COMPLEXITY AND SCORING FUNCTION OF MS/MS PEPTIDE DE NOVO SEQUENCING

Changjiang Xu and Bin Ma*
Department of Computer Science, University of Western Ontario, London, ON N6A 5B7, Canada
Email: cjxu@csd.uwo.ca
**Email: bma@csd.uwo.ca*

Tandem mass spectrometry (MS/MS) has become a standard way for identifying peptides and proteins. A scoring function plays an important role in the MS/MS data analysis. De novo sequencing is the computational step to derive a peptide sequence from an MS/MS spectrum, normally by constructing the peptide that maximizes the scoring function. A number of polynomial time algorithms have been developed based on scoring functions that consider only either the N-terminal or C-terminal fragment ions of the peptide. It remains unknown whether the consideration of the internal fragment ions will still be polynomial time solvable. In this paper, we prove that the internal fragment ions make the de novo sequencing problem NP-complete. We also propose a regression model based scoring method to incorporate correlations between the fragment ions. Our scoring function is combined with PEAKS de novo sequencing algorithm and tested on ion trap data. The experimental results show that the regression model based scoring method can remarkably improve the de novo sequencing accuracy.

1. INTRODUCTION

Identification of the proteins existing in a tissue is frequently a key step in proteomics research. In recent years, tandem mass spectrometry (MS/MS) has become a powerful analytical tool for protein and peptide identification[1, 2]. It is difficult to identify intact proteins directly. Hence the proteins are digested into short peptides and the individual peptides are identified separately by using MS/MS.

The peptide identification is to deduce a peptide sequence that best matches an MS/MS spectrum. This technique can give accurate peptide identifications provided that a high quality MS/MS spectrum is available. However, currently only a fraction of the acquired spectra lead to positive peptide identifications. The reasons involve various factors[3–5], which include poor fragmentation of the selected precursor ions, chemical contaminants obscuring peptide fragment ions, unanticipated residues caused by post-translational modifications.

Over the past decade, numerous computational approaches and software programs have been developed for MS/MS peptide identification. These can be categorized into four classes[6]: sequence database searching, de novo sequencing, sequence tagging, and consensus of multiple search engines. The database searching finds the best matching peptide from a protein sequence database. The popular algorithms using this approach include Sequest[7], Mascot[8], Tandem[9], and Omssa[10]. The de novo sequencing is equivalent to searching for the optimal peptide from an universal peptide database that includes all linear combinations of amino acids. The efficient algorithms for computing the optimal peptide are required to avoid the explicit searching. Among the de novo algorithms are Lutefisk[11, 12], Sherenga[13], Compute-Q[14], PEAKS[15, 16], PepNovo[17], and NovoHMM[18]. The sequence tagging is to find the best peptide by searching a database with sequence tags that may be inferred by de novo sequencing. The existing algorithms include GutenTag[19], OpenSea[20], SPIDER[21], and DeNovoID[22]. The consensus method combines several different programs to increase the confidence and coverage[23, 24]. A review of most of the protein and peptide identification algorithms can be found in Ref. 4.

Scoring function, which is used to evaluate the matches between candidate peptides and the MS/MS spectrum, is a key component in peptide identification. The scoring function is usually described by a mathematical model that quantifies the likelihood that a given sequence is the correct peptide

*Corresponding author.

sequence that generates the MS/MS spectrum. The basic principle of the MS/MS peptide identification is that the peaks in the spectrum are produced by the fragment ions of the peptide. The scoring function evaluates the peptide with the number and intensities of the peaks. A lot of scoring methods have been developed for database searching [7, 25–30] and for de novo sequencing [13–15, 17, 18]. Many scoring functions examine the correlations between the fragment ions. Different techniques such as likelihood test[13, 26, 27], Hidden Markov model[28], decision trees[29], and Bayesian network[17] are used.

However, all of the polynomial time de novo sequencing algorithms developed previously are base on scoring functions that only use the N-terminal or C-terminal ions of the peptide. None of them utilizes the internal fragment ions. Some de novo sequencing programs such as PEAKS[15] use a scoring function that takes into account of internal fragment ions to further re-evaluate the peptide candidates. However, the de novo sequencing step could not account for the internal fragment ions since the internal fragment ions will make de novo sequencing problem very similar to the two well-known open problems: the *partial digest problem* [31] and the problem 12.116 in Ref. 32. Neither of these two open problems has known polynomial time algorithms.

In this paper, we prove that the de novo sequencing with internal fragment ions is in fact NP-complete. Therefore, future research in utilizing internal fragment ions should focus on either heuristic algorithms or exponential time algorithms that run fast enough for smaller instances. This also justifies the two-step approach used in PEAKS [15] to utilize the internal fragment ions. The second contribution of the paper is to present a new scoring function that is based on a regression model (RM). The RM-based scoring method can efficiently exploit the relationship between different fragment ion types. The new scoring function is used to refine PEAKS' de novo sequencing results, and the significant improvement is achieved.

The remainder of this paper is organized as follows. Section 2 proves the complexity of the de novo sequencing with internal fragment ions. Section 3 presents the RM-based scoring method. Section 4 gives the comparison of the new scoring method with PEAKS and PepNovo.

2. COMPLEXITY OF DE NOVO SEQUENCING WITH INTERNAL FRAGMENT IONS

2.1. Notations and preliminaries

There are 20 common amino acid residues, denoted by 20 different single-letter codes. Normally, a peptide is a string over the alphabet of these 20 letters. The mass of a residue a is denoted by $m(a)$. For a string of residues $a_1a_2\ldots a_n$, define $m(a_1a_2\ldots a_n) = \sum_{i=1}^{n} m(a_i)$.

In an MS/MS, a peptide is fragmented into different ion types. In low energy collision-induced dissociation (CID), the fragmentation produces mostly y-ions (the fragment with C-terminus) and b-ions (the fragment with N-terminus). These are the most interesting ion types in de novo sequencing algorithms. However, the peptide is frequently fragmented more than once. This causes the internal fragment ions (the fragments without the terminus). Internal fragment ions are more often observed when the collision energy is high, such as in the TOF/TOF mass spectrometers. When the internal fragment ion contains only one amino acid, it is also called the immonium ion. In this paper we only consider the internal fragment ions with two or more amino acids. For example, the peptide AGEDK has four b-ions A, AG, AGE, and AGED; four y-ions K, DK, EDK, and GEDK; and three internal fragment ions GE, ED, and GED.

The mass value of each ion is the total mass value of its residues plus a constant associated with the ion type. In practice, when the ions retain only one positive charge, the constants for b, y and internal ions are 1, 19, and 1, respectively. Therefore, the b-ions of the peptide $a_1a_2\ldots a_n$ have mass values $B = \{1 + m(a_1a_2\ldots a_k) \,|\, k = 1,\ldots,n-1\}$; the y-ions of the peptide have mass values $Y = \{19 + m(a_ka_{k+1}\ldots a_n) \,|\, k = 2,\ldots,n\}$; and the internal fragment ions have mass values $I = \{1 + m(a_ka_{k+1}\ldots a_j) \,|\, 1 < k < j < n\}$.

An MS/MS spectrum provides the signal intensity at every mass value (in fact, mass to charge ratio). This can be used to define a scoring function to select peptides. First, three functions $f_y(x)$, $f_b(x)$,

and $f_I(x)$ are used to define the score of that a y, b, and internal fragment ion is at mass x, respectively. Suppose the sets of mass values of the y, b, and internal fragment ions of a peptide P are Y, B, and I, respectively. Then the score of the peptide is defined by:

$$\text{score}(P) = \sum_{x \in Y} f_y(x) + \sum_{x \in B \setminus Y} f_b(x) + \sum_{x \in I \setminus (Y \cup B)} f_I(x)$$

The de novo sequencing problem is to compute the peptide sequence P over an alphabet Σ such that score(P) is maximized. Notice that when the mass values of multiple ions overlap, only the score of one (the most important one) is counted in the scoring function.

A simplified version is to let $f_y(x) = c_y f(x)$, $f_b(x) = c_b f(x)$, and $f_I(x) = c_I f(x)$ for some constants c_y, c_b, c_I and a function $f(x)$. In next section we will prove that even this simplified version is NP-complete.

2.2. NP-completeness

The de novo sequencing problem has been extensively studied. Polynomial time algorithms have been proposed. However, all polynomial time algorithms consider only the ions with either the N or C-terminus. In another word, $f_I(x) = 0$ for all x. Some software systems such as PEAKS software use internal fragment ions to refine the results after the de novo sequencing algorithm finds a list of candidates. But internal fragment ions are not used in the de novo sequencing algorithm. An MS/MS spectrum usually contains a large number of ions that are neither y or b ions. A significant portion of these additional ions are internal fragments. The use of these ions will inevitably improve the accuracy. However, it is unknown whether an efficient algorithm exists when these ions are taken into account. Our result in this section answers the question negatively. That is, the finding of optimal sequence is NP-complete when the internal fragment ions are counted. Our result suggests that when internal fragment ions are counted, most likely no polynomial time algorithm exists (unless P=NP). And therefore, research efforts should be put to design either heuristic algorithms or exponential time algorithms that run fast enough when the sequence is short.

Theorem 2.1. *De novo sequencing is NP-complete if internal fragment ions are counted.*

Proof. Obviously the problem is in NP because given any peptide sequence, the score can be calculated in polynomial time. In what follows we reduce the Max-Cut-3 problem to our problem. A Max-Cut-3 instance is a graph $\langle V, E\rangle$, where $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$. Each vertex has degree exactly 3. The optimal solution is two disjoint vertex sets V_1 and V_2 such that (a) $V_1 \cup V_2 = V$ and (b) the number of edges with the two adjacent vertices in two different sets is maximized. It is well known that Max-Cut-3 is NP-hard [33].

Our constructed instance of de novo sequencing has only five letters in the alphabet Σ. They are G, E, A, D, and W. The mass values of the five letters are 57, 129, 71, 115, and 186, respectively. [a] Therefore, $m(\mathtt{G}) + m(\mathtt{E}) = m(\mathtt{A}) + m(\mathtt{D}) = m(\mathtt{W})$.

For each vertex v_i, suppose it is adjacent with three edges e_j, e_k, and e_l, we construct the following string

$$s_i = t_i \underbrace{\mathtt{WW\ldots W}}_{j-1} \mathtt{AD} \underbrace{\mathtt{WW\ldots W}}_{k-j-1} \mathtt{AD} \underbrace{\mathtt{WW\ldots W}}_{l-k-1} \mathtt{AD} \underbrace{\mathtt{WW\ldots W}}_{2m-l},$$

where t_i can be one of W, GE and EG, all having the same mass. Let

$$s_0 = \mathtt{W}^{2m+1} = \underbrace{\mathtt{WW\ldots W}}_{2m+1},$$

and S be the concatenation of the constructed string: $S = s_0 s_1 s_2 \ldots s_n s_0$.

The idea of our construction is to define a spectrum so that the optimal solution of the de novo sequencing problem has the form of S. This will be achieved by carefully design the y and b ion scores $f_y(x) = c_y f(x)$ and $f_b(x) = c_b f(x)$. Then we will use the internal fragment ion score $f_I(x) = c_I f(x)$ to "fine tune" the t_i in each s_i. Depending on whether t_i takes GE or EG, s_i will produce different internal fragment ions EW...WA or GW...WA. For an edge $e_k = (v_i, v_j)$, if t_i and t_j are different, then both of $\mathtt{EW}^{k-1}\mathtt{A}$ and $\mathtt{GW}^{k-1}\mathtt{A}$ will contribute to the score. However, if t_i and t_j are the same, then only one of

[a] These are the nominal mass values of the five real amino acid residues coded with the five letters.

the two will contribute to the score. Thus, the solution of the de novo sequencing is connected to the solution of the Max-Cut-3. The detail of the construction follows.

Let Y, Y' and Y'' be the y-ions of S by setting all t_i to W, GE and EG, respectively. Similarly, let B, B' and B'' be the b-ions of S by setting all t_i to W, GE and EG, respectively. Obviously $Y \subset Y' \cap Y''$ and $B \subset B' \cap B''$.

Furthermore, let $I = \{1 + m(\mathtt{EW}^{k-1}\mathtt{A}), 1 + m(\mathtt{GW}^{k-1}\mathtt{A}) \mid k = 1, \ldots, m\}$. Clearly, I consists of the internal fragment ions resulted from the fragments in t_i and one of the three AD in the same s_i. Each pair of the internal fragment ions in I corresponds to an edge e_k. Therefore, $|I| = 2m$. Because of the existence of s_0 at both ends of S. It is easy to verify the three sets Y, B, and I do not overlap each other. Let $I^* = \{1 + m(a\mathtt{W}^j b) \mid a \in \{\mathtt{E}, \mathtt{G}, \mathtt{W}\}, a \in \{\mathtt{A}, \mathtt{D}, \mathtt{W}\}, 0 \leq j \leq 2mn + 4m + n\}$. One can easily verify that all the the internal fragment ions of S are in I^*, no matter how individual t_i's take their values.

We assign values of $f(x)$ as follows:

$$f(x) = \begin{cases} 1, & x \in Y \cup B \\ \frac{1}{4m}, & x \in I \\ 0, & x \in (Y' \cup Y'' \cup B' \cup B'' \cup I^*) \\ & \quad \backslash (Y \cup B \cup I) \\ -1, & otherwise \end{cases}$$

Let $f_y(x) = f_b(x) = f(x)$, and $f_I(x) = \frac{1}{n} f(x)$.

Because $f(x) = 1$ for $x \in Y \cup B$ and $f(x) = 0$ for $x \in (Y' \cup Y'' \cup B' \cup B'' \cup I^*) \setminus (Y \cup B \cup I)$, the y and b ion scores will enforce the optimal solution to have the form of S, as proved in the following lemma.

Lemma 2.1. *Any optimal solution can be modified to have the form of S. In addition, t_i must be either* GE *or* EG.

Proof. According to the definition of $f(x)$, all the y, b and internal ions of the sequences with form S have scores greater than or equal to 0 in the definition; and all of the ions in $Y \cup B$ will contribute score 1. Therefore, any sequence with form S will have a score no less than $|Y \cup B|$. Because $Y \cap B = \emptyset$, $|Y \cup B| = |Y| + |B|$.

On the other hand, even if all the positive positions are matched by y and b ions, the score is no more than $|Y| + |B| + \frac{1}{2}$ because $|I| = 2m$. Consequently, an optimal solution needs to match all mass values in $Y \cup B$ using its y and b ions. This ensures that it has the form of $X^{(2m+1)(n+2)}$, where each segment X can independently take one of W, EG, GE, AD, and DA. If this optimal solution does not satisfy the lemma, then for every segment X that contradicts the lemma, there are two possible cases.

Case 1. X takes W but S asks for GE or EG as a t_i. In this case, we simply change X from W to either GE or EG. Because of the definition of $f(x)$, this will not reduce the score.

Case 2. X is a two letter segment, and is different from what S asks for. In this case, one can easily check that the y-ion caused by the fragmentation of the two letter segment in X will give a -1 value. This will make the total score less than $|Y| + |B|$. Therefore, this case does not exist.

Thus, the lemma is proved. □

The following lemma concludes our proof of Theorem 2.1.

Lemma 2.2. *The spectrum has an optimal solution with value $4mn + 8m + 2n + 2 + \frac{m+K}{4m}$ if and only if the Max-Cut-3 instance has an optimal solution that cuts K edges.*

Proof. Note that $|Y| + |B| = 4mn + 8m + 2n + 2$. That is, any solution that satisfies Lemma 2.1 should gain score $4mn + 8m + 2n + 2$ using the y and b ions. The $\frac{m+K}{4m}$ portion is determined by the internal cleavage ions in I.

"⇐" Suppose the optimal cut is $V = V_1 \cup V_2$. For each $v_i \in V_1$, let $t_i =$ GE. For each $v_i \in V_2$, let $t_i =$ EG. Then S is a solution of the de novo sequencing problem. All the mass values in $Y \cup B$ are matched. For each edge e_k, if it is cut, the pair of mass values $1 + m(\mathtt{EW}^{k-1}\mathtt{A})$ and $1 + m(\mathtt{GW}^{k-1}\mathtt{A})$ in I are both matched. If e_k is not cut, then exactly one is matched. This gives score $\frac{m+K}{4m}$ from internal fragment ions.

"⇒" Because of Lemma 2.1, each t_i is either GE or EG. Let V_1 consist of all v_i such that $t_i =$ GE, and V_2 consist of all v_i such that $t_i =$ EG. We get a cut for the Max-Cut-3. This way, it is clear that an edge e_k is cut if and only if the pair of mass values $1 + m(\mathtt{EW}^{k-1}\mathtt{A})$ and $1 + m(\mathtt{GW}^{k-1}\mathtt{A})$ in I are both

matched. The score $\frac{m+K}{4m}$ contributed by the ions in I ensures that exactly K pairs are both matched. That is, exactly K edges are cut. □

The proof of Lemma 2.2 finishes the proof of Theorem 2.1. □

3. REGRESSION MODEL BASED SCORING METHODS FOR DE NOVO SEQUENCING

3.1. Relationship between fragment ions

When peptides are fragmented by collision-induced dissociation (CID) in a tandem mass spectrometer, the resulting fragment ions can be categorized into three classes. One is the complementary fragment ions generated from one backbone cleavage, which include the N-terminal fragments (a, b, and c ions) and the C-terminal fragments (x, y, and z ions). Another is the derivatives of fragment ions that include the neutral loss of water or ammonia, multiple charged ions, and isotopic ions. The last is internal fragments and immonium ions generated from double backbone cleavage. The typical fragment ions in low energy CID are summarized in Table 1. The notations used in this paper are also listed in Table 1. Notice that b^i and y^i denote the derivative ions from b and y-ions. This is different from the conventional notation of b_i and y_i, which represents the b-ion and y-ion with i residues, respectively.

The fragment ions observed in an MS/MS spectrum have various intensities. Many are low and even below noise. It is therefore difficult to directly distinguish the fragment ions with low intensity from the contaminants and noise. However, the fragment ions occur correlatively with each other. This relationship between the fragment ions is helpful to correctly identify the fragment ions. The dependencies and correlations between types of fragment ions may be categorized into two classes. One is between the complementary fragments (such as b and y ions). The other is between fragments and their derivatives (such as b, b-NH_3, and b-H_2O, or y, y-NH_3, and y-H_2O). The relationship between the fragment ions can be examined via their statistical distributions. Table 2 lists the conditional probabilities calculated by examining the fragment ions in ion trap data sets. From the statistical results, we can clearly see the dependencies between different types of fragment ions. For example, b and y ions mostly occur together. The derivatives of fragment ions strongly depend on the fragment ions.

3.2. Regression model for scoring function

First, the peak intensities in the mass spectrum are normalized so that each peak has intensity between 0 and 1. Let p be the r-th highest peak in the spectrum, which r is refered to as the ranking of peak p. Then the normalized intensity of p is defined by $s(r) = (r_0 + 1)/(r_0 + r)$. The constant r_0 may be taken in the range $[50, 100]$.

Suppose a peptide $P = a_1 a_2 \dots a_n$. Each fragmentation between a_k and a_{k+1} is associated to a number of ions. The N-temrinal ions include the b-ion $a_1 a_2 \dots a_k$ and its derivative ions as in Table 1. The C-terminal ions include the y-ion $a_{k+1} \dots a_n$ and its derivative ions as in Table 1. We use the same notation b^i and y^i to denote both the derivative ions and the normalized intensity of the ions. In addition, there are internal fragment ions $a_i \dots a_k$ $(i = 2, \dots, k-1)$ and $a_{k+1} \dots a_j$ $(j = k+2, \dots, n-1)$ associated to the fragmentation. We sort these internal fragment ions according to their normalized intensities and denote the normalized intensities as $u^1, u^2, \dots$, from high to low. Thus for each fragmentation k we construct a score function via the following quadratic regression model:

$$f(k) = \sum_i \alpha_{1,i} y^i + \sum_i \beta_{1,i} b^i + \sum_i \gamma_{1,i} u^i + \sum_{i,j} \alpha_{2,ij} y^i y^j + \sum_{i,j} \beta_{2,ij} b^i b^j + \sum_{i,j} \gamma_{2,ij} y^i b^j \quad (1)$$

where α's, β's, and γ's are the regression coefficients, which are nonnegative and satisfy the following constraint,

$$\sum_i \alpha_{1,i} + \sum_i \beta_{1,i} + \sum_i \gamma_{1,i} + \sum_{i,j} \alpha_{2,ij} + \sum_{i,j} \beta_{2,ij} + \sum_{i,j} \gamma_{2,ij} = 1 \quad (2)$$

The last three terms are the quadratic regression part, which represents the dependencies between different ion types. If necessary, the model also allows to add the terms of triple regression.

Table 1. Fragment ions in low energy CID and notations. $m_c = M - m + 2$, where M is the precursor ion mass.

Fragments with N-terminus			Fragments with C-terminus		
Fragment type	Mass	Notation	Fragment type	Mass	Notation
b	m	b^0 or b	y	m_c	y^0 or y
b^{2+}	$(m+1)/2$	b^2	y^{2+}	$(m_c+1)/2$	y^2
b^{3+}	$(m+2)/3$	b^3	y^{3+}	$(m_c+2)/3$	y^3
b-NH_3	$m-17$	b^{17}	y-NH_3	m_c-17	y^{17}
b-H_2O	$m-18$	b^{18}	y-H_2O	m_c-18	y^{18}
b-2NH_3	$m-34$	b^{34}	y-2NH_3	m_c-34	y^{34}
b-NH_3-H_2O	$m-35$	b^{35}	y-NH_3-H_2O	m_c-35	y^{35}
b-2H_2O	$m-36$	b^{36}	y-2H_2O	m_c-36	y^{36}
a (b-CO)	$m-28$	b^{28}			

Table 2. Statistical probabilities of fragment ions in ion trap data set.

b^i :	b^2	b^3	b^{17}	b^{18}	b^{34}	b^{35}	b^{36}	b^{28}
P(b^i observed):	0.23	0.10	0.45	0.45	0.23	0.26	0.23	0.34
P(b^i observed | b observed):	0.24	0.11	0.58	0.59	0.29	0.33	0.29	0.46
y^i :	y^2	y^3	y^{17}	y^{18}	y^{34}	y^{35}	y^{36}	
P(y^i observed):	0.32	0.12	0.36	0.37	0.22	0.24	0.20	
P(y^i observed | y observed):	0.32	0.11	0.47	0.48	0.28	0.30	0.25	
P(b observed):	0.68							
P(b observed | y observed):	0.78							

In practice, we do not need to consider all combinations of all fragment ions. The regression model can be simplified according to the statistical characterization. In low energy CID, it is known that b and y-ions are dominant ion types. Moreover, for tryptic peptides, y-ions in general have stronger intensities than b-ions, and the derivatives of fragment ions strongly depend on the fragment ions. Taking this into account, we simplify the above model as follows:

$$f(k) = \sum_{i=0,2,3} \alpha_{1,i} y^i + \sum_{i=0,2,3} \beta_{1,i} b^i + \sum_{i\le 5} \gamma_{1,i} u^i + y \sum_{i\neq 0} \alpha_{2,i} y^i + b \sum_{i\neq 0} \beta_{2,i} b^i + y \sum_{i\neq 0} \gamma_{2,i} b^i \quad (3)$$

In this simplified model, the neutral loss of water or ammonia is not considered in the linear regression terms, and only the top five internal ions are used for each fragmentation. We also ignore the relationship between the derivative ions because their effects are too week. For clarity, we rewrite the above scoring model as

$$f(k) = X_k^T \cdot w \quad (4)$$

where $X_k = [y^i, b^i, u^i, yy^i, bb^i, yb^i]^T$ is a column vector associated to the fragmentation between a_k and a_{k+1}, $w = [\alpha_{1,i}, \beta_{1,i}, \gamma_{1,i}, \alpha_{2,i}, \beta_{2,i}, \gamma_{2,i}]^T$ is a column vector of the regression coefficients, and the superscription T stands for the transpose of a vector. Notice that because each i can take several different values, both X_k and w are 34-dimension vectors.

Let N_k be the number of unobserved b and y ions associated to the fragmentation k. N_k can be 0, 1, or 2. Introducing a penalty for the unobserved b and y ions, we further modify the scoring model as

$$f'(k) = f(k) - \mu N_k = X_k^T \cdot w - \mu N_k \quad (5)$$

where $0 \le \mu \le 1$ is a penalty coefficient. For a peptide P of n amino acids. The score of the spectrum S matched by the peptide P is calculated by

$$\text{score}(S,P) = \sum_{k=1}^{n-1} f'(k) = \sum_{k=1}^{n-1} X_k^T \cdot w - \mu \sum_{k=1}^{n-1} N_k \quad (6)$$

We train the regression coefficients by a linear programming. Suppose we have K mass spectra as training dataset. For each spectrum S_k, there are one positive peptide P_k, and L negative peptides P_{kl}, $l = 1, \ldots, L$. The linear programming formulation is

given as

$$\begin{aligned} &\max \sum_{k=1}^{K} e_k \qquad \text{subject to} \\ &\text{score}(S_k, P_k) - \text{score}(S_k, P_{kl}) \geq e_k, \\ &\sum w_i = 1, \\ &0 \leq w_i \leq 1, \\ &0 \leq \mu \leq 1, \\ &e_k \leq c \end{aligned} \tag{7}$$

This formulation is very similar to the linear programming used in Ref. 30. However, Ref. 30 concerned about the "database search" approach of MS/MS peptide identification, and therefore all the negative peptides are usually very different from the positive peptide. We concern about the de novo sequencing approach, where the negative peptides often differ from the positive peptide by only a few amino acids. As a result, the selected ion types and dependencies in the regression model are very different in the two approaches. Furthermore, we use the normalized intensities which depends on the ranking of a peak but not the actual signal intensity. This is a novel approach, which produces better accuracy. Before this work, we did not know whether a similar linear programming formulation as in Ref. 30 can work for improving de novo sequencing results.

The current PEAKS program first computes a y-ion matching score and a b-ion matching score at each mass value according to the peaks around it, and then efficiently computes thousands of amino acid sequences that maximize the total scores at the mass values of b-ions and y-ions [15]. These candidate sequences are then re-evaluated by a refined scoring function and the top scoring sequence is output. Here, we add another step to further use the regression model based scoring function to re-evaluate the 100 top-scoring sequences computed by PEAKS. We refer to this modified approach as PEAKS-RM.

4. EXPERIMENTS

In this section, we give experimental results to show that the regression model based scoring method can significantly improve the de novo sequencing accuracy over two existing high performance de novo programs: PEAKS and PepNovo.

The performance is measured using the ratio between the number of correctly predicted amino acids and the total length of the peptides. The ratio is refered to as the identification accuracy. Two types of the ratio are considered and defined as follows:

$$\text{(I)} \quad \frac{\text{number of correctly predicted amino acids}}{\text{number of amino acids in the real peptides}}$$

$$\text{(II)} \quad \frac{\text{number of correctly predicted amino acids}}{\text{number of amino acids in the prediction}}$$

An amino acid is correctly predicted if the amino acid appears at the same mass position of both the predicted and the real peptides. Given a test data set, the total length of real peptides is fixed. Therefore Type I accuracy only depends on the number of correctly predicted amino acids. However, because PepNovo only outputs partial sequences for some peptides, the number of the predicted amino acids may be significantly less than the total amino acids in the real peptides. Therefore, Type II accuracy may be very different from Type I accuracy. We note that software can increase Type II accuracy by missing the amino acids that do not appear in the MS/MS spectra, and only outputing the amino acids that are easy to be determined.

All the data sets used in our experiments are ion trap MS/MS data. The mass error in the ion trap data is around 0.5 dalton. Therefore, we do not make a distinction between the amino acids leucine and isoleucine (which have identical mass) and between lysine and glutamine (which have a small difference of 0.04 dalton in their masses).

Three ion trap datasets are used in the experiment. The training dataset contains 168 positive MS/MS spectra obtained from the first LC/MS/MS runs on "mixture A" as described in Ref. 34. The peptide sequences of these 168 MS/MS were all identified in Ref. 34. The two datasets used for testing are denoted by dataset 1 and dataset 2, respectively. Dataset 1 has 400 spectra provided by the authors of Ref. 17. 280 of the 400 spectra were used to compare PepNovo and other software in Ref. 17. Dataset 2 has 144 LCQ spectra. The three datasets were obtained in different labs with different protein mixtures.

Experimental results are given in Tables 3, 4 and 5. Table 3 shows the results for dataset 1. All the spectra in this dataset are doubly charged. Table 4

shows the results for dataset 2. This dataset contains singly, doubly and triply charged spectra. Because PepNovo's parameters were only trained for doubly charged spectra [17], we also list the results for the doubly charged spectra of dataset 2 in Table 5. By comparing with PEAKS and PepNovo algorithms, it is clear that our regression model based scoring function can significantly improve the de novo sequencing accuracy.

Table 3. Accuracies of PEAKS-RM, PEAKS, and PepNovo for dataset 1. (The average length of real peptides is 10.55.)

Algorithm	Type I	Type II	Average length
PEAKS-RM	**0.708**	**0.701**	10.66
PEAKS	0.655	0.665	10.38
PepNovo	0.652	0.697	9.87

Table 4. Accuracies of PEAKS-RM, PEAKS, and PepNovo for dataset 2. (The average length of real peptides is 11.82.)

Algorithm	Type I	Type II	Average length
PEAKS-RM	**0.639**	**0.638**	11.83
PEAKS	0.623	**0.638**	11.54
PepNovo	0.518	0.547	11.19

Table 5. Accuracies of PEAKS-RM, PEAKS, and PepNovo for only the doubly charged spectra of dataset 2. (The average length of real peptides is 12.085.)

Algorithm	Type I	Type II	Average length
PEAKS-RM	**0.666**	0.663	12.14
PEAKS	0.655	**0.667**	11.86
PepNovo	0.567	0.602	11.40

5. CONCLUSION AND DISCUSSION

This paper first proved that the de novo sequencing with internal fragment ions is NP-complete. This explains the reason that all existing polynomial time de novo sequencing algorithms could not use internal fragment ions. The paper then studied the statistical correlations between different ion types in ion trap MS/MS spectra; and proposed a regression model based scoring function for de novo sequencing, which incorporates the correlations between the fragment ion types. The experimental results showed that the regression model is a very effective scoring method in peptide de novo sequencing.

The authors also compared the regression models with and without internal fragment ions using our datasets. In the regression model without internal fragment ions, the coefficients were re-trained using the training data. The results showed that the inclusion of internal fragment ions improved the accuracy quite a bit in dataset 1 but only very slightly in datasets 2 and 3. The improvement mostly happens when there are some y and b ions missing for one peptide, and the internal fragment ions can then help to deduce the missing information. The experiments do not prove or disapprove that the consideration of internal fragment ions will *significantly* improve the peptide identification accuracy. This is because (a) the training and testing data were selected by currently available software that does not utilize internal fragment ions; (b) as illustrated by the NP-hardness result, there is no efficient algorithm (unless P=NP) to find the optimal solution with internal fragment ions, and the regression model is only a heuristic method. Consequently, the detailed comparison is omitted and the results in Section 3 should be purely regarded as a regression model instead of the discussion of internal fragment ions.

Acknowledgment

This research was undertaken, in part, thanks to funding from NSERC, PREA, and the Canada Research Chairs Program. The authors thank Dr. Kaizhong Zhang and Dr. Gilles Lajoie for discussions. The authors also thank the authors of Ref. 34 for providing the training dataset; Dr. Pavel Pevzner, Ari Frank for providing dataset 1 and PepNovo program; and Dr. Richard Johnson for providing dataset 2.

References

1. Snyder, A.P. Interpreting Protein Mass Spectra: A Comprehensive Resource, Oxford University Press. 2000.
2. Aebersold, R. and Mann, M. Mass spectrometry-based proteomics. *Nature* 2003 , 422, 198–207.
3. Johnsona, R.S. et al. Informatics for protein identification by mass spectrometry. *Methods* 2005, 35, 223–236.
4. Shadforth, I. et al. Protein and peptide identification algorithms using MS for use in high-throughput, automated pipelines. *Proteomics* 2005, 5, 4082–4095.

5. Zhang, N. et al. ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *Proteomics* 2005, 5, 4096–4106.
6. Xu, C. and Ma, B. Software for computational peptide identification from MS-MS data. *Drug Discovery Today*, July 2006.
7. Eng, J.K. et al. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Amer. Soc. Mass Spectrom.* 1994, 5, 976–989.
8. Perkins, D.N. et al. Probability-based protein identification by searching sequence database using mass spectrometry data. *Electrophoresis* 1999, 20, 3551–3567.
9. Craig, R. and Beavis, R.C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* 2003, 17, 2310–2316.
10. Geer, L.Y. et al. Open Mass Spectrometry Search Algorithm. *J. Proteome Research* 2004, 3, 958–964.
11. Taylor, J.A. and Johnson, R.S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 1997, 11, 1067–1075.
12. Taylor, J.A. and Johnson, R.S. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem.* 2001, 73, 2594–2604.
13. Dancik, V. et al. De novo protide sequencing via tandem mass-spectrometry. *J. Comp. Biology* 1999, 6, 327–342.
14. Chen, T. et al. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comp. Biology* 2001, 8, 325–337.
15. Ma, B. et al. PEAKS: powerful software for peptide de novo sequencing by MS/MS. Rapid Commun. *Mass Spectrom.* 2003,17, 2337–2342.
16. Ma,B. et al. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *Journal of Computer and System Sciences* 2005, 70, 418–430.
17. Frank, A. and Pevzner, P. Pepnovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* 2005, 77, 964–973.
18. Fischer, B. et al. NovoHMM: A Hidden Markov Model for de novo peptide sequencing. *Anal. Chem.* 2005, 77, 7265–7273.
19. Tabb, D. et al. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem* 2003, 75, 6415–6421.
20. Searle, B.C. et al. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS de novo sequencing results. *Anal. Chem.* 2004, 76, 2220–2230.
21. Han, Y. et al. SPIDER: Software for protein identification from sequence tags with de novo sequencing error. *Journal of Bioinformatics and Computational Biology* 2005, 3, 697–716.
22. Halligan, B. D. et al. DeNovoID: a web-based tool for identifying peptides from sequence and mass tags deduced from de novo peptide sequencing by mass spectroscopy. *Nucleic Acids Research* 2005, 33, 376–381.
23. Searle, B. C. Improving sensitivity by combining results from multiple search methodologies. Workshop Computational Proteomics and Mass Spectrometry 2004. Ohio State University, Ohio.
24. Rogers, I. Assessment of an amalgamative approach to protein identification, ASMS Conference on Mass Spectrometry 2005. San Antonio, Texas.
25. Fu, Y. et al. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* 2004, 20, 1948–1954.
26. Bafna, V. et al. A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 2001, 17, S13–S21.
27. Havilio, M. et al. Intensity-Based Statistical Scorer for Tandem Mass Spectrometry. *Anal. Chem.* 2003, 75, 435–44.
28. Colinge, J. et al. OLAV: Towards high-throughput tandem mass spectrometry data identification. *J. Proteomics* 2003, 3, 1454–63.
29. Elias, J. E. et al. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotechnol.* 2004, 22, 214–219.
30. J. Liu, B. Ma, M. Li: PRIMA: Peptide robust identification from MS/MS spectra. *APBC'05*, 181–190, 2005.
31. Pevzner, P.A. and Waterman, M.S. Open Combinatorial Problems in Computational Molecular Biology. Proceedings of the 3rd Israel Symposium on Thoery of Computing and Systems 1995, 158–173.
32. Pevzner, P.A. Computational Molecular Biology: An Algorithmic APproach. MIT Press. 2000.
33. Papadimitrou, C. and Yannakakis M. Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences* 1991, 43, 425–440.
34. Keller, A. et al. Experimental Protein Mixture for Validating Tandem Mass Spectra Analysis, OMICS 2002, 6(2), 207–212.