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 cannot 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 \(^{12}\text{C}\) 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 XQtandem (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 mInspect (Bellew et al., 2006), MZmine 2 (Pluskal et al., 2010), MultiAlign (LaMarche et al., 2013), SuperHirn (Mueller et al., 2008), ChromAlign (Nelson 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
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 mInspect 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 and could have wide applications in biological and clinical research based on LC-MS/MS.

To evaluate the performance of PeakLink, 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 \( P_{M_i} \), a peptide located in \( M_i \) through tandem MS identification, our goal is to find its corresponding peaks, \( P_{M_i}'P_{M_i} \), \( \ldots \) 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 process 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,
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 mz information in the LC-MS runs that 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 $\{p_{M_1}^1, p_{M_1}^2, ..., p_{M_1}^n\}$ in $M_2$, where we know $P_{M_1}$ is its corresponding peak through tandem MS. Consequently, the pair $(P_{M_1}, p_{M_2}^i)$, is considered as corresponding, and $(P_{M_1}, p_{M_2}^j)$, $(P_{M_1}, p_{M_2}^k)$, ..., $(P_{M_1}, p_{M_2}^n)$ 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, $(\{r_{M_1}, r_{M_2}\})_{i=1:n}$, where $n$ is the total number of corresponding peaks in the training set, $r_{M_1}$ is the $rt$ of the $i$-th training peptide in $M_1$ and $r_{M_2}$ 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

$$\Delta r_{ti} = f(r_{M_1}^i) - r_{M_2}^i, i = 1 : n$$  \hspace{1cm} (1)

Here, $r_{M_1}^i$ is the $rt$ of a peptide in $M_1$, and $r_{M_2}^i$ 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 $\Delta rt$ scores.

2.2 **$PL_1$ processing.**

4. **Fit a normal distribution on $\Delta rt$.** After preprocessing, $\Delta rt$ has been calculated for all peptides in the training set. We find that $\Delta rt$ 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, $(\mu_+, \sigma_+)$, and $(\mu_-, \sigma_-)$ for non-corresponding features based on the training set. Once these parameters are estimated, then, given the $\Delta rt$ of a candidate corresponding pair, we can estimate its likelihood probabilities of being corresponding and non-corresponding as $P(\Delta rt|\mu_+, \sigma_+)$ and $P(\Delta rt|\mu_-, \sigma_-)$, respectively. We define the following likelihood ratio as $\Delta rt \text{LR}$

$$\Delta rt \text{LR} = \frac{P(\Delta rt|\mu_+, \sigma_+)}{P(\Delta rt|\mu_-, \sigma_-)}$$  \hspace{1cm} (2)

Typically, $\Delta rt \text{LR}$ scores should be $>1$ for corresponding features.

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

After rejecting candidate pairs with small $\Delta rt \text{LR}$ 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.
By applying wavelet transform to remaining candidates, LC peaks after \(PL_1\), each peak will be represented by wavelet coefficients \([L_0, H_0, H_1, \ldots, H_{M_1}]\). We focus on \(L_0\), the low-frequency subsignal that holds the main information of the peak. High-frequency coefficients, \(H_0, \ldots, H_2, H_3\), 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 \(2^p\) samples to perform db12 Level 6 decomposition. The Matlab function, `interpl', 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 \(L_0\):

\[
AW_i = |\text{corr}(L_0^M (1 : 7), L_0^M (1 : 7))|, \quad i = 1 : n
\]

where \(L_0^M\) is the wavelet coefficient of an LC peak in \(M_1\), \(L_i^M\) 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 \(\Delta t\) and peak shape correlation of a candidate peak pair. Based on PeakLink, two probability scores are short noted as \(\text{PL}_{1}^\text{M} \) and \(\text{PL}_{2}^\text{M} \). From Supplementary Figure S2, we can see that the separation increases significantly after wavelet decomposition.

The separation is examined using the receiver operating characteristic (ROC) curves. Our test shows that the first seven coefficients of \(L_0\) 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 \(t\), 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 \(\Delta t\) 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 \(\Delta t\) 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 \(\Delta t\) 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 \(\Delta t\). \(AW\) 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 \(\alpha\) and \(\beta\) are the model parameters and need to be estimated like in Step 1.

We estimate two sets of parameters \((\alpha, \beta)\) for the corresponding training set and \((\alpha_n, \beta_n)\) for the non-corresponding set. Then, \(AW\) scores are converted to probability scores \(P(AW|\alpha, \beta)\) and \(P(AW|\alpha_n, \beta_n)\). These probability scores are short noted as \(P_{AW}^\text{C}\) and \(P_{AW}^\text{NC}\). Given a candidate peak pair, we have two probability scores based on \(\Delta t\) and two probability scores based on \(AW\).

### 8. SVM classification

After training the statistical models for \(\Delta t\) and \(AW\), we can calculate the likelihood ratios \(\{P_{AW|\Delta t}, P_{AW|\Delta t}^\text{C}, P_{AW|\Delta t}^\text{NC}, P_{AW|\Delta t}^\text{NC}\}\) 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 \(\{P_{AW|\Delta t}, P_{AW|\Delta t}^\text{C}, P_{AW|\Delta t}^\text{NC}, P_{AW|\Delta t}^\text{NC}\}\) 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 \(\Delta t\) and \(AW\) 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 \(\Delta t\) and \(PL_1\) 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.
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 both SCFIA (Cui et al., 2011) and SMFM/SMFM-g (Lin et al., 2013) have superior 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.

Table 1. Test cases for performance analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Alignment</th>
<th>Fractions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal versus Ductal</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>2</td>
<td>Normal (Orbi) versus Lobular (Orbi)</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>3</td>
<td>Lobular versus Tumor</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Normal versus Tumor</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>Ductal versus Tumor</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Different fractions</td>
<td>1, 3</td>
</tr>
<tr>
<td>7</td>
<td>Normal (Orbi.) versus Lobular (Velos5)</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>Mann.1 versus Coon2.F4</td>
<td>--</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Case Number</th>
<th>SCFIA</th>
<th>P-4</th>
<th>PL$_1$</th>
<th>PL$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 3</td>
<td>0.80</td>
<td>0.661</td>
<td>0.831</td>
<td>0.895</td>
</tr>
<tr>
<td>Case 4</td>
<td>0.79</td>
<td>0.766</td>
<td>0.828</td>
<td>0.889</td>
</tr>
<tr>
<td>Case 5</td>
<td>0.841</td>
<td>0.744</td>
<td>0.852</td>
<td>0.903</td>
</tr>
<tr>
<td>Case 6</td>
<td>0.945</td>
<td>0.949</td>
<td>0.96</td>
<td>0.923</td>
</tr>
<tr>
<td>Case 7</td>
<td>0.804</td>
<td>0.66</td>
<td>0.793</td>
<td>0.863</td>
</tr>
<tr>
<td>Case 8</td>
<td>0.85</td>
<td>0.827</td>
<td>0.884</td>
<td>0.918</td>
</tr>
</tbody>
</table>

Note: The bold values are the highest accuracy between different methods. Maximum accuracy is 1.
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 \( \frac{374 \times 35\%}{100} \approx 130 \) 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 \( \approx 100 \) 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 \( ri \) 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 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.

<table>
<thead>
<tr>
<th>Case number</th>
<th>Linked by PL</th>
<th>Linked through tandem MS</th>
<th>Link coverage increase (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case1 Fraction1</td>
<td>1854</td>
<td>1258</td>
<td>147</td>
</tr>
<tr>
<td>Case 3</td>
<td>3044</td>
<td>1731</td>
<td>776</td>
</tr>
<tr>
<td>Case 5</td>
<td>3029</td>
<td>1837</td>
<td>161</td>
</tr>
<tr>
<td>Case 7</td>
<td>5109</td>
<td>1072</td>
<td>476</td>
</tr>
</tbody>
</table>
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


