ADEPTS: Advanced Peptide De novo Sequencing with a Pair of Tandem Mass Spectra

Lin He
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
l22he@uwaterloo.ca

Bin Ma∗
David R. Cheriton School of Computer Science, University of Waterloo
Waterloo, Ontario, Canada N2L 3G1
binma@uwaterloo.ca

De novo sequencing is an important task in proteomics to identify novel peptide sequences. Traditionally, only one MS/MS spectrum is used for the sequencing of a peptide, however, the use of multiple spectra of the same peptide with different types of fragmentation has the potential to significantly increase the accuracy and practicality of de novo sequencing. Research into the use of multiple spectra is in a nascent stage. We propose a general framework to combine the two different types of MS/MS data. Experiments demonstrate our method significantly improves the de novo sequencing of existing software.

Keywords: peptide de novo sequencing; tandem mass spectrometry; CID; ETD.

1. Introduction

Currently, tandem mass spectrometry (MS/MS) analysis is the standard method for the identification of peptide sequences. In an MS/MS spectrometer a peptide is fragmented into fragment ions, and from these fragments an MS/MS spectrum is produced. The fragmentation is frequently achieved by the collision induced dissociation (CID) method, however, new fragmentation techniques—such as infrared multiphoton dissociation (IRMPD), electron capture dissociation (ECD), electron transfer dissociation (ETD), and higher energy collisionally activated dissociation (HCD)—have been developed in hope of improving the performance. In particular, ETD is getting widely adopted due to the fact that it is better suited for fragmentation of longer peptide.

Different fragmentation methods produce different ion types, and may fragment at slightly different locations of the same peptide sequence. Therefore, two independent measurements of the same peptide using two different fragmentation methods can provide more information of the peptide than

∗To whom correspondence should be addressed.
only one of the measurements. This approach has recently drawn some attention in peptide identification.\textsuperscript{2,3,4}

There are two typical computational approaches for peptide identification from MS/MS: database search, which tries to identify the sequence by finding the best match in a protein database, and \textit{de novo} sequencing, which constructs the best peptide sequence directly from the MS/MS (without using a protein database). \textit{De novo} sequencing is ideal for the discovery of novel peptide but generally requires higher quality data to obtain satisfactory results. If some of the expected fragment ions are not produced by the fragmentation, then the corresponding peaks will be absent from the spectrum, leading to ambiguity for the determination of some local segments of the peptide. The ambiguity can often be removed by searching a protein database in the database search approach, but it remains in the \textit{de novo} sequencing approach and leads to a partially correct solution.

The possibility of using multiple fragmentation methods provides an ideal solution to reduce the aforementioned ambiguity in the \textit{de novo} sequencing. When a fragment ion is absent from one fragmentation method, a different type of fragment ion with the same segment of residues may be present in another fragmentation method. Thus, the ambiguity is removed and the local structure can be determined. In addition, peaks from the multiple spectra can confirm each other, greatly increasing the confidence to distinguish a signal peak from noise. Indeed, in three recent publications\textsuperscript{2,3,4} the use of two spectra with different fragmentation methods significantly improved the \textit{de novo} sequencing accuracy.

In general, calculating a confidence score to match each peptide candidate with its corresponding MS/MS spectrum is the crucial part of the \textit{de novo} sequencing. In this paper, a fragmentation site refers to all types of ions generated from a fragmentation between two adjacent amino acid residues. The peaks matched by a fragmentation site in the spectrum are used to calculate a score for this fragmentation site, and thus, the score of a candidate peptide is a sum of the scores at all the fragmentation sites. The peptide candidate with the highest score will finally be selected as the result of \textit{de novo} sequencing.

Although both Datta \textit{et al.}\textsuperscript{2} and Savitski \textit{et al.}\textsuperscript{3} presented new algorithms for \textit{de novo} sequencing based on two types of spectra, our method, which is also based on two spectra, has the novel capability that majority of existing \textit{de novo} sequencing algorithms\textsuperscript{5,6,7} can be utilized with some minor modifications. Imperative to the use of two spectra, is the modification of the score for a fragmentation site so that the score is now based on two spectra. Hence, one of the major contributions of this work is a more accurate scoring function that we discuss in thorough detail in Section 2.

Savitski \textit{et al.}\textsuperscript{3} used a simple criterion to determine the correct fragmentation sites and used a greedy algorithm to derive the peptide sequence. Datta \textit{et al.}\textsuperscript{2} computed the score of a fragmentation site by using a TAN-structured Bayesian network to combine the events that different fragment ions occur in the spectrum. Both the network structure and the parameters are trained from the training data.
This differs from the Bayesian network used in PepNovo\textsuperscript{5} algorithm where the structure is hand selected by a human expert. The advantage of using a Bayesian network is that the correlation between different fragment ion types is considered in the model. However, in order to use Bayesian network model, Datta \textit{et al.} had to discard majority of the intensity information because the events in a Bayesian network are required to be discrete in order to learn the joint probabilities from the training data.

Bertsch \textit{et al.}\textsuperscript{4} utilized the intensity information in another way. A divide-and-conquer algorithm, which was similar to the one described by Zhang\textsuperscript{8}, was adopted to generate the peptide candidates for each spectrum. Then a theoretical spectrum was generated according to each peptide candidate to score against the experimental spectrum. However, Bertsch \textit{et al.} require an accurate prediction of the theoretical spectrum, which is an exceedingly difficult problem.\textsuperscript{9} The performance of such a scoring scheme highly depends on the accuracy of a predictor.

In this paper, we propose a new scoring method named ADEPTS to do \textit{de novo} sequencing using a pair of CID and ETD spectra. ADEPTS uses a simple model to incorporate the peak intensity information into the score but pays less attention to the correlation between different ion types. We show that this approach increases the ability to distinguish between the true and false fragmentation sites. In addition, we propose a simple framework to adopt a traditional \textit{de novo} sequencing program to do \textit{de novo} sequencing based on two spectra. Experiments show that ADEPTS improves upon the \textit{de novo} sequencing accuracy over other existing software.

2. Method

Our method requires a pair of MS/MS spectra (one with CID and the other ETD) for the \textit{de novo} sequencing of each target peptide. The computational analysis requires the definition of a new scoring function based on a pair of spectra. The general framework of the analysis is the following:

1. use PEAKS\textsuperscript{6} \textit{de novo} sequencing software to find 1,000 peptide candidates from both CID and ETD spectra;
2. use the new scoring function to evaluate each of the 2,000 peptide candidates against the spectrum pair;
3. report the peptide candidate with the highest score.

We note that PEAKS \textit{de novo} sequencing in step (1) can be replaced by any other existing \textit{de novo} sequencing software that can output multiple peptide candidates from one spectrum, such as Lutefisk\textsuperscript{10} and the algorithm in Lu and Chen\textsuperscript{11}. The more important aspect of the analysis is the definition of the new scoring function that is defined as follows:

1. each peak in a spectrum is assigned with a non-negative \textit{significance} value, according to the intensity of the peak \textit{relative to} other surrounding peaks in the spectrum;
(2) the peak significance is converted to a likelihood score for each frequent ion type;
(3) all fragment ion types at the same site from both spectra are calculated; their likelihood scores form a score vector and an SVM is used to convert the score vector to a score for the site;
(4) similar to (3), an SVM score is calculated for each residue in the peptide sequence;
(5) the scores for all fragmentation sites and all residues of a peptide are added up to give the peptide score.

More details for each step are given in the next subsections.

2.1. Peak Significance Value

As in Liu et al., we calculate four features, the rank, the relative intensity, the local rank, and the local relative intensity for each peak. The rank is the number of peaks in the spectrum with intensity higher than or equal to the current peak. The relative intensity is the ratio between the average intensity of the top few peaks in the spectrum and the intensity of the current peak. The local versions of the above two features are defined the same except that only the peaks within ±56 Da from the current peak are counted, rather than the whole spectrum. The peak significance value is defined as the linear combination of the logarithms of these four values. More specifically,

\[
significance = c_1 \cdot \log(\text{rank}) + c_2 \cdot \log(\text{relative intensity}) + c_3 \cdot \log(\text{local rank}) + c_4 \cdot \log(\text{local relative intensity}).
\]

Notice that under this definition a smaller significance value indicates a stronger peak.

The data refinement function in PEAKS 5.2 is used to preprocess the MS/MS spectra as follows: the peaks are centroided, de-isotoped and de-convoluted. We then calculate the significance value of each peak using Eq.(1). The four coefficients, \(c_1\), \(c_2\), \(c_3\), and \(c_4\) are trained with our training data to maximize the area under the ROC curve (see Section 4.2).

2.2. Likelihood Score of Frequent Ion Types

The frequencies of observed peaks for several common ion types in our training data are listed in Table 1. We select the most frequent five ion types in the table for CID spectra, and also, the most frequent five for ETD spectra. As a result, \(y\), \(b\), \(y\)-H\(_2\)O, \(b\)-H\(_2\)O and \(b\)-NH\(_3\) ions from the CID spectrum, and \(c\), \(y\), \(z\)', \(z\)'+1 and \(c\)-NH\(_3\) ions from the ETD spectrum, are considered in the scoring function.

\[\text{Implementation details and justifications can be found in Liu et al.}^{12}\]
Table 1. Frequencies of matches for different fragment ions in the training data

<table>
<thead>
<tr>
<th>Ion</th>
<th>Frequency (%)</th>
<th>Ion</th>
<th>Frequency (%)</th>
<th>Ion</th>
<th>Frequency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>52.4</td>
<td>y-H₂O</td>
<td>31.9</td>
<td>b</td>
<td>31.7</td>
</tr>
<tr>
<td>b</td>
<td>47.3</td>
<td>b-H₂O</td>
<td>31.7</td>
<td>b-NH₃</td>
<td>27.6</td>
</tr>
<tr>
<td>a-NH₃</td>
<td>25.3</td>
<td>a</td>
<td>23.1</td>
<td>a</td>
<td>22.7</td>
</tr>
<tr>
<td>a</td>
<td>23.7</td>
<td>a-NH₃</td>
<td>22.5</td>
<td>b-NH₃</td>
<td>22.5</td>
</tr>
<tr>
<td>y-NH₃</td>
<td>56.6</td>
<td>c</td>
<td>20.9</td>
<td>z</td>
<td>19.5</td>
</tr>
<tr>
<td>c</td>
<td>50.4</td>
<td>y</td>
<td>20.8</td>
<td>z</td>
<td>18.7</td>
</tr>
<tr>
<td>y</td>
<td>42.6</td>
<td>z'</td>
<td>18.5</td>
<td>z'</td>
<td>18.5</td>
</tr>
<tr>
<td>z</td>
<td>32.4</td>
<td>z'+1</td>
<td></td>
<td>z'+1</td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>31.6</td>
<td>c-NH₃</td>
<td></td>
<td>c-NH₃</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 1. (a) The distributions of the peak significance for the true $z'$-ions and random matches. (b) The $z'$-ion likelihood scoring function with respect to the significance value.

For each of the selected ion type $t$ we define a likelihood score function $f_t(\cdot)$. $f_t(x)$ represents the score contribution of the event that a type-$t$ ion matches a peak with a significance value $x$.

The significance values of the matched peaks by type-$t$ ions in the training data are divided into four intervals, each containing the same number of the matched peaks. Then the likelihood score at the centroid of each interval, denoted by $x_o$, is

$$f_t(x_o) = \log \left( \frac{\Pr(\text{significance value falls in the interval} \mid \text{true site})}{\Pr(\text{significance value falls in the interval} \mid \text{random site})} \right).$$  \hspace{1cm} (2)

The likelihood scores for other significance values are computed by linear interpolation. If a theoretical ion does not match any peak in the spectrum, the likelihood score can also be calculated as

$$f_t(null) = \log \left( \frac{\Pr(\text{no peak matching} \mid \text{true site})}{\Pr(\text{no peak matching} \mid \text{random site})} \right).$$  \hspace{1cm} (3)

For example, the distributions of the significance of $z'$-ion and random matches, as well as the likelihood scoring function of the $z'$-ions, are given in Figure 1.
2.3. Score for Each Fragmentation Site

A peptide candidate of length $n$ contains $n - 1$ fragmentation sites. For each site $i$ and ion type $t$, let the significance value of the matched peak be $x_{i,t}$, then the likelihood score of this match is $s_{i,t} = f_t(x_{i,t})$, where $f_t$ is defined in Section 2.2. The score of the fragmentation site $i$, denoted as $s_i$, is defined as the linear combination of likelihood score of all ion types at this site. Hence,

$$ s_i = \sum_{t=1}^{k} c_t \cdot s_{i,t}, $$

where $k$ is the number of selected ion types, and $c_1, \ldots, c_k$ are constant coefficients that are trained with an SVM.

The standard LIBSVM library\textsuperscript{13} is used to train the SVM. The training of the coefficients maximizes the distinction between the fragmentation sites of true peptide sequences and randomly generated peptide sequences using the training data.

2.4. Score for Each Residue

Each residue in the peptide sequence defines two fragmentation sites: the one at left and right side of the residue. In an MS/MS spectrum, there are often strong correlations between the ions from two adjacent fragmentation sites. Figure 2 shows such a spectrum. Thus, a residue score is used to incorporate this correlation into the scoring function. For each ion type, let $x_i$ and $x_{i+1}$ be the significance values of the two peaks at the two sites determined by a residue. Then the significance $x_r$...
for this residue and the ion type is defined as:

$$x_r = \sqrt{\frac{x_i^2 + x_{i+1}^2}{2}}.$$  \hspace{1cm} (5)

When $$x_i + x_{i+1}$$ is fixed, the least $$x_r$$ (smaller is better) is achieved when $$x_i = x_{i+1}$$. This property results in the fact that the events where two adjacent ions have similar significance are scored high, and thus, incorporates the aforementioned correlation between ions.

Using the same procedure as the score for a fragmentation site, the residue significance values for different ions are converted to likelihood scores and then combined into a residue score using SVM.

2.5. **Peptide Score**

Suppose the peptide candidate has length $$n$$, then the scores calculated for the $$n-1$$ fragmentation sites are $$s_1, \ldots, s_{n-1}$$, and the scores calculated for the $$n$$ residues are $$r_1, \ldots, r_n$$. Let the peptide score be defined as:

$$\text{score} = \sum_{i=1}^{n-1} s_i + \lambda \cdot \sum_{i=1}^{n} r_i,$$ \hspace{1cm} (6)

where $$\lambda$$ is a positive coefficient that balances the weights between the fragmentation site scores and the residue scores. In our method, an enumerative search method is performed to determine the optimal $$\lambda$$ by using the training data.

3. **Experimental Results**

ADEPTS was tested with two independent datasets. The first dataset was obtained from LTQ Orbitrap XL ETD (Thermo Fisher Scientific™, Bremen, Germany), and the second dataset was from an Iontrap mass spectrometer with implemented ETD module (Model HCTultra PTM discovery system, Bruker Daltonik GmbH, Bremen, Germany). The second dataset was previously used by Bertsch et al. for their CompNovo software.

3.1. **Experiments on LTQ Orbitrap Dataset**

A mixture of yeast proteins digested with LysC was measured with both CID and ETD fragmentation modes of an LTQ Orbitrap XL (Thermo Fisher Scientific™, Bremen, Germany). The precursor ions were measured with FT mode and the fragment ions were measured with Iontrap.

The database search module of PEAKS 5.2 was used to identify real peptide sequences from Swiss-Prot database using the spectra in the two LC-MS/MS runs. A pair of spectra from the two runs are considered as from the same peptide if: (i) the two peptide sequences identified for both the CID and ETD spectra are the same; (ii) the PEAKS database search confidence score is at least 60% on both
spectra; and (iii) the two spectra have the same charge state and similar retention time (subject to a fluctuation of at most ±10 minutes). The fairly large LC retention time fluctuation (±10 minutes) is chosen to account for the fact that the two spectra are from two separate LC-MS/MS runs, and thus, the same peptide may not be fragmented at the same LC retention time. The other two conditions ensure that the selected pair of spectra are indeed from the same peptide, with high probability.

317 peptide sequences (and their pairs of CID and ETD spectra) were selected from the PEAKS database search results. 148 of them were randomly chosen as training data, and the remaining 169 were used as testing data. No peptide was in both training and testing data. There were 2,291 total residues in the training data and 2,648 residues in the testing data.

After consultation with the authors of CompNovo software, a known issue prevents the processing of the data in this context. Thus, for this dataset we compared our method with PepNovo and PEAKS, which are widely used de novo sequencing software packages that use only one type of fragmentation at a time. PEAKS 5.2 was used to process CID and ETD data in two different runs. PepNovo release 20091029 was only used to process the CID data since it does not have an ETD parameter.

<table>
<thead>
<tr>
<th>PEAKS (CID)</th>
<th>PEAKS (ETD)</th>
<th>PepNovo (CID)</th>
<th>ADEPTS (CID+ETD)</th>
<th>TOTAL (DB Search)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct peptides</td>
<td>12 (7.1%)</td>
<td>10 (5.9%)</td>
<td>8 (4.7%)</td>
<td>32 (18.9%)</td>
</tr>
<tr>
<td>≤ 1 incorrect residue</td>
<td>12 (7.1%)</td>
<td>13 (7.7%)</td>
<td>11 (6.5%)</td>
<td>33 (19.5%)</td>
</tr>
<tr>
<td>≤ 2 incorrect residues</td>
<td>25 (14.8%)</td>
<td>25 (14.8%)</td>
<td>27 (16.0%)</td>
<td>66 (39.1%)</td>
</tr>
<tr>
<td>≤ 3 incorrect residues</td>
<td>28 (16.6%)</td>
<td>34 (20.1%)</td>
<td>33 (19.5%)</td>
<td>74 (43.8%)</td>
</tr>
<tr>
<td>Total correct residues</td>
<td>945 (35.7%)</td>
<td>1,157 (43.7%)</td>
<td>972 (36.7%)</td>
<td>1,580 (59.7%)</td>
</tr>
</tbody>
</table>

The precursor mass error tolerance and fragment error tolerance were 0.1 Da and 0.5 Da, respectively. The difference in the error tolerances is because the precursor was measured by Orbitrap and the fragment was measured by Iontrap. Table 2 illustrates that ADEPTS outperformed both PEAKS and PepNovo.

\[^b\]When a new version CompNovo software is available we will publish an edition to this paper with the updated results.
3.2. Experiments on Iontrap Dataset

The second dataset was obtained from the authors of CompNovo. It contained 156 CID/ETD spectrum-pairs as training data and 2,405 \(^4\) pairs as testing data. The total number of residues in the training and testing peptides were 1,906 and 32,186, respectively. The spectra were collected from an Iontrap mass spectrometer with implemented ETD module (Model HCT ultra PTM discovery system, Bruker Daltonik GmbH, Bremen, Germany). We observed that the fragmentation pattern of this dataset was quite different from the pattern of the first dataset, which is expected since the different types (and manufactures) of the mass spectrometers were used to collect the data.

The same experiment as in Section 3.1 was conducted on this dataset except that the precursor mass error tolerance and fragment error tolerance were set 1.5 Da and 0.4 Da, respectively. The change in the parameters was done to correspond to what had been used in the CompNovo paper. Again, CompNovo failed to run on our machines and thus, the de novo sequencing results published from Bertsch et al. were used in the comparison. The results of the comparison are summarized in Table 3. Not only ADEPTS significantly improved the de novo sequencing accuracy over PEAKS and PepNovo, but it was noticeably better than CompNovo software, which also uses CID/ETD pairs.

<table>
<thead>
<tr>
<th></th>
<th>PEAKS (CID)</th>
<th>PEAKS (ETD)</th>
<th>PepNovo (CID)</th>
<th>CompNovo (CID)</th>
<th>CompNovo (CID+ETD)</th>
<th>ADEPTS (CID+ETD)</th>
<th>TOTAL (DB Search)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct peptides</td>
<td>286 (11.9%)</td>
<td>57 (2.4%)</td>
<td>63 (2.6%)</td>
<td>676 (28.1%)</td>
<td>820 (34.1%)</td>
<td>2,405</td>
<td></td>
</tr>
<tr>
<td>≤ 1 incorrect residue</td>
<td>301 (12.5%)</td>
<td>74 (3.1%)</td>
<td>99 (4.1%)</td>
<td>697 (29.0%)</td>
<td>844 (35.1%)</td>
<td>2,405</td>
<td></td>
</tr>
<tr>
<td>≤ 2 incorrect residues</td>
<td>779 (32.4%)</td>
<td>216 (9.0%)</td>
<td>404 (16.8%)</td>
<td>1,243 (51.7%)</td>
<td>1,350 (56.1%)</td>
<td>2,405</td>
<td></td>
</tr>
<tr>
<td>≤ 3 incorrect residues</td>
<td>971 (40.4%)</td>
<td>362 (15.1%)</td>
<td>664 (27.6%)</td>
<td>1,445 (60.1%)</td>
<td>1,528 (63.5%)</td>
<td>2,405</td>
<td></td>
</tr>
<tr>
<td>Total correct residues</td>
<td>20,069 (62.4%)</td>
<td>13,767 (42.8%)</td>
<td>19,175 (59.6%)</td>
<td>23,721 (73.7%)</td>
<td>24,378 (75.7%)</td>
<td>32,186</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3(a) compares the identification rates of different methods as the function of the number of allowed incorrect residues in each peptide. Figure 3(b) compares the identification rates of different methods as the function of the length of the longest consecutively correct subsequence. Clearly, the two methods (ADEPTS and

\(^4\)There were 2,406 pairs in the testing data of the CompNovo paper. One pair was removed since it is clearly an error: the two spectra had the different precursor m/z value.
CompNovo) that combine CID/ETD spectra pair performed significantly better than the methods that use only one type of fragmentation (PEAKS and PepNovo). In addition the performance of ADEPTS is noticeably better in both figures.

4. Discussion

In this section we justify the choices we made in our method and experiments (Section 4.1 – 4.3), and discuss how the software can be used (Section 4.4 – 4.6).

4.1. The Use of PEAKS Software for Candidate Generation

It is possible that any de novo sequencing tool which outputs multiple peptide candidates may be used in step (1) in the framework in Section 2. PEAKS is used because it is regarded as the most superior general de novo sequencing software with respect to the accuracy and efficiency, and it also supports de novo sequencing for both CID and ETD. In addition, it has the capability of outputting as many as 1,000 candidates for each spectrum; the relatively large number of candidates is important to our framework. However, PEAKS itself cannot perform de novo sequencing on a pair of CID and ETD spectra. As shown in the experimental results, ADEPTS significantly outperformed PEAKS (CID) and PEAKS (ETD). This indicates the importance of the scoring function on a pair of CID and ETD spectra.

In this paper the peptide candidates were generated using CID and ETD separately. There is a possibility that the performance can be further improved if the candidate generation is done based on the combination of CID and ETD spectra as well. We are currently investigating this possibility.

4.2. The Use of Significance Value

In our framework, the peak significance value rather than the peak intensity is calculated and used in the analysis. This is important to the performance of ADEPTS.
To illustrate this, we compare the discriminative power of this significance value and the relative intensity (the ratio between the current peak intensity and the maximum peak intensity in the spectrum) by their ROC curves in Figure 4. To generate the curves, the positive peaks are those matched by the $y$- and $b$-ions (if CID spectrum) or by the $c$- and $z'$-ions (if ETD spectrum) in the testing data; and the negative peaks are those matched by random $m/z$ values. The figure shows that the significance value is a better measure than the relative intensity.

4.3. The Use of SVM

There are correlations between different ion types at the same fragmentation sites, and between two adjacent fragment ions of the same type. To better account for these correlations, the Bayesian network model used by Datta et al. or PepNovo is more appropriate. However, this requires to discretize the peak intensity (or significance value), resulting in a loss of information that is valuable to the accuracy of the de novo software. Instead, an SVM is used to remove partial correlations between different ion types at the same fragmentation site, and the residue score defined in Section 2.4 is used to account for the correlations between the adjacent fragment ions. Determining a model that has the power of a Bayesian network, without the accompanied weakness, is a problem that warrants future investigation.

Both RBF kernel and the linear kernel were tried in the SVM approach. The procedure to optimize the SVM kernel parameters suggested by Hsu et al.\textsuperscript{15} was followed. Both kernels achieved very similar performance, and thus, the linear kernel is used for its simplicity and efficiency.

Instead of using the likelihood scores as the vector for SVM, we also tried to use the peak significance values to form the vector, but the accuracy is remarkably reduced. This also indicates the conversion from significance value to the likelihood...
score is necessary.

4.4. Obtaining Spectrum Pairs

Our method requires a pair of CID and ETD spectra for the same peptide. This can be obtained either by (1) two separate LC-MS/MS runs of the same protein digest, or by (2) programming the mass spectrometer to fragment the same precursor ion twice in two consecutive scans both with CID and ETD. For the first setting, some treatments on the data are needed to obtain spectrum pairs. This can usually be done by pairing up the spectra with the same precursor m/z value and similar retention time from the two LC-MS/MS experiments. The pairing of peptide features from two LC-MS experiments is well studied in the label-free quantification method\textsuperscript{16}. For those unpaired spectra, we can still use the traditional \textit{de novo} sequencing methods for the analysis. However, ADEPTS is used for better \textit{de novo} sequencing accuracy on the paired spectra.

4.5. \textit{De Novo} Sequencing Versus Database Search

\textit{De novo} sequencing does not require the target peptide to be in an existing protein database, and therefore needs higher quality data than a database search method does. When the target peptide is in a given database, the accuracy of a \textit{de novo} sequencing algorithm will be no better than the database search method with the same peptide scoring function. Thus, \textit{de novo} sequencing is most-frequently used for the study of organisms whose genomes are unknown, or synthesized peptide, where database search will inevitably fail. Following the general practice in \textit{de novo} sequencing research, we use database search to provide a testing data for \textit{de novo} sequencing but do not compare the performance of these two approaches.

4.6. Availability of ADEPTS

ADEPTS software is freely available at http://monod.uwaterloo.ca/adepts. For each pair of CID and ETD spectra, the software takes a list of candidates, calculates their scores under the two spectra, and outputs the one with the best score. The candidate generation can be carried out by PEAKS or any other \textit{de novo} sequencing software for single spectrum. In addition to the commercial version used to carry out this research, a free online version of PEAKS is available at http://bioinfor.com/peaksonline.

Acknowledgment

This work is supported in part by NSERC (RGPIN 238748-2006), a start-up grant at University of Waterloo, a China National High-tech R&D program (2008AA02Z313), and a MITACS Acclerate PhD Fellowship program. We thank Drs. Baozhen Shan and Lei Xin for providing the mass spectrometry data and helping run PEAKS; Andreas Bertsch for providing CompNovo software and the mass
spectrometry data; Xi Han for useful discussions; and Christina Boucher for editing the manuscript. We also thank the anonymous reviewers for the constructive suggestions.

References

13. Chang CC, Lin CJ, LIBSVM: A library for support vector machines, Department of Computer Science, National Taiwan University, 2006.
15. Hsu CW, Chang CC, Lin CJ, A practical guide to support vector classification, Department of Computer Science, National Taiwan University, 2009.