Engineering Letters, 13:3, EL_13_3_15 (Advance online publication: 4 November 2006)

A Two-way Parallel Searching for Peptide Identification via Tandem Mass Spectrometry

Jung Hun Oh and Jean Gao *

Abstract

De novo peptide sequencing that determines the amino acid sequence of a peptide via tandem mass spectrometry (MS/MS) has been increasingly used nowadays in proteomics for protein identification. Current *de novo* methods generally employ a graph theory, which usually produces a large number of candidate sequences and causes heavy computational cost while trying to determine a sequence with less ambiguity. We present a novel *de novo* sequencing algorithm that greatly reduces the number of candidate sequences. By utilizing certain properties of b- and y-ion series in MS/MS spectrum, we propose a reliable two-way parallel searching algorithm to filter out the peptide candidates that are further pruned by an intensity evidence based screening criterion. In addition, we define an adjusted value required to determine the end node positions for b- and y-ion series in the case of +2 charged precursor. Experimental results demonstrate the efficiency and potency of the proposed algorithm.

Keywords: tandem mass spectrometry, MS/MS spectrum, proteomics, de novo peptide sequencing, two-way searching algorithm

1 Introduction

Tandem mass spectrometry (MS/MS) is a mass spectrometry that has more than one analyzer. It has been recognized as one of the most powerful tools in proteomics for protein identification [1, 2, 3, 4]. Prior to an MS/MS experiment, proteins are digested into peptides by ionization process such as electrospray ionization (ESI) or matrix-assisted laser desorption ionization (MALDI). Tandem mass spectrometer usually has two analyzers. The first analyzer selects ions of a particular charged peptide called *precursor* or *parent peptide* according to the mass-charge ratio (m/z). The selected peptide ions are fragmented by a process known as collision-induced dissociation (CID). Once fragmented ions pass through the second analyzer, they are detected by an ion detector which is connected with a data system where mass-charge ratios are stored together with their relative abundances to generate the MS/MS spectrum.

The fragmentation of a precursor peptide bond is determined by the properties of the peptide and the energy of CID. Fig.1 illustrates how a peptide with four amino acids can be cleft into different fragmentations [5, 6]. There are three different types of bonds in a peptide, *i.e.* CH-CO, CO-NH and NH-CH bonds. Each bond breakage produces two pieces. Therefore, there are six likely types of fragment ions for each amino acid residue: the N-terminal a, b, c fragments and C-terminal x, y, z fragments. The most common cleavage happens at CO-NH bonds by low-energy CID, which makes b- and y-ions dominant fragment types. For a peptide consisting of n amino acids, the possible number of fragment ions is $6(n-1)$.

Though the m/z ratios of fragment ions provide informative information for protein identification, no knowledge about the cleavage positions or charge of fragments is provided. In addition, contamination and inaccuracy of instrument may generate fake m/z ratio peaks. Consequently successful identification of protein structure still remains a challenging task. Current endeavors to unambiguous peptide identification can be generalized into two classes: *database search algorithms* and *de novo sequencing algorithms*. In the first, one seeks to determine a peptide sequence by the best match by comparing the experiment spectrum with the theoretical spectrum generated from a candidate peptide list which is obtained from a protein database [4, 6, 7, 8, 9, 10]. The early popular SEQUEST [8] computes the cross correlation value (Xcorr) between the experimental MS/MS spectrum and the hypothetical spectra generated from candidate peptides in the database with the same mass.

*Department of Computer Science and Engineering, The University of Texas, Arlington, TX 76019, USA Email: {joh, gao} @cse.uta.edu

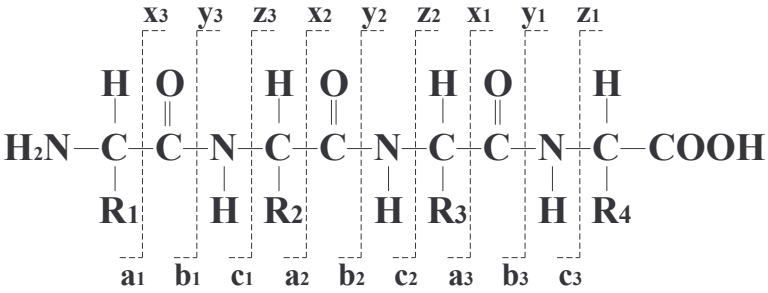

Figure 1: Structure of a peptide consisting of four amino acids linked by the C-N bonds. Six different types of fragment ions are produced by CID, *i.e.* a, b, c type fragment ions with N-terminal and x, y, z type fragment ions with C-terminal.

The candidate peptide producing the highest Xcorr value comes to be the first hit. However, a drawback of this algorithm is that its scoring function is not based on a rigorous metric. ProteinProspector [11] is composed of several tools for mining sequence databases in conjunction with mass spectrometry experiments. It takes into account the impact of mass measurement accuracy on protein identification experiment. Mascot [12] computes a probability based scoring to obtain the significance of the observed match between the experimental data and mass values calculated from a candidate peptide. SCOPE [7] calculates the probability density function based on a two-step model, *i.e.* the probability of a particular fragmentation pattern of a peptide and the probability that the observed spectrum is generated by the fragmentation pattern of the peptide. It is assumed that fragments are independent in order to make its complex probability problem computable. ProbID [4] makes use of the Bayesian approach as the basis for the probabilistic score function. It calculates the final posterior probability by considering several contributing factors. Despite the simple approach, the performance of this algorithm is comparable to industry-standard software. Lu & Chen [9] propose a suffix tree based approach to identify peptide sequence. The construction and search of a suffix tree are performed within a reasonable time. To rank candidate peptide sequences, a SEQUEST-like scoring function is used. Fu *et al.* [6] introduce a scoring algorithm by considering the correlative information among fragment ions to improve the peptide identification accuracy. The Kernel Spectral Dot Product (KSDP) extended from SDP is used as a scoring method. The success of all these algorithms depends largely on the completeness of database and the robustness of the scoring metrics, and can not be used for the identification of proteins from unknown genomes.

On the other hand, *de novo* algorithms rely heavily on the MS/MS spectrum for the determination of peptide sequence, and often do not use a database. SHERENGA [1] constructs an optimal path scoring in the spectrum graph, and automatically learns fragment types and intensity thresholds from test spectra. Lutefisk [13] converts an experimental spectrum into a spectrum graph with corresponding b type ion masses to make a sequence graph. To identify variants of known proteins in database, sequence candidates obtained from Lutefisk can be used as input. PEAKS [3] uses dynamic programming to compute 10000 sequences with the highest scores. It shows not only the confidence level of each output sequence but also the confidence level of each amino acid in the sequence. For each mass, this method first computes the reward and penalty. The reward is given, if there is a peak close to the mass; otherwise penalty. This algorithm tries to find a sequence such that its y and b ions maximize the total rewards at their mass values. Yan *et al.* [14] propose a novel graph approach to solve the problem of separating b-ions from y-ions, in which two types of edges are considered: a type-1 edge connects two peaks possibly of the same ion types and a type-2 edge connects two peaks possibly of different ion types. This algorithm does not deal with the PTM (Post-Translational Modification) problem. Jarman *et al.* [15] present a partial peptide identification based on a model of random sequence probability and the evidence defined as the instances of consecutive subsequences. In the approach, sequence hierarchy is used to represent a family of partial peptide candidates. Recently, Frank and Pevzner [16] present a new peptide sequencing algorithm using a probabilistic network as a scoring scheme that assigns a relevance score to peptide prefix masses. The probabilistic network represents three different types of relations such as correlations between fragment ions, the posi-

tional influence of the cleavage site and the influence of flanking amino acids to the cleavage site. These factors help to improve the accuracy of the peptide sequencing algorithm. Majority of *de novo* algorithms employ graph theory through which the experimental spectrum is transformed into a spectrum graph. Each peak in the spectrum is converted into several nodes representing different ion types. Two nodes are connected by an edge if the mass of an amino acid is approximately equal to the difference between the two nodes. For the final resulting directed acyclic graph (DAG), each path from start node to end node corresponds to a candidate sequence.

Regardless of the different mechanisms, a lot of research focus has been put on effective scoring metrics which is doubtlessly essential for unambiguous peptide identification. However, robust selection of peptide candidates can not only improve the computation, but also can eliminate false positives. In this paper, we will present an effective and efficient two-way *de novo* searching algorithm to reduce the number of candidate sequences dramatically by utilizing the properties of MS/MS spectrum and the confidence measurement based on the intensity values of spectrum. Moreover, to make a spectrum graph, the decision of start and end position is very important. We will also introduce our new positioning method for precursors with charge +1 and +2. Experimental results demonstrate the performance and efficacy of our novel approach.

This paper is organized as follows. In Section 2, we give a detailed description of our algorithm which embodies properties of MS/MS spectrum, relation between the precursor *m/z* and mass of peptide, normality test, and an elaboration of our two-way searching algorithm. Then in Section 3, we test our algorithm on public data, and finally we close our paper by conclusion and future work.

2 Algorithms

2.1 Random peptide sequence denotation

We will first introduce several peptide notations before moving on to the proposed algorithms. Let Σ be the alphabet set consisting of 20 amino acids. Each amino acid with a distinctive mass is represented by $m(a)$, $a \in \Sigma$. By denoting the product of Σ_1 and Σ_2 as $\Sigma_1 \times \Sigma_2 = \{ab \mid a \in \Sigma_1 \text{ and } b \in \Sigma_2\}$, the following expressions, $\Sigma^1 = \Sigma$, $\Sigma^2 = \Sigma \times \Sigma$, and $\Sigma^n = \Sigma \times \Sigma^{n-1}$, are possible. Σ^n means a set that includes all possible sequences with length n, whose number of elements $|\Sigma^n|$ is 20^n. So the set Σ^+ consisting of all possible sequences with different lengths made by 20 amino acids can be written as,

$$\Sigma^+ = \Sigma^1 \cup \Sigma^2 \cup \cdots \cup \Sigma^n \cup \cdots = \bigcup_{i=1}^{\infty} \Sigma^i. \quad (1)$$

Obviously $\Sigma = \{\text{A, C, } \ldots, \text{ Y}\}$, and the power expressions of Σ indicating peptides with different lengths are defined recursively as $\Sigma^1 = \{\text{A, C, } \ldots, \text{ Y}\}$, $\Sigma^2 = \{\text{AA, AC, } \ldots, \text{ AY, CA, CC, } \ldots, \text{ CY, YA, YC, } \ldots, \text{ YY}\}$, etc.. The union of all possible peptides is $\Sigma^+ = \{\text{A, C, } \ldots, \text{ Y, AA, AC, } \ldots, \text{ YY, AAA, AAC, } \ldots, \text{ YYY, } \ldots\}$.

A parent peptide $P = p_1 p_2 \ldots p_n$ is a sequence of amino acids, which consists of n amino acids. The mass of peptide P is $m(P) = \sum_{i=1}^{n} m(p_i)$ where $P \in \Sigma^n$ and the mass of each amino acid is $m(p_i)$. Let T be an element of Σ^+, *i.e.* $T \in \Sigma^+$. If the mass of T is approximately equal to that of the target peptide P, *i.e.* $|m(T) - m(P)| < \epsilon$, T will be one candidate sequence with respect to the parent peptide. ϵ is the tolerant error the mass spectrometer has.

2.2 Properties of MS/MS spectrum

The actual mass of peptide P can be written as $18+m(P)$. Number 18 comes from two extra hydrogen atoms and one extra oxygen atom at the C- and N-terminals where the mass of one hydrogen atom is approximately 1 Da (Dalton) and the mass of one oxygen atom is approximately 16 Da. The mass of b-ion with i amino acids, represented by b_i, can be computed as

$$b_i = 1 + m(p_1) + \cdots + m(p_i) = 1 + \sum_{j=1}^{i} m(p_j), \quad (2)$$

where mass 1 comes from one hydrogen atom attached to the b-ion type fragments. Similarly, the mass of y-ion with i amino acids, denoted by y_i, can be calculated by

$$y_i = 19 + m(p_{n-i+1}) + \cdots + m(p_n) = 19 + \sum_{j=n-i+1}^{n} m(p_j), \quad (3)$$

where mass 19 is due to three hydrogen atoms and one oxygen atom linked to the y-ion type fragments.

The difference of *m/z* ratio of two adjacent singly-charged b- or y-ions is the exact mass of one residue.

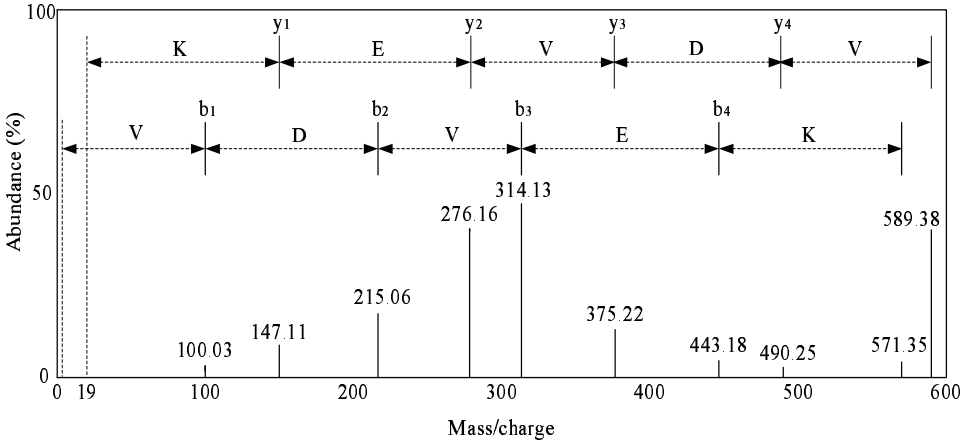

Figure 2: Hypothetical MS/MS spectrum and amino acid sequences.

However, in real MS/MS spectra, an ion may be charged with different values (+1, +2, ...), which make several different peaks. In addition, there is no information about the fragment ion type (b, y, ...) and fragmenting position. For a tandem mass spectrometry, we suppose that each fragment ion has a unique mass-charge ratio and each amino acid is fragmented [6, 7]. Therefore, the m/z value of an ion is equal to the mass of the ion. Before introducing our Two-way Searching Algorithm in Section 2.5, following two properties will be presented:

Property One: If the mass-charge ratio of precursor is given, then the positions of start and end nodes for both b- and y-ion series are known. (More details will be provided in Section 2.3).
Proof: Start nodes are always at 1 m/z for b-ion series and 19 m/z for y-ion series in the spectrum because of the extra attachments as explained above. And the end nodes of b- and y-ion series will happen at $m(P)+1$ m/z and $m(P)+19$ m/z, respectively by Eqs. (2) and (3). Therefore, the m/z positions of b_1, y_1, b_{n-1}, and y_{n-1} can be expressed as $b_1 = 1+m(p_1)$, $y_1 = 19 + m(p_n)$, $b_{n-1} = m(P) + 1 - m(p_n)$, and $y_{n-1} = m(P) + 19 - m(p_1)$, respectively.

Furthermore, let $S = \{(S_i, I_i) \mid 1 \leq i \leq k\}$ be an MS/MS spectrum ordered by m/z values where S_i and I_i denote the position and intensity of the i-th peak, and k is the number of peaks. Then the position of b_i in the spectrum is determined by the m/z value with the largest intensity I_j within the tolerant error.

$$b_i = \underbrace{\text{argmax}}_{S_j}\{I_j \mid |S_j - b_{i-1} - m(a)| < \epsilon\} \quad (4)$$

where $a \in \Sigma$. The same rule is applied to y_i position locating.

Property Two: There exists a pair-wise relationship between peaks in the b- and y-ion series. That is, the sequence identified in the b-ion series is the same as that identified in the y-ion series in the reversed order.
Proof: We can derive the equation $b_i + y_{n-i} = 20 + m(P)$ from Eqs. (2) and (3). Therefore, b_i and y_i can be expressed as $b_i = 20 + m(P) - y_{n-i}$ and $y_i = 20 + m(P) - b_{n-i}$, respectively. This means that the sequence of b-ion series is the same as that of y-ion series in the reverse order. Fig. 2 shows the two properties of an MS/MS spectrum.

2.3 Relation between the precursor m/z and mass of peptide

To simultaneously identify peptide from both the start and end nodes, knowledge of the positions of these nodes is important. Since the end nodes of b- and y-ion series happen at $m(P) + 1$ m/z and $m(P) + 19$ m/z, information about the mass of the target peptide sequence is required which can be obtained based on the precursor information. Let the m/z of precursor be P^z, where z is the charge of the ion. When $z = 1$, the positions of end node of b- and y-ion series are $P^1 - 18$ and P^1, because of $P^1 - 19 \approx m(P)$.

For $z \geq 2$ the above equation can not be used any more. Through experiment, the following equation is formed based on some heuristic δ value:

$$P^2 \times 2 - \delta - 19 \approx m(P), \quad (5)$$

where $0.9 \leq \delta \leq 1.0$. Therefore, for P^2 the positions

Table 1: Upper tail percentage points for Anderson-Darling statistic A^*.

α	0.2	0.15	0.1	0.05	0.025	0.01	0.005
A^*_α	0.509	0.561	0.631	0.752	0.873	1.035	1.159

Table 2: Normality test for distribution of measured mass-charge ratios.

i	$X_{(i)}$	$(X_{(i)}-\mu)/\sigma$	$F(Z_{(i)})$	$\ln(F(Z_{(i)}))+\ln(1-F(Z_{(r+1-i)}))$
1	0.006	-1.168	0.121	-5.314
2	0.009	-1.119	0.132	-12.626
3	0.028	-0.844	0.199	-17.774
4	0.037	-0.715	0.237	-16.947
5	0.043	-0.634	0.263	-19.084
6	0.089	0.029	0.512	-15.255
7	0.094	0.110	0.544	-11.885
8	0.109	0.321	0.626	-11.097
9	0.161	1.065	0.857	-6.411
10	0.171	1.210	0.887	-4.958
11	0.208	1.744	0.959	-3.588

of end nodes of b- and y-ion series are $P^2 \times 2 - \delta - 18$ and $P^2 \times 2 - \delta$, respectively. The position of start node in P^2 is the same as that of P^1. We used $\delta = 0.95$ in our experiment and showed δ can be used as the pertinent adjusted value.

2.4 Normality test

One assumption we used in our *de novo* peptide sequencing is that the measured mass-charge ratio can be modelled as a normal distribution as other researchers have done [7, 11]. We performed a goodness of fit test to confirm whether we can use the normal distribution as analysis model for our experiment data set or not. There are several approaches for assessing the underlying distribution of a data set. Among them, Anderson-Darling (AD) test which belongs to a class of distance test is known as a more powerful test than other distance tests. AD test shows a good performance in small samples as well as large samples. When the number of measured mass-charge ratios with respect to the center one within the tolerant error is sparse, AD test is more appropriate because of applying the cumulative distribution function(CDF) of the data set.

To test the normality, we define hypotheses:
H_0 : The distribution for the data set is a normal distribution.
H_1 : The distribution for the data set is a non-normal distribution.
Let X be a random sample with $X = (X_{(1)}, X_{(2)}, \ldots, X_{(r)})$ sorted in the ascending order with sample size r. The standardized value is $Z_{(i)} = (X_{(i)} - \mu)/\sigma$ where μ and σ denote mean and standard deviation for the sample data. The AD normality test is calculated by the following function:

$$AD = -\frac{1}{r}\{\sum_{i=1}^{r}(2i-1)\ln(F[Z_{(i)}](1-F[Z_{(r+1-i)}]))\} - r \tag{6}$$

where F is the standard normal cumulative probability and ln is the natural logarithm (base e). Eq. (6) is further modified by computing:

$$A^* = AD(1 + \frac{0.75}{r} + \frac{2.25}{r^2}). \tag{7}$$

If A^* exceeds the selected critical values given in Table 1, we will reject the null hypothesis at the $100\alpha\%$ level.

As an example for peptide YLYELAR spectrum shown in Fig. 3, we normalize the intensities of all peaks such that the highest peak is one and keep peaks with more than 1 % (0.00208) of maximum intensity (0.208). So we obtain the sample data (0.006, 0.009, 0.043, 0.089, 0.161, 0.171, 0.208, 0.109, 0.094, 0.037, 0.028). Mean and standard deviation of these intensities are $\mu = 0.087$ and $\sigma = 0.069$. We obtained $AD = 0.358$ and $A^* = 0.389$ through the procedure shown in Table 2. At significance level of 0.1, $A^* = 0.389 \leq \alpha = 0.631$. Therefore, we can assume normality for distribution of measured mass-charge ratios. Fig. 3 represents the distribution of measured mass-charge ratios and normal distribution. In general, if the p-value is 0.1 or more, we can assume normality.

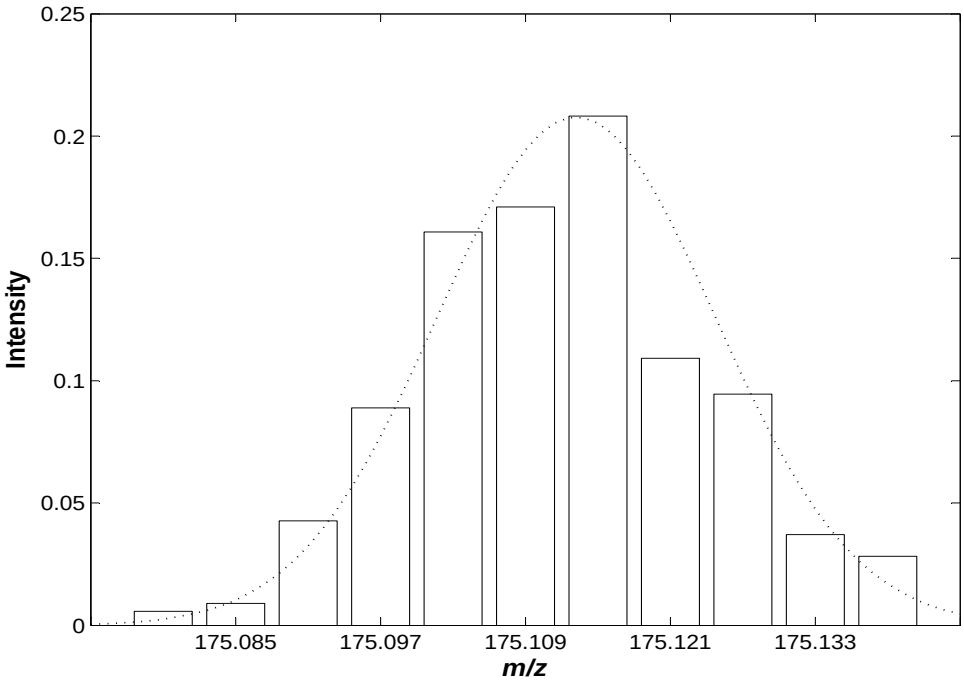

Figure 3: Distribution of measured mass-charge ratios (solid line bar) and normal distribution (dotted curve).

We found most measured mass-charge ratios are self-centered normal distribution. Therefore the application of normal distribution as the fundamental frame in the following scoring function is legitimate.

2.5 Two-way searching algorithm

Peptide candidate initial filtering by two-way searching

Our new two-way searching algorithm for MS/MS peptide sequencing begins with both start and end position localizations. In our approach, the positions of start and end nodes for b-ion and y-ion are determined in advance in the MS/MS spectrum, *i.e.* at 1 and $m(P) + 1$ m/z for b-ion series, and at 19 and $m(P) + 19$ m/z for y-ion series as shown in Fig. 2. During our two-way parallel searching, these four initial nodes will extend simultaneously by scanning the whole spectrum, where start nodes for b- and y-ion series proceed simultaneously in the forward direction, and end nodes in the backward direction at the same time. This procedure keeps going until some requirements are met. We denote the direction from low m/z to high m/z as the forward direction and from high m/z to low m/z as the reverse direction.

Four amino acid sets generated in the process of graph extension are denoted as Fb (forward for b-ion), Rb (reverse for b-ion), Fy (forward for y-ion), and Ry (reverse for y-ion) as shown in Fig. 4. At every extension of the graph, the candidate amino acids of Fb are compared with those of Ry. The common amino acids are kept and new nodes are added in positions corresponding to the m/z values. And the amino acids which are not in common are eliminated to reduce the computational burden. Likewise, the amino acids of Rb are compared with those of Fy simultaneously.

In stead of exhaustively checking all possible paths, we reduce the number of nodes in the spectrum graph effectively through such a method. Consequently it reduces the number of candidate sequences and computational cost to determine a sequence with the best score. After the successive progress, when the nodes of Fb meet those of Rb within an error range, *i.e.* $|b_i - b_j| < \epsilon$, $bi \in Fb$, and $bj \in Rb$, we merge two nodes into one and trace back in the two-way direction while storing the amino acids.

Finally, the two partial amino acids are concatenated into one complete sequence which will be used as one candidate sequence. It is also possible that these processes confront with a distance, which corresponds to one amino acid between Fb and Rb nodes, *i.e.* $|bi - bj - m(a)| < \epsilon$, $bi \in Fb$, $bj \in Rb$, and $a \in \Sigma$. If nodes of both sides cross over, the nodes disappear

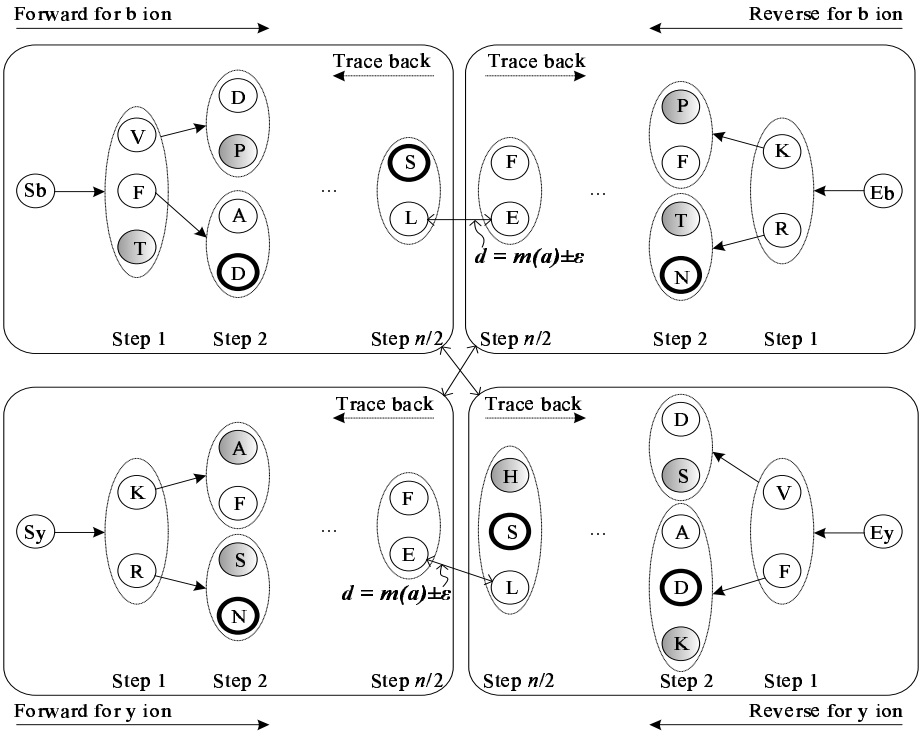

Figure 4: Two-way searching algorithm. Sb and Eb represent the start and end nodes of b-ion series, respectively. Sy and Ey represent the start and end nodes of y-ion series, respectively. Amino acid sets Fb and Ry always have the same amino acids. Same to sets Rb and Fy. Amino acids with gray color are deleted in the comparison procedure and amino acids with thick boundaries are removed in the further pruning procedure.

from the graph. The number of steps to obtain amino acid sequences of the same length as the target peptide is $\lceil n/2 \rceil$. The same rule is applied to Fy and Ry. Fig. 4 shows a diagram representing our algorithm.

Since the b- and y-ions may lose a water or ammonia molecule, it's necessary to employ all the related ion types for the series. Ions b, b - H_2O, and b - NH_3 for b-ion series are considered in the forward and reverse directions, which are the most frequent N-terminal ions, and y, y - H_2O, and y - NH_3 for y-ion series. In addition, we consider the a-ion which is also a dominant ion. After the initial filtering, we have a reduced amino acid candidate pool. Next we will introduce further pruning of candidate amino acids by a scoring function to determine the best sequence.

Scoring function for final candidate screening

By utilizing the piece-wise local region intensity values of MS/MS, we define an evidence based scoring function for screening out the best optimal peptide candidate. Based on the normality test verification introduced in Section 2.4, we can let I_{bi} be the Gaussian sum of all peak intensities close to b_i within a tolerant error ϵ:

$$I_{bi} = \sum_{b_i+\epsilon}^{b_i-\epsilon} \frac{I_j}{I_M} \exp(-(S_j - b_i)^2/2\sigma^2), \tag{8}$$

where I_M is the highest intensity in the whole spectrum, I_j is the individual peak intensity of the local region, and S_j is the corresponding m/z value. In the procedure of normalization, intensities of all peaks are divided by the highest intensity. Standard deviation σ represents to which extent the peak positions in the experimental spectrum deviate from the theoretical ones. Without loss of generality, based on the normality test, we can assume the peak having the highest intensity in the whole spectrum is most likely to be a real fragment and neighbor peaks close to the peak will form the normal distribution. Standard deviation σ can be heuristically determined by,

$$\sigma_{min} = \underbrace{\text{argmin}}_{\sigma} \sum_{b_i+\epsilon}^{b_i-\epsilon} (\exp(-(S_j - S_M)^2/2\sigma^2) - \frac{I_j}{I_M}), \tag{9}$$

where S_M is the m/z value corresponding to the highest intensity I_M.

Let x be the mass of b-ion, then the masses of a-ion, b - H_2O, and b - NH_3 are $x-28$, $x-18$, and $x-17$, respectively, *i.e.* differences $\triangle = \{-28, -18, -17\}$. Given the b_i we define:

$$I_{bi\ -\ NH_3} = \sum \frac{I_j}{I_M} \exp(-(S_j - S_{b_i - NH_3})^2 / 2\sigma^2), \tag{10}$$

where $S_{b_i-NH_3} = \underbrace{\text{argmax}}_{S_j}\{I_j \mid |b_i - S_j - 17| < \epsilon\}$.

Likewise, we can compute $I_{b_i-H_2O}$, I_{b_i-a}, I_{y_i}, $I_{y_i-H_2O}$ and $I_{y_i-NH_3}$. The total intensity of ions related to b_i is

$$I_{Tbi} = I_{bi} + I_{bi-NH_3} + I_{bi-H_2O} + I_{bi-a}. \tag{11}$$

And I_{Tyi} is expressed as follows,

$$I_{Tyi} = I_{yi} + I_{yi-NH_3} + I_{yi-H_2O}. \tag{12}$$

The total intensity of b-ion series in the forward direction is:

$$I_{Fb} = \sum_{i=1}^{n/2} I_{Tbi}, \tag{13}$$

where n is the number of steps to obtain the peptide sequence. The total intensity of b-ion series in the reverse direction is

$$I_{Rb} = \sum_{i=(n/2)+1}^{n-1} I_{Tbi}. \tag{14}$$

Similarly, we can compute I_{Ry} and I_{Fy}. A legitimate intensity scoring function is obtained by summing the four direction total intensities and the top scoring sequence among all candidate sequences becomes the best candidate:

$$Scoring = I_{Fb} + I_{Rb} + I_{Fy} + I_{Ry}. \tag{15}$$

This scoring function is reasonable because many of the noise peaks have a low intensity value. By incorporating abundance difference of b- and y-ions, *i.e.*, y-ions are usually more ample than b-ions, we may apply different weights to the total summing intensity of b-ion series and the total summing intensity of y-ion series. Then, the scoring function can be modified as

$$Scoring = \lambda(I_{Fb} + I_{Rb}) + (1-\lambda)(I_{Fy} + I_{Ry}). \tag{16}$$

Obviously λ is set as less than or equal to 0.5.

Before the scoring function is applied, for further pruning of candidate pool, we define a screening criterion from the calculation of total intensities I_{Tbi} and $I_{Ty(n-i)}$ of ions in the i-th step of graph extension where $|b_j - b_{j-1}| \approx |y_{n-j+1} - y_{n-j}| \approx m(a)$ with $b_{-1} = \text{S}_\text{b}$, $y_n = \text{Ey}$ and $1 \le j \le i$. This screening criterion is used at every step of node extension of the spectrum graph. Suppose there exist q candidate amino acids in the i-th step after eliminating amino acids which are not in common between Fb and Ry. Let I_i be a set whose elements indicate the sum of I_{Tbi} and $I_{Ty(n-i)}$ where b_i and y_{n-i} are a complementary ion pair.

$$I_i = \{I_{Tbi}{}^1 + I_{Ty(n-i)}{}^1, \ldots, I_{Tbi}{}^q + I_{Ty(n-i)}{}^q\}. \tag{17}$$

The elements in set I_i are sorted in ascending order, and the first β% amino acids are removed from Fb and Ry. The same rule is applied to every step of Fy and Rb.

3 Experimental Results

The two-way peptide sequencing algorithm was implemented by using C++ codes. We employed nine peptide data sets whose ground truths were given, among which four data sets with precursor charge +1 and five data sets of +2. These data sets were obtained from QSTAR instrument which is a hybrid quadrupole/time-of-flight (Q-TOF) tandem mass spectrometer. The machine is operated in DDA mode that is a commercial bovine Cytochrome-C. In this study, we used $\epsilon = 0.3$ and $\lambda = 0.5$.

Table 3 shows results of our algorithm with $\beta = 35\%$ ($\beta = 20\%$ for spectrum 678.3 m/z) as a screening criterion ($\delta = 0.95$ for precursor of charge +2). To show the result which is most optimal to benchmark, instead of listing the most optimal identified peptide, *i.e.*, the one with highest scoring value or with ranking 1, we show the sequence which is within ranking 3 range. We compared results of the proposed method with those of Lutefisk [13]. Lutefisk converts an experimental spectrum into a sequence graph where partial sequences are examined. To reduce candidates, after finding complete sequences, sequences that appear to have been derived from alternating b-type and y-type ions and that are derived mostly from the low-mass fragments are discarded. Finally, Lutefisk yields at most 50 candidates. Users can set the parameter. When we set the parameter as the maximum value, 50, 45, 50 and 50 candidates were made in 634.4 678.3 779.4 and 927.4 m/z, respectively.

Table 3: Experimental results of peptide sequencing with $\beta = 35\%$ ($\beta = 20\%$ for 678.3 m/z) in our method. For comparison, Lutefisk was performed. We cannot distinguish between the isobaric amino acid pair of leucine (L) and isoleucine (I), and pair of glutamine (Q) and lysine (K) as with most *de novo* methods, since m(I)=m(L)=113.16 Da, m(Q)=128.13 Da, m(K)=128.17 Da. The value shown in brackets of Lutefisk method represents an approximate mass of the remaining amino acid residues.

Spectrum	z	Correct sequence	Two-way searching algorithm		Lutefisk	
			Sequence	Rank	Sequence	Rank
634.4	1	IFVQK	**LFVQK**	1	**LFVQK**	1
678.3	1	YIPGTK	**YLPGTK**	1	**YLPGTK**	1
779.4	1	MIFAGIK	**MLFAGLK**	2	[244.12]**FAGLK**	1
927.4	1	YLYEIAR	**YLYELAR**	1	[276.11]**YE**[184.08]**R**	1
584.8	2	TGPNLHGLFGR	**TGPNLHGLFGR**	3	[409.25]**R**	1
689.9	2	HGTVVLTALGGILK	**HGTVVLTALGGLLK**	3	[194.08][WY]**K**	2
728.8	2	TGQAPGFSYTDANK	**TGQAPGFS**PQQP**NK**	1	AQGT[HS]**K**	3
792.9	2	KTGQAPGFSYTDAN	**KTG**AG**AP**AMAPQG**DAN**	1	[209.58]NH**ANK**	3
943.0	2	YLEFISDAIIHVLHSK	**YLEFL**ALTT**LHVLHSK**	3	[222.56]**HVLH**[215.12]	1

Table 4: The number of candidates and rankings as the screening ratio changes. The numerator indicates the ranking of the correct sequence our algorithm made, and the denominator represents the total number of candidate sequences. 0% means no screen is adopted.

Spectrum	0%	10%	20%	30%	40%	50%	60%
634.4	1/91	1/27	1/21	1/19	1/14	1/12	0/6
678.3	1/1	1/1	1/1	0/0	0/0	0/0	0/0
779.4	2/173	2/152	2/93	2/64	2/10	2/6	2/6
927.4	1/5197	1/819	1/360	1/239	1/195	1/126	1/24

For precursor of charge +2, we found that there exists an equation $P^2 \times 2 - \delta - 19 \approx m(P)$ where $0.9 \leq \delta \leq 1.0$. We used $\delta = 0.95$ to determine the position of end node in the graph extension. This value is greater than our tolerant error 0.3. Therefore if we use $P^2 \times 2 - 19 \approx m(P)$ as the position of end node as in other methods which proceed in the forward direction only, our algorithm will not provide good answers. Thus the positions of end node of b- and y-ion series are respectively $P^2 \times 2 - 0.95 - 18$ and $P^2 \times 2 - 0.95$. For example for spectrum 584.8 m/z, the end nodes of b- and y-ion series come to appear in the $584.8 \times 2 - 0.95 - 18 = 1150.65$ m/z and $584.8 \times 2 - 0.95 = 1168.65$ m/z. The position of start node in P^2 is the same as that of P^1. Out of the nine peptides, our algorithm came up six sequences identical with ground truth, and the remaining three turned out to be almost the same as benchmark.

Table 4 shows the number of candidates change as the screening criterion β varying from 0% to 60% for the charged +1 sequences. Although β changes, the rankings remain unchanged. For spectrum 678.3 m/z, since the number of candidates after initial two-way searching was reduced to only one as seen in Fig. 5, no screen was applied. With 10% further pruning at every step of graph extension, we can reduce the number of candidates significantly up to 70% and 84% for spectra 634.4 and 927.4 m/z, respectively. Therefore, by adopting a proper screening criterion to our algorithm, we can reduce the processing time effectively.

4 Conclusion

In this paper, we presented a novel *de novo* approach called two-way searching algorithm for determining the sequence of peptide. The main contribution of this paper lies in the greatly reduced number of peptide candidates. Based on the property that the same identification of peptide sequence will be resulted from b-ion or y-ion series, we obtained a list of peptide candidates by simultaneously searching from four differ-

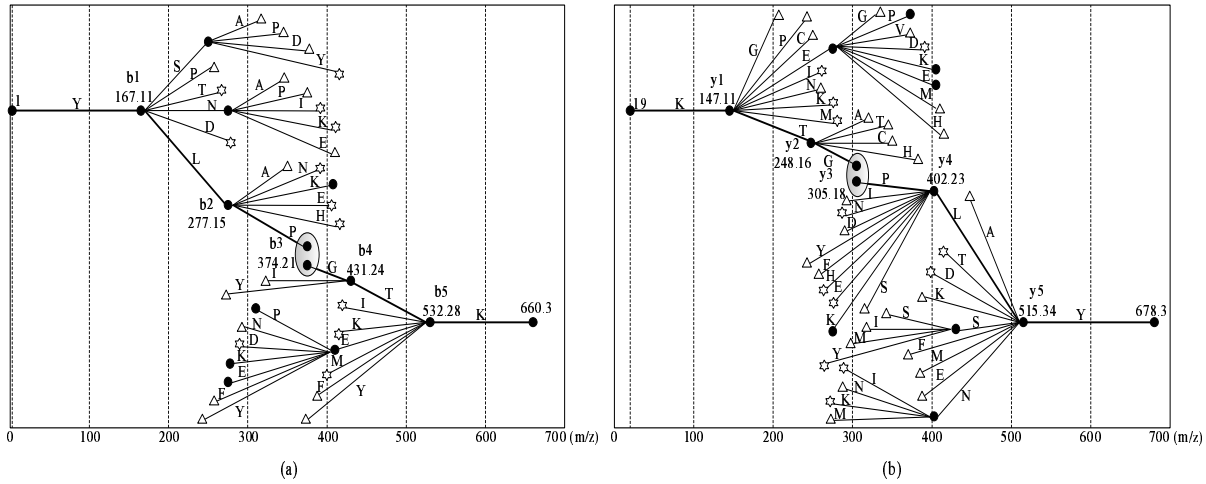

Figure 5: Example of peptide sequence YIPGTK in the graph extension: (a)The sequence of b-ion series; (b) The sequence of y-ion series. Amino acids with triangle symbols are deleted in the procedure of comparison and amino acids with star symbols are removed from the candidate amino acid pool in the procedure of further pruning. Amino acids represented in black circles are reserved. Thick line presents the path of one candidate sequence.

ent start and end positions and filtering out peptide sequences which violate this property. The initially filtered candidate pool is further pruned by incorporating a screening criterion based on the local region intensities of the spectrum. The final optimal best candidate is singled out based on the highest confidence defined as the global intensity from two-way search results. Contributions of this paper also come from the determination of the end nodes for b and y-ion series in the case of the charged +2 precursor. For the future work, we will improve our algorithm by introducing gap edges corresponding to the di- and tri-peptides in a spectrum graph. Also we will modify the scoring algorithm by using a probabilistic model to make the pruning more robust.

References

[1] Danick, V., Addona, T.A., Clauser, K.R., Vath, J.E., and Pevzner, P.A., "*De novo* peptide sequencing via tandem mass spectrometry," *J. Comput. Biol.*, V6, pp. 327-342, 1999.

[2] Lu, B., and Chen, T., "A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry," *J. Comput. Biol.*, V10, pp. 1-12, 2003.

[3] Ma, B., Zhang, K.Z., Hendrie, C., Liang, C.Z., Li, M., Doherty-Kirby, A., and Lajoie, G., "PEAKS: powerful software for peptide *de novo* sequencing by MS/MS," *Rapid Communications in Mass Spectrometry*, V17, pp. 2337-2342, 2003.

[4] Zhang, N., Aebersold, R., and Schwikowski, B., "ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data," *Proteomics* V2, pp. 1406-1412, 2002.

[5] Ma, B., Zhang, K., and Liang, C., "An Effective Algorithm for the Peptide *De Novo* Sequencing from MS/MS Spectrum," *CPM*, pp. 266-278, 2003.

[6] Fu, Y., Yang, Q., Sun, R., Li, D., Zeng, R., Ling, C.X, and Gao, W., "Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry," *Bioinformatics*, V20, pp. 1948-1954, 2004.

[7] Bafna, V., and Edwards, N., "SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database," *Bioinformatics*, V17, pp. S13-S21, 2001.

[8] Eng, J.K., McCormack, A.L., and Yates, J.R., "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database," *J. Am. Soc. Mass Spectrom*, V5, pp. 976-989, 1994.

[9] Lu, B., and Chen, T., "A suffix tree approach to the interpretation of tandem mass spectra: applications to peptides of non-specific digestion

and post-translational modifications," *Bioinformatics*, V19, pp. ii113-ii121, 2003.

[10] Taylor, J.A., and Johnson, R.S., "Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry," *Rapid Communications in Mass Spectrometry*, V11, 1067-1075, 1997.

[11] Clauser, B.T., Baker, P.R., and Burlingame, A.L., "Role of accurate mass measurement (+/- 10ppm) in protein identification strategies employing MS or MS/MS," *Analytical Chem.*, V71, pp. 2871-2882, 1999.

[12] Perkins, D.N., Pappin, D.J.C, Creagy, D.M., and Cottrell, J.S., "Probability-based protein indentification by searching sequence databases using mass spectrometry data," *Electrophoresis*, V20, pp. 3551-3567, 1999.

[13] Taylor, J.A., and Johnson, R.S., "Implementation and uses of automated *de novo* peptide sequencing by tandem mass spectrometry," *Anal. Chem.*, V73, pp. 2594-2604, 2001.

[14] Yan, B., Pan, C., Olman, V.N., Hettich, R.L., and Xu, Y., "A graph-theoretic approach to separation of b and y ions in tandem mass spectra," *Bioinformatics*, V21, pp. 563-574, 2005.

[15] Jarman, K.D., Cannon, W.R., Jarman, K.H., and Heredia-Langner, A., "A model of random sequences for *de novo* peptide sequencing," *Proceedings of IEEE Symposium on Bioinformatics and Bioengineering*, pp. 206-213, 2003.

[16] Frank, A., and Pevzner, P., "PepNovo: De novo peptide sequencing via probabilistic network modeling," *Analytical Chemistry*, V77, pp. 964-73, 2005.