-SILAC labeling (Geiger *et al*., 2010). The dataset was originally uploaded to the Tranche database, and it has been moved to ftp://MSV000074502:a@massive.ucsd.edu now. We downloaded the dataset, which is heavy labeled with a super-SILAC mix and mixed with the lysate of mammary carcinoma tissue from an individual with grade II lobular carcinoma. This sample is short noted as the (**Tumor**) sample. We also downloaded the dataset collected from lobular (**Lobular**) and ductal (**Ductal**) breast tumors, as well as the normal (**Normal**) tissue surrounding the ductal carcinoma using the super-SILAC mix as an internal standard for our investigations. These represent a set of typical samples collected in cancer studies. Please refer to the original paper for detailed descriptions of these datasets. In all of these datasets, six fractions are collected, and within each fraction, three technical replicates are collected.

Group 2 contains two publicly available LC-MS/MS datasets, referred to as **Coon2.F4** and **Mann.1** in (Lin *et al*., 2013), which are collected from yeast cells, which are provided by Coon's and Mann's laboratories, respectively. Yeast samples are generally

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

less complex than human samples, and alignment is not an issue by using simple algorithms. However, when samples are collected from different laboratories, the problem becomes more challenging. More details about data acquisition for these datasets can be found in (Nagaraj *et al.*, 2012; Swaney *et al.*, 2008).

Sample composition, LC-MS device and experimental conditions could all affect linking accuracy, and to evaluate the performance of *PL* in various challenging scenarios, we organize our tests in eight cases, as listed in Table 1. In Cases 1 and 2, Fractions 1, 2 and 3 of the Ductal and Lobular breast cancer data are linked with the Normal data. In Cases 3, 4 and 5, the Tumor data are linked with the Lobular, Ductal and the Normal data in the first fraction. These cases represent typical scenarios in clinical and biological studies. In Case 6, two Ductal datasets from different fractions (1 and 3) are selected. Peptide composition from different fractions is expected to vary considerably. In Cases 1–6, data are collected by the same LC-MS device. In Case 7, the Normal and Lobular samples are collected using different devices. Finally, in Case 8, two label-free yeast datasets collected from different laboratories are linked. The exact name of these datasets can be found in the Supplementary Table S1. Note that our test covers three types of MS instrument. In Case 7, the data are collected from Orbi6 and Velos5. In Cases 1–6, the data are collected on Orbi6. Mann's yeast dataset is collected on a Q Exactive mass spectrometer, and Coon's dataset is collected on Orbitrap.

In this article, we selected P-4 and SCFIA as benchmark algorithms for comparison because P-4 is an easily implementable algorithm with comparable performance to the best of WCAs in the literature, and SCFIA (Cui *et al.*, 2011) has superior performance (94.1%) than OpenMS1.7 (80%) and P-4 (termed Gwarping), as shown in (Cui *et al.*, 2011).

To evaluate different algorithms, we calculate the accuracy based on the testing datasets in the ground truth list. The percentage of correctly linked corresponding features among all testing corresponding peptide pairs is reported as linking accuracy. Note that both SCFIA (Cui *et al.*, 2011) and SMFM/SMFM-g (Lin *et al.*, 2013) have used the same evaluation method.

3.2 Linking accuracy

In Figure 2 and Table 2, we show the accuracy results for all cases with different samples compared with SCFIA and P-4. Figure 2 shows the box plot of accuracy results in Cases 1 and 2. In total, there are nine datasets in Cases 1 and 2. In both cases, PL_1 or PL_2 have the highest accuracy compared with the others. PL_2 has the smallest variance in Case 1 and the highest mean alignment accuracy in Case 2. PL_1 or PL_2 is selected as the final model based on maximum accuracy in each experiment. The results of the rest of the cases are listed in Table 2. In Case 3, PL_2 reports 20% higher accuracy than P-4 when linking the Lobular data with the Tumor data in Fraction 1. PL_2 consistently reports the best performance, except in Case 6, in which the *rt* shifts have minimal variations, and even a simple method like P-4 reports satisfactory performance. These results also confirm the consistent improvement of *PL* over SCFIA by using wavelet decomposition and SVM classification.

These tests also reveal that linking accuracy will be negatively affected when different instrument are used for data collection as in Case 7. Although the samples of Case 7 are the same as those in Case 2, the resulted linking accuracy is lower. However, *PL* is the most useful in Case 7, where it improves the accuracy by 20% over P-4.

The number of training and testing peptide pairs in all of these cases are listed in the Supplementary Table S1.

3.3 The effect of using different number of training peptides

In the following, we report the effect of selecting different number of training peptide pairs on performance in Case 8.

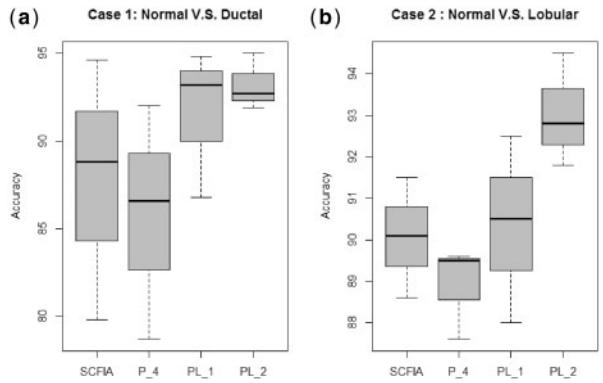

Fig. 2. Cases 1 and 2 contain in total nine samples, and the different methods alignment results are represented by using box plot (**a**) Ductal and Normal (**b**) Lobular and Normal

Table 1. Test cases for performance analysis

Case	Alignment	Fractions
1	Normal versus Ductal	1, 2, 3
2	Normal (Orbi) versus Lobular (Orbi)	1, 2, 3
3	Lobular versus Tumor	1
4	Normal versus Tumor	2
5	Ductal versus Tumor	1
6	Different fractions	1, 3
7	Normal (Orbi.) versus Lobular (Velos5)	1
8	Mann.1 versus Coon2.F4	–

Table 2. Accuracy results of applying different methods on dataset cases

Case Number	SCFIA	*P*-4	PL_1	PL_2
Case 3	0.80	0.661	0.831	**0.895**
Case 4	0.79	0.766	0.828	**0.889**
Case 5	0.841	0.744	0.852	**0.903**
Case 6	0.945	0.949	**0.96**	0.923
Case 7	0.804	0.66	0.793	**0.863**
Case 8	0.85	0.827	0.884	**0.918**

Note: The bold values are the highest accuracy between different methods. Maximum accuracy is 1.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

This case is selected because the same test is reported in (Lin *et al.*, 2013). In this case, we link the *Conn2.F4* and *Mann.1* datasets from different laboratories. Tandem MS search found 374 commonly identified peptides. To evaluate the effect of selecting different training set size on performance, we first randomly pick a percentage between 20 and 80% for choosing the number of training peptides from the ground truth list. For example, if we randomly pick (35%) as the percentage, then the training set will contain $374 \times 35\% \simeq 130$ peptides. The rest of the peptides are used for testing. *PL* is compared with SCFIA and P-4. We repeat this test 70 times, and the resulted accuracy is shown in Figure 3. We can see that the training set size does not affect accuracy significantly, and using ~100 peptides for training is sufficient. On average, PL_1 and PL_2 achieve 88.45 and 91.84% accuracy, respectively, which shows consistent improvement over SCFIA with an average of 85% and over P-4 with an average of 82.7% in accuracy. Note that P-4 is reported to have a 79% accuracy in (Lin *et al.*, 2013), which is lower than that of SMFM/SMFM-g (82%). However, in our test, P-4's performance is on par with that of SMFM/SMFM-g reported in (Lin *et al.*, 2013). Different tests return results with small variations in accuracy, and we should be careful when reporting small differences in performance.

3.4 The effect of peptide peak intensity on accuracy

Because peak shape correlation is affected by signal to noise ratios, linking accuracy could be affected by peptide peak intensity. We investigate this effect by applying a changing log intensity threshold and evaluate the accuracy of testing peptides with peak intensities below the threshold. We plot the accuracy of various algorithms in four cases as a log intensity threshold rises in Figure 4. We can see that PL_2 outperforms other algorithms significantly throughout all intensity ranges. In Cases 1 and 3, PL_2 maintains a good performance even when the intensity is low. In Cases 5 and 7, PL_2's performance increases as the intensity grows. Because *PL* is designed for peptides with single identifications (target peptides), and its testing is conducted on commonly identified peptides in both runs, it is necessary to investigate if there exists any significant difference between testing and targeted peptides in intensity. We have compared the intensity histograms of testing and target peptides in Supplementary Figure S4. There exists minimal difference at the lower end of the histograms before log10 intensities reach 7.5 in most cases. After 7.5, PL_2's performance stabilizes, and the difference in intensity distributions shall not cause significant deviation in accuracy estimation between the targeted and testing peptides.

3.5 Running time comparison

The total amount of time spent on *PL* is small compared with feature extraction in preprocessing, and users do not need to spend time to optimize parameters for *PL* because statistical information on *rt* shifts and peak shape correlations are obtained through the training process. The comparison of running time of *PL* with other algorithms can be found in Supplementary Table S2. Basically, the increased processing time of *PL* over P-4 can be attributed to the required wavelet transformation in PL_1. SVM classification in PL_2 only requires little extra time. Fortunately, the computational complexity of wavelet transformation is linear

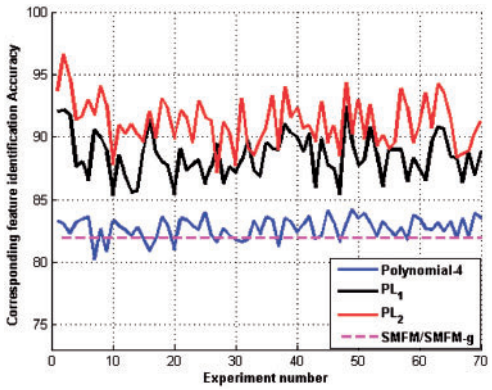

Fig. 3. PL_1 and PL_2 accuracy results in aligning *Conn2.F4 Vs Mann.1* compared with P-4 and SMFM/SMFM-g

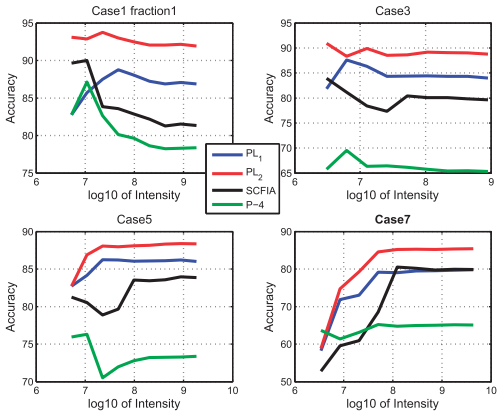

Fig. 4. Accuracy versus log peak intensity

Table 3. Number of linked features by *PL*

Case number	Linked by *PL*	Linked through tandem MS	Link coverage increase (%)
Case1 Fraction1	1854	1258	147
Case 3	3044	1731	176
Case 5	3029	1837	161
Case 7	5109	1072	476

to the number of peptide peaks to be linked. Overall, the total time required by *PL* is small (a few hundred seconds) compared with feature extraction (a few hours in Matlab).

3.6 Linking coverage

PL is designed to significantly increase the number of peptides that can be linked between two LC-MS/MS runs beyond those linked through common tandem MS identifications. We investigate how many more targeted peptides are linked successfully in four of the testing cases by using *PL*. The results are listed in Table 3.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

We can see that *PL* can and significantly increase the number of linked peptides with high accuracy.

4 CONCLUSION

In this article, we propose a new method *PL* for linking corresponding features across two LC-MS/MS maps. The novel aspects of the algorithm include using wavelet transform to reduce noise and using an SVM classifier for linking corresponding features. Compared with other methods in the literature, *PL* offers the highest accuracy in various challenging test cases including complex samples from different tissues, different instruments and different laboratories. Our results are shown to be reproducible across many runs using different training and testing sizes.

With the achieved good performance, the proposed method could have wide applications in biological and clinical studies when protein expression changes across different conditions need examination.

ACKNOWLEDGEMENT

The authors thank the Computational Biology Initiative (UTSA/UTHSCSA) for providing access and training to the analysis software used.

Funding: The National Institute on Minority Health and Health Disparities (G12MD007591) from the National Institutes of Health.

Conflict of Interest: none declared.

REFERENCES

Bellew,M. *et al.* (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics*, **22**, 1902–1909.

Craig,R. and BeavisR.,C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

Cox,J. *et al.* (2011a) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.

Cox,J. *et al.* (2011b) Software lock mass by two-dimensional minimization of peptide mass errors. *J. Am. Soc. Mass Spectrom.*, **22**, 1373–1380.

Cui,J. *et al.* (2011) SCFIA: a statistical corresponding feature identification algorithm for LC/MS. *BMC Bioinformatics*, **12**, 439.

Eng,J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Am. Soc. Mass Spectrom.*, **5**, 976–989.

Geiger,T. *et al.* (2010) Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nat. Methods*, **7**, 383–385.

LaMarche,L.B. *et al.* (2013) MultiAlign: a multiple LC-MS analysis tool for targeted omics analysis. *BMC Bioinformatics*, **14**, 49.

Lange,E. *et al.* (2008) Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements. *BMC Bioinformatics*, **9**, 375.

Lin,H. *et al.* (2013) A combinatorial approach to the peptide feature matching problem for label-free quantification. *Bioinformatics*, **29**, 1768–1775.

Meyer,D. *et al.* (2003) The support vector machine under test. *Neurocomputing*, **55**, 169–186.

Mortensen,P. *et al.* (2010) MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. *J. Proteome Res.*, **7**, 393–403.

Mueller,N.L. *et al.* (2008) SuperHirn- a novel tool for high resolution LC-MS-based peptide/protein profiling. *Proteomics*, **7**, 3470–3480.

Nagaraj,N. *et al.* (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top orbitrap. *Mol. Cell. Proteomics*, **11,** M111.013722.

Neelamani,R.N. *et al.* (2004) ForWaRD: Fourier-wavelet regularized deconvolution for ill-conditioned systems. *IEEE Trans. Signal Process*, **52**, 418–433.

Neilson,K.A. *et al.* (2011) Less label, more free: approaches in label-free quantitative mass spectrometry. *Proteomics*, **11**, 535–553.

Pasa-Toli,L. *et al.* (2004) Proteomic analyses using an accurate mass and time tag strategy. *Biotechniques*, **37**, 621–633.

Perkins,D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Pluskal,T. *et al.* (2010) MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*, **11**, 395.

Silva,S.J. *et al.* (2006) Simultaneous qualitative and quantitative analysis of the *Escherichia coli* proteome. *Mol. Cell. Proteomics*, **5**, 589–607.

Smith,R. *et al.* (2013) LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief. Bioinform.* [Epub ahead of print, doi: 10.1093/bib/bbt080].

Sturm,M. *et al.* (2008) OpenMS - an open-source software framework for mass spectrometry. *BMC Bioinformatics*, **9**, 163.

Swaney,D.L. *et al.* (2008) Decision tree-driven tandem mass spectrometry for shotgun proteomics. *Nat. Methods*, **5**, 959–964.

Vonesch,C. *et al.* (2007) Generalized Daubechies wavelet families. *IEEE Trans., Signal Process.*, **55**, 4415–4429.

Voss,B. *et al.* (2010) SIMA: simultaneous multiple alignment of LC/MS peak lists. *Bioinformatics*, **27**, 987–993.

Zhang,J. *et al.* (2009) Review of peak detection algorithms in liquid-chromatography-mass spectrometry. *Curr. Genomics*, **10**, 388–401.

Zhenqiu,L. *et al.* (2010) Sparse support vector machines with L-p penalty for biomarker identification. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **7**, 100–107.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016