JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 8, Number 3, 2001
Mary Ann Liebert, Inc.
Pp. 325–337

A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry

TING CHEN,[1] MING-YANG KAO,[2] MATTHEW TEPEL,[1] JOHN RUSH,[1]
and GEORGE M. CHURCH[1]

ABSTRACT

Tandem mass spectrometry fragments a large number of molecules of the same peptide sequence into charged molecules of prefix and suffix peptide subsequences and then measures mass/charge ratios of these ions. The *de novo peptide sequencing* problem is to reconstruct the peptide sequence from a given tandem mass spectral data of k ions. By implicitly transforming the spectral data into an *NC-spectrum graph* $G = (V, E)$ where $|V| = 2k + 2$, we can solve this problem in $O(|V||E|)$ time and $O(|V|^2)$ space using dynamic programming. For an ideal noise-free spectrum with only b- and y-ions, we improve the algorithm to $O(|V| + |E|)$ time and $O(|V|)$ space. Our approach can be further used to discover a modified amino acid in $O(|V||E|)$ time. The algorithms have been implemented and tested on experimental data.

Key words: dynamic programming, peptide sequencing, mass spectrometry, computational proteomics, protein identification, computational biology.

1. INTRODUCTION

THE DETERMINATION OF THE AMINO ACID SEQUENCE of a protein is an important step toward quantifying this protein and solving its structure and function. Conventional sequencing methods (Wilkins *et al.*, 1997) cleave proteins into peptides and then sequence the peptides individually using Edman degradation or ladder sequencing by mass spectrometry or tandem mass spectrometry (McLafferty *et al.*, 1999). Among such methods, tandem mass spectrometry combined with high-performance liquid chromatography (HPLC) has been widely used as follows. A large number of molecules of the same but unknown peptide sequence are separated using HPLCs and a mass analyzer, such as a Finnigan LCQ ESI-MS/MS mass spectrometer. They are ionized and fragmented by collision-induced dissociation. All the resulting ions are measured by the mass spectrometer for mass/charge ratios. In the process of collision-induced dissociation, a peptide bond at a random position is broken, and each molecule is fragmented into two *complementary* ions, typically an N-terminal ion called *b-ion* and a C-terminal ion called *y-ion*.

Figure 1 shows the fragmentation of a doubly charged peptide sequence of n amino acids ($NHHCHR_1CO - \cdots - NHCHR_iCO - \cdots - NHCHR_nCOOH$). The ith peptide bond is broken and the peptide is fragmented into an N-terminal ion which corresponds to a charged prefix subsequence ($NHHCHR_1CO - \cdots -$

[1]Department of Genetics, Harvard Medical School, Boston, MA 02115.
[2]Department of Computer Science, Yale University, New Haven, CT 06520.

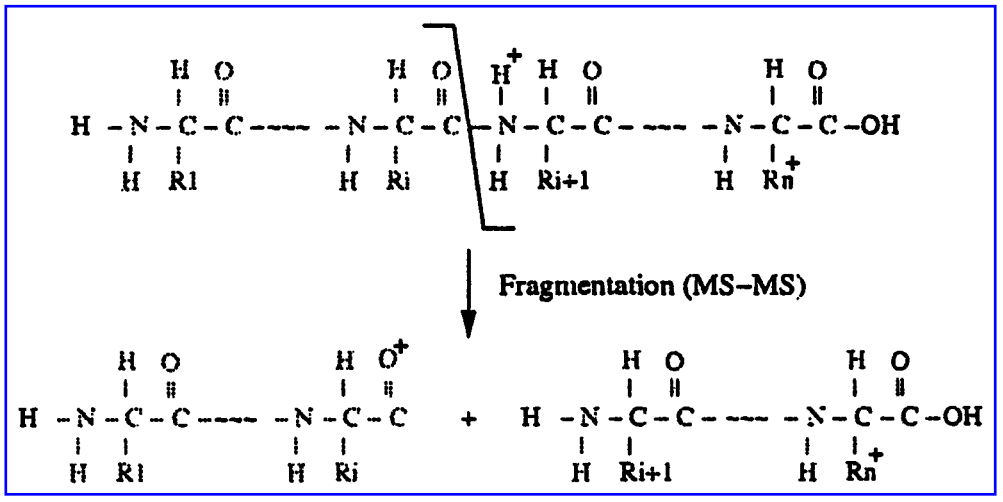

FIG. 1. A doubly charged peptide molecule is fragmented into a b-ion and a y-ion.

$NHCHR_iCO^+$) and a C-terminal ion which corresponds to a charged suffix subsequence ($NHHCHR_{i+1}CO - \cdots - NHCHR_n^+COOH$). These two ions are *complementary* because joining them determines the original peptide sequence. This dissociation process fragments a large number of molecules of the same peptide sequence, and ideally, the resulting ions contain all possible prefix subsequences and suffix subsequences. Table 1 shows all the resulting b-ions and y-ions from the dissociation of a peptide ($R_1 - R_2 - R_3$). These ions display a spectrum in the mass spectrometer, and each appears at the position of its mass because it carries a $+1$ charge. All the prefix (or suffix) subsequences form a sequence ladder where two adjacent sequences differ by one amino acid, and indeed, in the tandem mass spectrum, the mass difference between two adjacent b-ions (or y-ions) equals the mass of that amino acid. Figure 2 shows a hypothetical tandem mass spectrum of all the ions (including the parent ions) of a peptide SWR and the ladders formed by the b-ions and the y-ions.

We define an *ideal* tandem mass spectrum to be noise-free and containing only b- and y-ions, and every mass peak has the same height (or abundance). The interpretation of an ideal spectrum only deals with the following two factors: 1) it is unknown whether a mass peak (of some ion) corresponds to a prefix or a suffix subsequence; 2) some ions may be lost in the experiments and the corresponding mass peaks disappear in the spectrum. The *ideal de novo peptide sequencing problem* takes an input of a subset of prefix and suffix masses of an unknown target peptide sequence P and asks for a peptide sequence Q such that a subset of its prefixes and suffixes gives the same input masses. Note that, as expected, Q may or may not be the same as P, depending on the input data and the quality.

In practice, noise and other factors can affect a tandem mass spectrum. An ion may display two or three different mass peaks because of the distribution of two isotopic carbons, C^{12} and C^{13}, in the molecules. An ion may lose a water or an ammonia molecule and display a different mass peak from its normal one. The fragmentation may result in some other ion types such as a- and z-ions. Every mass peak displays a height that is proportional to the number of molecules of such an ion type. Therefore, the *de novo peptide sequencing problem* is, given a defined correlation function, to find a peptide sequence whose hypothetical prefix and suffix masses are optimally correlated to a tandem mass spectrum.

A special case of the peptide sequencing problem is the amino acid modification. An amino acid at an unknown location on the target peptide sequence is modified and its mass is changed. This modification

Table 1. Ionization and Fragmentation of Peptide ($R_1 - R_2 - R_3$)

B-ion sequences		*Y-ion sequences*	
b_1	$(R_1)^+$	y_2	$(R_2 - R_3)^+$
b_2	$(R_1 - R_2)^+$	y_1	$(R_3)^+$

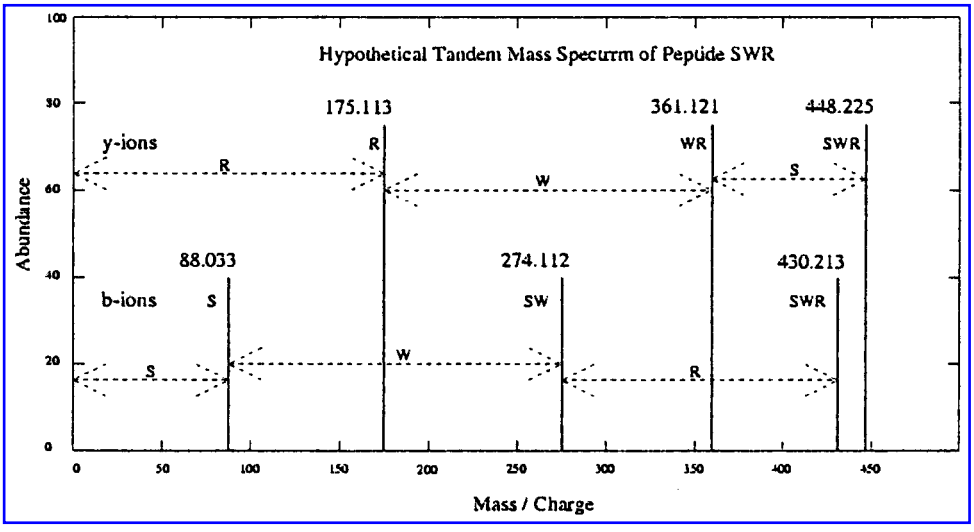

FIG. 2. Hypothetical tandem mass spectrum of peptide SWR.

appears in every molecule of this peptide, and all the ions containing the modified amino acid display different mass peaks from the unmodified ions. Finding this modified amino acid is of great interest in biology because modifications are usually associated with protein functions.

Several computer programs such as SEQUEST (Eng *et al.*, 1994), Mascot (Perkins *et al.*, 1999), and ProteinProspector (Clauser *et al.*, 1999), have been designed to interpret the tandem mass spectral data. A typical program like SEQUEST correlates peptide sequences in a protein database with the tandem mass spectrum. Peptide sequences in a database of over 300,000 proteins are converted into hypothetical tandem mass spectra, which are matched against the target spectrum using some correlation functions. The sequences with top correlation scores are reported. This approach gives an accurate identification, but cannot handle the peptides that are not in the database. Pruning techniques have been applied in some programs to screen the peptides before matching the database but at the cost of reduced accuracy.

An alternative approach (Dancik *et al.*, 1999 and Taylor and Johnson, 1997) is *de novo peptide sequencing*. Some candidate peptide sequences are extracted from the spectral data before they are validated in the database. First, the spectral data is transformed to a directed acyclic graph, called a *spectrum graph*, where 1) a node corresponds to a mass peak and an edge, labeled by some amino acids, connects two nodes that differ by the total mass of the amino acids in the label; 2) a mass peak is transformed into several nodes in the graph, and each node represents a possible prefix subsequence (ion) for the peak. Then, an algorithm is called to find the highest-scoring path in the graph or all paths with scores higher than some threshold. The concatenation of edge labels in a path gives one or multiple candidate peptide sequences. However, the well-known algorithms (Cormen *et al.*, 1990) for finding the longest path tend to include multiple nodes associated with the same mass peak. This interprets a mass peak with multiple ions of a peptide sequence, which is rare in practice. This paper provides efficient sequencing algorithms for a general interpretation of the data by restricting a path to contain at most one node for each mass peak.

For this purpose, we introduce the notion of an *NC-spectrum graph* $G = (V, E)$ for a given tandem mass spectrum, where $V = 2k+2$ and k is the number of mass peaks in the spectrum. In conjunction with this graph, we develop a dynamic programming approach to obtain the following results for previously open problems:

- The de novo peptide sequencing problem can be solved in $O(|V||E|)$ time and $O(|V|^2)$ space, and in $O(|V|+|E|)$ time and $O(|V|)$ space if the given spectrum is ideal.
- A modified amino acid can be found in $O(|V||E|)$ time.

Our paper is organized as follows. Section 2 formally defines the NC-spectrum graph and the peptide sequencing problem. Section 3 describes the dynamic programming algorithms for the peptide sequencing problem for three kinds of spectra: ideal spectra, noisy spectra, and spectra with a modified amino acid.

Section 4 reports the implementation and testing of our algorithms on experimental data. Section 5 mentions further research.

2. SPECTRUM GRAPHS AND THE PEPTIDE SEQUENCING PROBLEM

An amino acid unit in a peptide is called a *residue*. In forming the peptide bonds, an ionized amino acid molecule loses an oxygen and two hydrogens, so the mass of a residue is approximately 18 Daltons less than the mass of an ionized amino acid molecule. The structures of both molecules are shown in Figure 3. In this paper, we use the amino acid mass referring to the residue mass.

Given the mass W of a target peptide sequence P, k ions $I_1, \ldots, I_k$ of P, and the masses $w_1, \ldots, w_k$ of these ions, we create an *NC-spectrum graph* $G = (V, E)$ as follows.

For each I_j, it is unknown whether it is an N-terminal ion or a C-terminal ion. If I_j is a C-terminal ion, it has a complementary N-terminal ion, denoted as I_j^c, with a mass of $W - (w_j - 2)$, where the 2-Dalton mass is from the two extra hydrogens of the y-ion shown in Fig. 1. Therefore, we create a pair of nodes N_j and C_j to represent I_j and I_j^c, one of which must be an N-terminal ion. We also create two auxiliary nodes N_0 and C_0 to represent the zero mass and the total mass of all amino acids of P, respectively. Let $V = \{N_0, N_1, \ldots, N_k, C_0, C_1, \ldots, C_k\}$. Each node $x \in V$ is placed at a real line, and its coordinate $\text{cord}(x)$ is the total mass of its amino acids, i.e.,

$$\text{cord}(x) = \begin{cases} 0 & x = N_0; \\ W - 18 & x = C_0; \\ w_j - 1 & x = N_j \quad \text{for } j = 1, \ldots, k; \\ W - w_j + 1 & x = C_j \quad \text{for } j = 1, \ldots, k. \end{cases}$$

This coordinate scheme is adopted for the following reasons. An N-terminal b-ion has an extra hydrogen (approximately 1 Dalton), so $\text{cord}(N_j) = w_j - 1$ and $\text{cord}(C_j) = (W - (w_j - 2)) - 1 = W - w_j + 1$; and the full peptide sequence of P has two extra hydrogens and one extra oxygen (approximately 16 Daltons), so $\text{cord}(C_0) = W - 18$. If $\text{cord}(N_i) = \text{cord}(C_j)$ for some i and j, I_i, and I_j are complementary, one of them corresponds to a prefix sequence and another corresponds to the complementary suffix sequence. In the spectrum graph, they are merged into one pair of nodes. We say that N_j and C_j are *derived* from I_j. For convenience, for x and $y \in V$, if $\text{cord}(x) < \text{cord}(y)$, then we say $x < y$.

The edges of G are specified as follows. For x and $y \in V$, there is a directed edge from x to y, denoted by (x, y) and $E(x, y) = 1$, if the following conditions are satisfied: 1) x and y are not derived from the same I_j; 2) $x < y$; and 3) $\text{cord}(y) - \text{cord}(x)$ equals the total mass of some amino acids. Figure 4 shows a tandem mass spectrum and its corresponding NC-spectrum graph. In Figure 4, the path $N_0 - C_2 - N_1 - C_0$ that contains exactly one of every pair of complementary nodes derived from the same ion corresponds to the original peptide sequence SWR.

Since G is a directed graph along a line and all edges point to the right on the real line, we list the nodes from left to right according to their coordinates as $x_0, x_1, \ldots, x_k, y_k, \ldots, y_1, y_0$, where x_i and y_i, $1 \le i \le k$, are complementary. In practice, a tandem mass spectrum may contain noise such as mass peaks of other types of ions from the same peptide, mass peaks of ions from other peptides, and mass peaks of unknown ions. A general way to deal with these situations is to use a predefined edge (and node) scoring

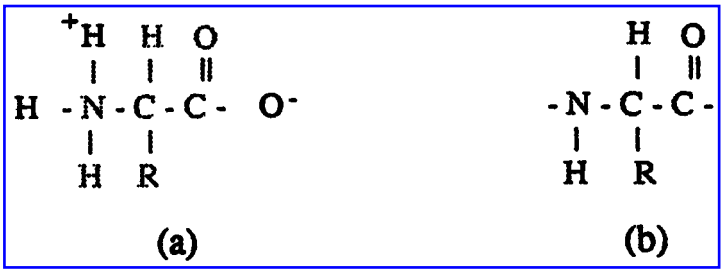

FIG. 3. (**a**) An ionized amino acid molecule and (**b**) a residue.

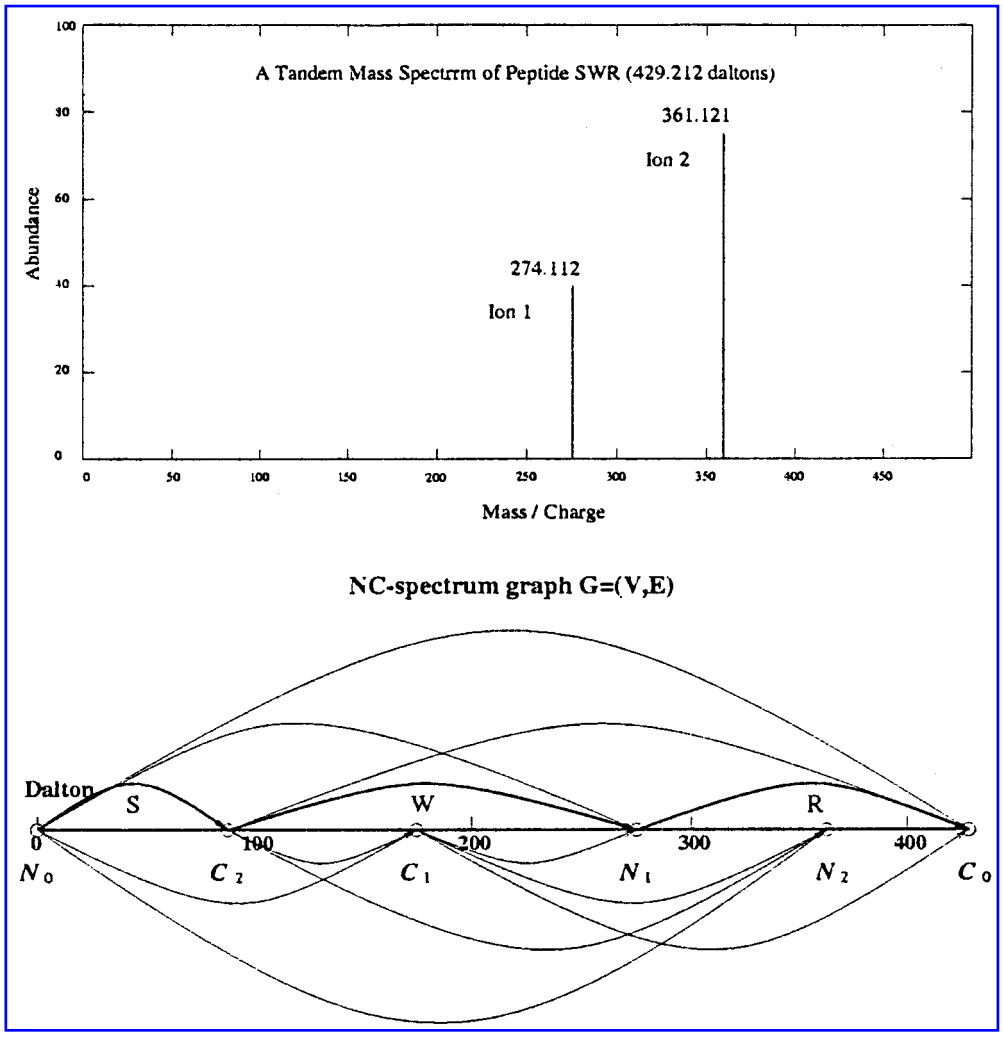

FIG. 4. A tandem mass spectrum and its corresponding NC-spectrum graph.

function $s(\cdot)$ such that nodes corresponding to high peaks and edges labeled with single amino acid receive higher scores. We define the score of a path to be the sum of the scores of the edges (and the nodes) on the path. Therefore, we have the following.

Definition 1. *The peptide sequencing problem is, given an NC-spectrum graph* $G = (V, E)$ *and an edge scoring function* $s(\cdot)$*, find a maximum score path from* x_0 *to* y_0*, such that at most one of* x_j *and* y_j *for every* $1 \leq j \leq k$ *is on the path.*

If the peptide sequence is known, we can identify the nodes of G corresponding to the prefix subsequences of this peptide. These nodes form a directed path from x_0 to y_0. Generally, the mass of a prefix subsequence does not equal the mass of any suffix subsequence, so the path contains at most one of x_j and y_j for each $j > 0$. On the other hand, a satisfying directed path from x_0 to y_0 contains observed prefix subsequences. If each edge on the path is labeled with some amino acids, we can visit the edges on the path from left to right and concatenate these amino acids to form one or multiple peptide sequences that display the tandem mass spectrum. If an appropriate scoring function is given, finding the maximum score path is equivalent to finding a peptide sequence that is optimally correlated to the spectrum.

Even if the mass of a prefix subsequence coincidently equals the mass of a suffix subsequence, which means the directed path contains both x_j and y_j, we can remove either x_j or y_j from the path and form

a new path corresponding to multiple peptide sequences which contain the real sequence. We call such a directed path a *feasible reconstruction* of P or *a feasible solution* of G.

To construct the edges of G, we use a mass array $\mathcal{A}$, which takes an input of mass m and returns 1 if m equals the total mass of some amino acids and 0 otherwise. Let h be the maximum mass under construction. Let δ be the measurement precision for mass. Then, we have the following.

Theorem 1. *Assume that we are given the maximum mass h and the mass precision δ.*

1. *The mass array $\mathcal{A}$ can be constructed in $O\left(\frac{h}{\delta}\right)$ time.*
2. *Given a spectrum of k mass peaks, G can be constructed in $O(k^2)$ time.*

Proof. These statements are proved as follows.

Statement 1. Given a mass m, $0 < m \leq h$, $\mathcal{A}[m] = 1$ if and only if m equals one amino acid mass or there exists an amino acid mass $r < m$ such that $\mathcal{A}[m - r] = 1$. If $\mathcal{A}$ is computed in the order from $\mathcal{A}[0]$ to $\mathcal{A}\left[\frac{h}{\delta}\right]$, each entry can be determined in constant time since there are only 20 amino acids and all the previous entries have been determined. The total time is $O\left(\frac{h}{\delta}\right)$.

Statement 2. For any two nodes v_i and v_j of G, we create an edge for v_i and v_j, $E(v_i, v_j) = 1$, if and only if $0 < \text{cord}(v_j) - \text{cord}(v_i) < h$ and $\mathcal{A}[\text{cord}(v_j) - \text{cord}(v_i)] = 1$. There are a total of $O(k^2)$ pairs of nodes. With $\mathcal{A}$, G can be constructed in $O(k^2)$ time. ■

In current practice, $\delta = 0.2$ Dalton, and $h = 400$ Daltons, roughly the total mass of four amino acids. The efficiency of our algorithm will allow biologists to consider much larger h and much smaller δ.

3. ALGORITHMS FOR PEPTIDE SEQUENCING

An ideal tandem mass spectrum is noise-free and contains only b- and y-ions, and every mass peak has the same height. This section starts with algorithms for ideal spectra in Section 3.1 and Section 3.2, and then describes algorithms for noisy spectra in Section 3.3 and spectra with a modified amino acid in Section 3.4.

3.1. Algorithm for ideal peptide sequencing

Given an ideal spectrum, we want to find a peptide sequence such that every mass peak of the spectrum matches with some b- or y-ion of the peptide. Therefore, we have the following.

Definition 2. *The ideal peptide sequencing problem is equivalent to the problem which, given $G = (V, E)$, asks for a directed path from x_0 to y_0 which contains exactly one of x_j and y_j for each $j > 0$.*

We list the nodes of G from left to right as $x_0, x_1, \ldots, x_k, y_k, \ldots, y_1, y_0$. Let $M(i, j)$ be a two-dimensional matrix with $0 \leq i, j \leq k$. Let $M(i, j) = 1$ if and only if in G there is a path L from x_0 to x_i and a path R from y_j to y_0, such that $L \cup R$ contains exactly one of x_p and y_p for every $p \in [1, i] \cup [1, j]$. Denote the two paths $L \cup R$ as the LR paths for $M(i, j) = 1$. Let $M(i, j) = 0$ otherwise. Table 2 shows the matrix M for the NC-spectrum graph in Figure 4.

Table 2. Matrix M for the NC-Spectrum Graph in Fig. 4

M	*0*	*1*	*2*
0	1	0	0
1	1	0	1
2	1	0	0

Algorithm Compute-M(G)
1. Initialize $M(0,0) = 1$ and $M(i, j) = 0$ for all $i \neq 0$ or $j \neq 0$;
2. Compute $M(1, 0)$ and $M(0, 1)$;
3. For $j = 2$ to k
4. For $i = 0$ to $j - 2$
   (a) if $M(i, j-1) = 1$ and $E(x_i, x_j) = 1$, then $M(j, j-1) = 1$;
   (b) if $M(i, j-1) = 1$ and $E(y_j, y_{j-1}) = 1$, then $M(i, j) = 1$;
   (c) if $M(j-1, i) = 1$ and $E(x_{j-1}, x_j) = 1$, then $M(j, i) = 1$;
   (d) if $M(j-1, i) = 1$ and $E(y_j, y_i) = 1$, then $M(j-1, j) = 1$.

Theorem 2. *The following statements hold.*

1. *Given $G = (V, E)$, Algorithm Compute-M computes the matrix M in $O(|V|^2)$ time.*
2. *Given $G = (V, E)$ and M, a feasible solution of G can be found in $O(|V|)$ time.*
3. *Given $G = (V, E)$, a feasible solution of G can be found in $O(|V|^2)$ time and $O(|V|^2)$ space.*
4. *Given $G = (V, E)$, all feasible solutions of G can be found in $O(|V|^2 + n|V|)$ time and $O(|V|^2 + n|V|)$ space, where n is the number of solutions.*

Proof. These statements are proved as follows.

Statement 1. Without loss of generality, assume that $i < j$ and $M(i, j) = 1$. By definition, either x_{j-1} or y_{j-1} (but not both) must be on the LR paths for $M(i, j) = 1$, and there exists a node y_p such that $E(y_j, y_p) = 1$ and $M(i, p) = 1$. Thus either $i = j - 1$ or $p = j - 1$, corresponding to Steps 4(b) and 4(d) respectively in the algorithm. A similar analysis holds for $M(j, i) = 1$ and $i < j$ in Steps 4(a) and 4(c). Therefore, every entry in M is correctly computed in the algorithm. Note that $|V| = 2k + 2$ and Steps 4(a), 4(b), 4(c), and 4(d) take $O(1)$ time, and thus the total time is $O(|V|^2)$.

Statement 2. Note that $|V| = 2k + 2$. Without loss of generality, assume that a feasible solution S contains node x_k. Then there exists some $j < k$, such that edge $(x_k, y_j) \in S$ and $M(k, j) = 1$. Therefore, we search the nonzero entries in the last row of M and find a j that satisfies both $M(k, j) = 1$ and $E(x_k, y_j) = 1$. This takes $O(|V|)$ time. With $M(k, j) = 1$, we backtrack M to search the next edge of S as follows. If $j = k-1$, the search starts from $i = k-2$ to 0 until both $E(x_i, x_k) = 1$ and $M(i, j) = 1$ are satisfied; otherwise $j < k - 1$, and then $E(x_{k-1}, x_k) = 1$ and $M(k-1, j) = 1$. We repeat this process to find every edge of S. A similar process holds for feasible solutions that contain node y_k. Using a common data structure such as link lists or a two-dimensional matrix, this algorithm visits every node of G at most once in the order form x_k to x_0 and from y_k to y_0 at a total cost of $O(|V|)$ time.

Statement 3. We compute M by means of Statement 1 and find a feasible solution by means of Statement 2. The total cost is $O(|V|^2)$ time and $O(|V|^2)$ space.

Statement 4. The proof is similar to that of Statement 2. For feasible solutions that contain node x_k, we search every j that satisfies both $M(k, j) = 1$ and $E(x_k, y_j) = 1$, and each j corresponds to different feasible solutions. For every $M(k, j) = 1$, we backtrack M to search the next edges as follows. If $j = k-1$, the search starts from $i = k - 2$ to 0 to find every i that satisfies both $E(x_i, x_k) = 1$ and $M(i, j) = 1$; otherwise $j < k - 1$, and then $E(x_{k-1}, x_k) = 1$ and $M(k-1, j) = 1$. Every edge found in this process corresponds to different feasible solutions. We repeat this process to find all feasible solutions that contain node x_k. A similar process holds for feasible solutions that contain node y_k. Finding one feasible solution costs $O(|V|)$ time and $O(|V|)$ space because the algorithm visits every node of G at most once for each solution. Computing M and finding n solutions cost $O(|V|^2 + n|V|)$ time and $O(|V|^2 + n|V|)$ space in total. ■

3.2. An improved algorithm for ideal peptide sequencing

To improve the time and space complexities in Theorem 2, we encode M into two linear arrays. Define an edge (x_i, y_j) with $0 \leq i, j \leq k$ to be a *cross edge* and an edge (x_i, x_j) or (y_j, y_i) with $0 \leq i < j \leq k$ to be an *inside edge*. Let lce(z) be the length of the longest consecutive inside edges starting from node z; i.e.,

$$\begin{cases} \text{lce}(x_i) = j - i & \text{if} \quad E(x_i, x_{i+1}) = \ldots = E(x_{j-1}, x_j) = 1 \text{ and } (j = k \text{ or } E(x_j, x_{j+1}) = 0); \\ \text{lce}(y_j) = j - i & \text{if} \quad E(y_j, y_{j-1}) = \ldots = E(y_{i+1}, y_i) = 1 \text{ and } (i = 0 \text{ or } E(y_i, y_{i-1}) = 0). \end{cases}$$

Let dia(z) be two diagonals in M, where

$$\begin{cases} \text{dia}(x_j) = M(j, j-1) & \text{for } 0 < j \le k; \\ \text{dia}(y_j) = M(j-1, j) & \text{for } 0 < j \le k; \\ \text{dia}(x_0) = \text{dia}(y_0) = 1. \end{cases}$$

Lemma 3. *Given* lce($\cdot$) *and* dia($\cdot$), *any entry of* M *can be computed in* $O(1)$ *time.*

Proof. Without loss of generality, let the $M(i, j)$ be the entry we want to compute where $0 \le i < j \le k$. If $i = j-1$, $M(i, j) = \text{dia}(y_j)$ as defined; otherwise $i < j-1$ and $M(i, j) = 1$ if and only if $M(i, i+1) = 1$ and $E(y_j, y_{j-1}) = \ldots = E(y_{i+2}, y_{i+1}) = 1$, which is equivalent to $\text{dia}(y_{i+1}) = 1$ and $\text{lce}(y_j) \ge j - i - 1$. Thus both cases can be solved in $O(1)$ time. ■

Lemma 4. *Given* $G = (V, E)$, lce($\cdot$) *and* dia($\cdot$) *can be computed in* $O(|V| + |E|)$ *time.*

Proof. We retrieve consecutive edges starting from y_k, $y_{k-1}, \ldots,$ until the first y_p with $p \le k$ and $\text{RE}(y_p, y_{p-1}) = 0$. Then we can fill $\text{lce}(y_k) = k - p$, $\text{lce}(y_{k-1}) = k - p - 1, \ldots,$ and $\text{lce}(y_p) = 0$ immediately. Next, we start a new retrieving and filling process from y_{p-1}, and repeat this until y_0 is visited. Eventually we retrieve $O(k)$ consecutive edges. A similar process can be applied to x. Using a common graph data structure such link lists, a consecutive edge can be retrieved in constant time, and thus lce($\cdot$) can be computed in $O(|V|)$ time.

By definition, $\text{dia}(x_j) = M(j, j-1) = 1$ if and only if there exists some i with $0 \le i < j - 1$, $M(i, j-1) = 1$ and $E(x_i, x_j) = 1$. If we have computed $\text{dia}(x_0), \ldots, \text{dia}(x_{j-1})$ and $\text{dia}(y_{j-1}), \ldots, \text{dia}(y_0)$, then $M(i, j-1)$ can be computed in constant time by means of the proof in Lemma 3. To find the x_i for $E(x_i, x_j) = 1$, we can visit every inside edge that ends at x_j. Thus $\text{dia}(x_j)$ can be computed and so can $\text{dia}(y_j)$. Therefore the computation of dia($\cdot$) visits every inside edge exactly once, and the total time is $O(|V| + |E|)$. ■

Theorem 5. *Assume that* $G = (V, E)$ *is given.*

1. *A feasible solution of* G *can be found in* $O(|V| + |E|)$ *time and* $O(|V|)$ *space.*
2. *All feasible solutions of* G *can be fond in* $O(n|V| + |E|)$ *time and* $O(n|V|)$ *space, where* n *is the number of solutions.*

Proof. These statements are proved as follows.

Statement 1. By Lemma 4, lce($\cdot$) and dia($\cdot$) can be computed in $O(|V| + |E|)$ time and $O(|V|)$ space. By Lemma 3, the last row and the last column of M can be reconstructed from lce($\cdot$) and dia($\cdot$) in $O(|V|)$ time. By Theorem 2 and Lemma 3, a feasible solution of G can be found in $O(|E|)$ time. Therefore, finding a feasible solution takes $O(|V| + |E|)$ time and $O(|V|)$ space.

Statement 2. The proof is similar to the proof of Statement 4 in Theorem 2. Finding an additional feasible solution takes $O(|V|)$ time and $O(|V|)$ space. Thus finding n solutions takes $O(n|V| + |E|)$ time and $O(n|V|)$ space. ■

A feasible solution of G is a path of $k + 1$ nodes and k edges, and therefore there must exist an edge between any two nodes on the path by the edge transitive relation. This implies that there are at least $(k + 1)k/2$ or $O(|V|^2)$ edges in the graph. However, in practice, a threshold is usually set for the maximum length (mass) of an edge, so the number of edges in G could be much smaller than $O(|V|^2)$ and may actually equal $O(|V|)$ sometimes. Thus, Theorem 5 actually finds a feasible solution in linear time for a sparse graph G.

3.3. Algorithm for peptide sequencing

In practice, a tandem mass spectrum contains noise and other types of ions. This section describes an algorithm for the peptide sequencing problem (Definition 1). We first compute an NC-spectrum graph G from this spectrum. Let $s(\cdot)$ be the edge scoring function for G. Let $Q(i, j)$ be a two-dimensional matrix

with $0 \le i, j \le k$. $Q(i, j) > 0$ if and only if in G, there is a path L from x_0 to x_i and a path R from y_j to y_0, such that at most one of x_p and y_p is in $L \cup R$ for every $p \in [1, i] \cup [1, j]$; $Q(i, j) = 0$ otherwise. If $Q(i, j) > 0$, $Q(i, j) = \max_{L,R}\{s(L) + s(R)\}$, the maximum score among all L and R pairs. Table 3 shows the matrix Q for the NC-spectrum graph in Figure 4 using a scoring function $s(e) = 1$ for every edge $e \in G$.

Algorithm Compute-Q(G)
1. Initialize $Q(i, j) = 0$ for all $0 \le i, j \le k$;
2. For $j = 1$ to k
3. If $E(y_j, y_0) = 1$, then $Q(0, j) = \max\{Q(0, j), s(y_j, y_0)\}$;
4. If $E(x_0, x_j) = 1$, then $Q(j, 0) = \max\{Q(j, 0), s(x_0, x_j)\}$;
5. For $i = 1$ to $j - 1$
   (a) For every $E(y_j, y_p) = 1$ and $Q(i, p) > 0$, $Q(i, j) = \max\{Q(i, j), Q(i, p) + s(y_j, y_p)\}$;
   (b) For every $E(x_p, x_j) = 1$ and $Q(p, i) > 0$, $Q(j, i) = \max\{Q(j, i), Q(p, i) + s(x_p, x_j)\}$.

Theorem 6. *The following statements hold.*

1. *Given $G = (V, E)$, Algorithm Compute-Q computes the matrix Q in $O(|V||E|)$ time.*
2. *Given $G = (V, E)$, a feasible solution of G can be found in $O(|V||E|)$ time and $O(|V|^2)$ space.*

Proof. These statements are proved as follows.

Statement 1. Let L and R be the maximum score paths that correspond to $Q(i, j) > 0$ for $i < j$. By definition, after removing node y_j from R, $L \cup R - \{y_j\}$ contains at most one of x_q and y_q for all $1 \le q \le j - 1$. Let $(y_j, y_p) \in R$ such that $Q(i, j) = Q(i, p) + s(y_j, y_p)$ corresponding to Steps 3 and 5(a) in the algorithm. A similar analysis holds for $Q(j, i) = 1$ and $i < j$ in Steps 4 and 5(b). The loop at Step 2 uses the previously computed maximum scores $Q(0, j-1), \ldots, Q(j-1, j-1)$ and $Q(j-1, 0), \ldots, Q(j-1, j-1)$ to fill up the maximum scores in $Q(0, j), \ldots, Q(j, j)$ and $Q(j, 0) \ldots, Q(j, j)$. Thus every entry in Q is correctly computed in a correct order. For every j, Steps 5(a) and 5(b) visit every edge of G at most once, so the total time is $O(|V||E|)$.

Statement 2. Algorithm Compute-Q computes Q in $O(|V||E|)$ time and $O(|V|^2)$ space. For every i and j, if $Q(i, j) > 0$ and $E(x_i, y_j) = 1$, we compute the sum $Q(i, j) + s(x_i, y_j)$. Let $Q(p, q) + s(x_p, y_q)$ be the maximum value, and we can backtrack $Q(p, q)$ to find all the edges of the feasible solution. The total cost is $O(|V||E|)$ time and $O(|V|^2)$ space. ■

3.4. Algorithm for one-amino acid modification

Amino acid modifications are related to protein functions. There are a few hundred known modifications. For example, some proteins are active when some amino acid is phosphorylated but inactive when it is dephosphorylated. In most experiments, a protein is digested into multiple peptides, and most peptides have at most one modified amino acid. This section discusses how to find one modified amino acid from a tandem mass spectrum. For the simplicity of our explanation, we assume that a given tandem mass spectrum is ideal. The methodology works for a noisy spectrum too.

We make two assumptions about the modification: 1) the modified mass is unknown and is not equal to the total mass of any number of amino acids; otherwise, it is information-theoretically impossible to detect an amino acid modification from tandem mass spectral data; 2) there is no feasible reconstruction for the given spectral data because a modification is rare if there is a feasible solution.

TABLE 3. MATRIX Q FOR THE NC-SPECTRUM GRAPH IN FIG. 4

Q	*0*	*1*	*2*
0	0	0	0
1	1	0	2
2	2	0	0

Definition 3. *The one-amino acid modification problem is equivalent to the problem which, given $G = (V, E)$, asks for two nodes v_i and v_j, such that $E(v_i, v_j) = 0$ but adding the edge (v_i, v_j) to G creates a feasible solution that contains this edge.*

Suppose the peptide sequence and the position of the modification are given. The modified mass can be determined by the difference between the experimentally measured peptide mass and the unmodified mass. Thus, in the NC-spectrum graph G, we can identify the nodes corresponding to the prefix subsequences, among which there is only one pair of adjacent nodes v_i and v_j, such that $E(v_i, v_j) = 0$ and node v_j contains the modified amino acid. By adding the edge (v_i, v_j) to G, these nodes form a directed path from x_0 to y_0. This path is a feasible solution.

On the contrary, suppose adding an edge (v_i, v_j) to G creates a feasible solution that contains this edge. Edge (v_i, v_j) is labeled by α indicating a modified amino acid. If each edge on the path corresponds to one amino acid, we can visit the edges on the path from left to right and concatenate these amino acids to form a peptide sequence that display the tandem mass spectrum. If some edge corresponds to multiple amino acids, we obtain more than one peptide sequence. With additional information, such as a protein database or a modification database, we can predict the original amino acid(s) for α.

Let $G = (V, E)$ be an NC-spectrum graph with nodes from left to right as $x_0, \ldots, x_k, y_k, \ldots, y_0$. Let $N(i, j)$ be a two-dimensional matrix with $0 \le i, j \le k$, where $N(i, j) = 1$ if and only if there is a path from x_i to y_j which contains exactly one of x_p and y_p for every $p \in [i, k] \cup [j, k]$. Let $N(i, j) = 0$ otherwise. Table 4 shows the matrix N for the NC-spectrum graph in Figure 4.

Algorithm Compute-N(G)

1. Initialize $N(i, j) = 0$ for all i and j;
2. Compute $N(k, k-1)$ and $N(k-1, k)$;
3. For $j = k - 2$ to 0
4. For $i = k$ to $j + 2$
   (a) if $N(i, j+1) = 1$ and $E(x_j, x_i) = 1$, then $N(j, j+1) = 1$;
   (b) if $N(i, j+1) = 1$ and $E(y_{j+1}, y_j) = 1$, then $N(i, j) = 1$;
   (c) if $N(j+1, i) = 1$ and $E(x_j, x_{j+1}) = 1$, then $N(j, i) = 1$;
   (d) if $N(j+1, i) = 1$ and $E(y_i, y_{j+1}) = 1$, then $N(j+1, j) = 1$.

Theorem 7. *The following statements hold.*

1. *Given $G = (V, E)$, Algorithm Compute-N computes the matrix N in $O(|V|^2)$ time.*
2. *Given $G = (V, E)$, all possible amino acid modifications can be found in $O(|V||E|)$ time and $O(|V|^2)$ space.*

Proof. These statements are proved as follows.

Statement 1. Let L and R be the paths that correspond to $N(i, j) = 1$ and $i > j$. By definition, after removing node y_j from R, $L \cup R - \{y_j\}$ contains exactly one of x_q and y_q for all $j + 1 \le q \le k$. Let $(y_p, y_j) \in R$, then $N(i, p) = 1$. Therefore, either $i = j + 1$ or $p = j + 1$, corresponding to Step 4(d) or 4(b) respectively in the algorithm. A similar analysis holds for $N(j, i) = 1$ and $i > j$ in Steps 4(a) and 4(c), and thus every entry in N is correctly computed in the algorithm. The loop at Step 3 uses previously computed $N(k, j+1), \ldots, N(j+1, j+1)$ and $N(j+1, k), \ldots, M(j+1, j+1)$ to fill up

Table 4. Matrix N for the NC-Spectrum Graph in Fig. 4

N	*2*	*1*	*0*
2	0	1	0
1	1	0	1
0	1	1	0

$N(k, j), \ldots, N(j, j)$ and $N(j, k), \ldots, N(j, j)$. Thus the algorithm computes N in a correct order. Note that $|V| = 2k + 2$ and Steps 4(a), 4(b), 4(c), and 4(d) take $O(1)$ time, and thus the total time is $O(|V|^2)$.

Statement 2. Let M and N be two matrices for G computed from Algorithm Compute-M and Algorithm Compute-N, respectively, at a total cost of $O(|V|^2)$ time and $O(|V|^2)$ space. Without loss of generality, let the modification be between two prefix nodes x_i and x_j with $0 \leq i < j \leq k$ and $E(x_i, x_j) = 0$. All the prefix nodes to the right of x_j have the same mass offset from the normal locations because the corresponding sequences contain the modified amino acid. By adding a new edge (x_i, x_j) to G, we create a feasible solution S that contains this edge: 1) If $i + 1 < j$, then $y_{i+1} \in S$, and thus $M(i, i + 1) = 1$ and $N(j, i + 1) = 1$. Finding all such x_i and x_j pairs takes $O(|V|^2)$ time because there are $O(k^2)$ possible combinations of i and j. 2) If $1 < i+1 = j < k$, then there exists an edge $(y_q, y_p) \in S$ and $q > j > i > p$, such that $E(y_q, y_p) = 1$ and $M(i, p) = 1$ and $N(j, q) = 1$. There are at most $O(|E|)$ edges that satisfy $E(y_q, y_p) = 1$, and checking $O(|V|)$ possible $i + 1 = j$ costs $O(|V||E|)$ time. 3) If $0 = i = j - 1$, then there exists an edge $(y_q, y_0) \in S$ and $q > j > i$, such that $E(y_q, y_0) = 1$ and $N(1, q) = 1$, which can be examined in $O(|V|)$ time. 4) If $i + 1 = j = k$, then there exists an edge $(x_k, y_p) \in S$ and $j > i > p$, such that $E(x_k, y_p) = 1$ and $M(k - 1, p) = 1$, which can be examined in $O(|V|)$ time. The case that the modification is between two prefix nodes x_k and y_j can be examined for $E(x_k, y_j) = 0$ and $M(k, j) = 1$ in $O(|V|)$ time. Thus the total complexity is $O(|V||E|)$ time and $O(|V|^2)$ space. ■

Note that the condition in Theorem 7 does not require that all ions in the spectrum are observed. If some ions are lost but their complementary ions appear, G still contains all prefix and suffix nodes of the target sequence. Furthermore, if G does not contain all prefix and suffix nodes because of many missing ions, this algorithm still finds the position of the modification but the result is affected by the quality of the data.

4. EXPERIMENTAL RESULTS

We have presented algorithms for reconstructing peptide sequences from tandem mass spectral data with noise and loss of ions. This section reports experimental studies which focus on cases of b-ions losing a water or ammonia molecule and cases of isotopic varieties for an ion. We treat the rare occurrence such as y-ions losing a water or ammonia molecule, b-ions losing two water or ammonia molecules, and other types of ions, as noise and apply Algorithm Compute-Q to reconstruct peptide sequences.

Isotopic ions come from isotopic carbons to C^{12} and C^{13}. An ion usually has a couple of isotopic forms, and the mass difference between two isotopic ions is generally one or two Daltons. Their abundance reflects the binomial distribution between C^{12} and C^{13}. This distribution can be used for identification. Isotopic ions can be merged to one ion of either the highest intensity or a new mass.

It is very common for a b-ion to lose a water or ammonia molecule. In the construction of an NC-spectrum graph, we add two types of edges when 1) the distance between two nodes equals the total mass of some amino acids plus the mass of one water molecule and 2) the distance between two nodes equals the total mass of some amino acids minus the mass of one water molecule. The first type includes the case that the distance equals the mass of exactly one water molecule. Therefore, a feasible path may contain edges of these two types, but the number of the first type of edges should equal the number of the second type edges, so the net number of water molecules on the path equals zero. The scording function for each edge is based on the abundance of two nodes and the error from a standard mass of some amino acids. We have implemented Algorithm Compute-Q and tested it on the data generated by the following process:

> The Chicken Ovalbumin proteins were digested with trypsin in 100 mM ammonium bicarbonate buffer pH 8 for 18 hours at 37°C. Then 100 $\mu\ell$ are injected in acetonitrile into a reverse phase HPLC interfaced with a Finnigan LCQ ESI-MS/MS mass spectrometer. A 1% to 50% acetonitrile 0.1%TFA linear gradient was executed over 60 minutes.

Figure 5 shows one of our prediction results. The ions labeled in the spectrum were identified successfully. We use a resolution of 1.0 Dalton and a relative-abundance threshold of 5.0 in our program.

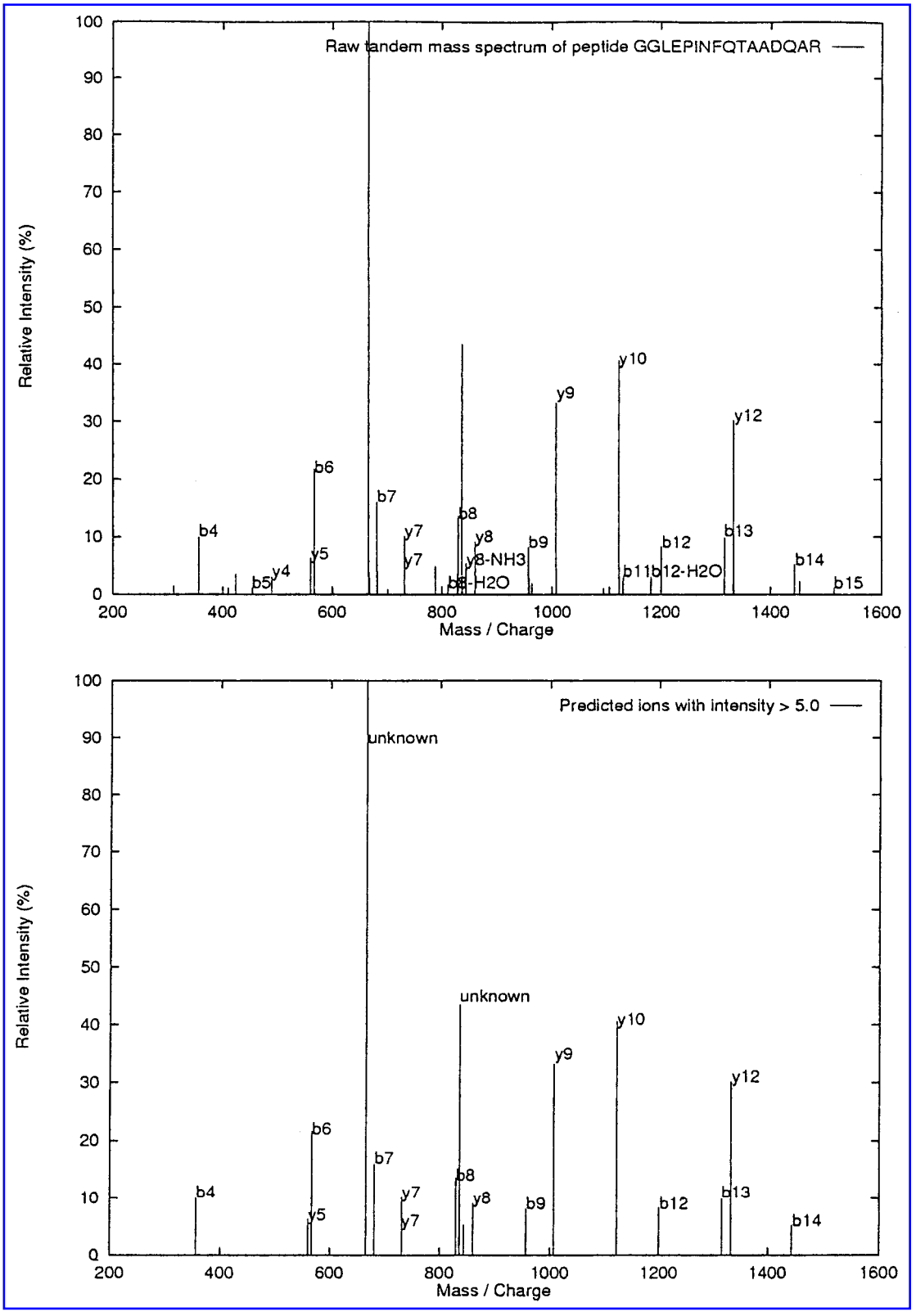

FIG. 5. Raw tandem mass spectrum and predicted ions of the Chicken Ovalbumin peptide GGLEPINFQTAADQAR.

5. FURTHER RESEARCH

We are working on a generalized scoring function which gives the best prediction, and the cases of multiple peptides.

REFERENCES

Clauser, K.R., Baker, P.R., and Burlingame, A.L. 1999. Role of accurate mass measurement (+/− 10ppm) in protein identification strategies employing MS or MS/MS. *Analytical Chem.* 71, 14, p. 2871.

Comen, T.H., Leiserson, C.E., and Rivest, R.L. 1990. *Introduction to Algorithms*, MIT Press, Cambridge, MA.

Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E., and Pevzner, P.A. 1999. De novo peptide sequencing via tandem mass spectrometry: A graph-theoretical approach. *J. Comp. Biol.* 6, 327–342.

Eng, J.K., McCormack, A.L., and Yates, J.R. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. for Mass Spectrometry* 5, 976–989.

McLafferty, F.W., Fridriksson, E.K., Horn, D.M., Lewis, M.A., and Zubarev, R.A. 1999. Biomolecule mass spectrometry. *Science* 284, 1289–1290.

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., and Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567.

Taylor, J.A., and Johnson, R.S. 1997. Sequence database searches via *de Novo* peptide sequencing by tandem mass spectrometry. Rapid Communications in Mass Spectrometry 11, 1067–1075.

Wilkins, M.R., Williams, K.L., Appel, R.D., and Hochstrasser, D.F. 1997. *Proteome Research: New Frontiers in Functional Genomics*, Springer-Verlag, New York.

Address correspondence to:
George M. Church
Department of Genetics
Harvard Medical School
Boston, MA 02115

E-mail: church@arep.med.harvard.edu

This article has been cited by:

1. Mark Alber, Adrian Buganza Tepole, William R. Cannon, Suvranu De, Salvador Dura-Bernal, Krishna Garikipati, George Karniadakis, William W. Lytton, Paris Perdikaris, Linda Petzold, Ellen Kuhl. 2019. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Medicine* **2**:1. . [Crossref]
2. Seungjin Na, Jihyung Kim, Eunok Paek. 2019. MODplus: Robust and Unrestrictive Identification of Post-Translational Modifications Using Mass Spectrometry. *Analytical Chemistry* **91**:17, 11324-11333. [Crossref]
3. Andrey Stavrianidi. 2019. A classification of liquid chromatography mass spectrometry techniques for evaluation of chemical composition and quality control of traditional medicines. *Journal of Chromatography A* 460501. [Crossref]
4. Fusong Ju, Jingwei Zhang, Dongbo Bu, Yan Li, Jinyu Zhou, Hui Wang, Yaojun Wang, Chuncui Huang, Shiwei Sun. 2019. De novo glycan structural identification from mass spectra using tree merging strategy. *Computational Biology and Chemistry* **80**, 217-224. [Crossref]
5. Fomin Eduard. 2019. A Simple Approach to the Reconstruction of a Set of Points from the Multiset of Pairwise Distances in n2 Steps for the Sequencing Problem: III. Noise Inputs for the Beltway Case. *Journal of Computational Biology* **26**:1, 68-75. [Abstract] [Full Text] [PDF] [PDF Plus]
6. Javid Ahmad Parray, Mohammad Yaseen Mir, Nowsheen Shameem. Advancement in Sustainable Agriculture: Computational and Bioinformatics Tools 465-547. [Crossref]
7. Yves Frank, Tomas Hruz, Thomas Tschager, Valentin Venzin. 2018. Improved de novo peptide sequencing using LC retention time information. *Algorithms for Molecular Biology* **13**:1. . [Crossref]
8. Thilo Muth, Bernhard Y Renard. 2018. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification?. *Briefings in Bioinformatics* **19**:5, 954-970. [Crossref]
9. Thilo Muth, Felix Hartkopf, Marc Vaudel, Bernhard Y. Renard. 2018. A Potential Golden Age to Come-Current Tools, Recent Use Cases, and Future Avenues for De Novo Sequencing in Proteomics. *PROTEOMICS* **18**:18, 1700150. [Crossref]
10. Ngoc Hieu Tran, Xianglilan Zhang, Ming Li. 2018. Deep Omics. *PROTEOMICS* **18**:2, 1700319. [Crossref]
11. Thomas Tschager, Simon Rösch, Ludovic Gillet, Peter Widmayer. 2017. A better scoring model for de novo peptide sequencing: the symmetric difference between explained and measured masses. *Algorithms for Molecular Biology* **12**:1. . [Crossref]
12. Hatem Loukil, Mohamed Tmar, Mahdi Louati, Afif Masmoudi, Faiez Gargouri. 2017. Impact of a priori MS/MS intensity distributions on database search for peptide identification. *Digital Signal Processing* **67**, 52-60. [Crossref]
13. Ngoc Hieu Tran, Xianglilan Zhang, Lei Xin, Baozhen Shan, Ming Li. 2017. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences* **114**:31, 8247-8252. [Crossref]
14. Han Hu, Kshitij Khatri, Joseph Zaia. 2017. Algorithms and design strategies towards automated glycoproteomics analysis. *Mass Spectrometry Reviews* **36**:4, 475-498. [Crossref]
15. Hao Yang, Hao Chi, Wen-Jing Zhou, Wen-Feng Zeng, Kun He, Chao Liu, Rui-Xiang Sun, Si-Min He. 2017. Open-pNovo: De Novo Peptide Sequencing with Thousands of Protein Modifications. *Journal of Proteome Research* **16**:2, 645-654. [Crossref]

16. Xiaoyan Guan, Naomi C. Brownstein, Nicolas L. Young, Alan G. Marshall. 2017. Ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry and tandem mass spectrometry for peptide de novo amino acid sequencing for a seven-protein mixture by paired single-residue transposed Lys-N and Lys-C digestion. *Rapid Communications in Mass Spectrometry* **31**:2, 207-217. [Crossref]
17. Sujata Baral, Swakkhar Shatabda, Mahmood A Rashid. CycloAnt 4-11. [Crossref]
18. Nathan J. Edwards. Protein Identification from Tandem Mass Spectra by Database Searching 357-380. [Crossref]
19. Fengchao Yu, Ning Li, Weichuan Yu. 2016. PIPI: PTM-Invariant Peptide Identification Using Coding Method. *Journal of Proteome Research* **15**:12, 4423-4435. [Crossref]
20. Fomin Eduard. 2016. A Simple Approach to the Reconstruction of a Set of Points from the Multiset of n2 Pairwise Distances in n2 Steps for the Sequencing Problem: II. Algorithm. *Journal of Computational Biology* **23**:12, 934-942. [Abstract] [Full Text] [PDF] [PDF Plus]
21. Vladimir Gorshkov, Stéphanie Yuki Kolbeck Hotta, Thiago Verano-Braga, Frank Kjeldsen. 2016. Peptide de novo sequencing of mixture tandem mass spectra. *PROTEOMICS* **16**:18, 2470-2479. [Crossref]
22. Thilo Muth, Erdmann Rapp, Frode S. Berven, Harald Barsnes, Marc Vaudel. Tandem Mass Spectrum Sequencing: An Alternative to Database Search Engines in Shotgun Proteomics 217-226. [Crossref]
23. Ludovic Gillet, Simon Rösch, Thomas Tschager, Peter Widmayer. A Better Scoring Model for De Novo Peptide Sequencing: The Symmetric Difference Between Explained and Measured Masses 185-196. [Crossref]
24. Sujun Li, Haixu Tang. Computational Methods in Mass Spectrometry-Based Proteomics 63-89. [Crossref]
25. Matúš Mihalák, Rastislav Šrámek, Peter Widmayer. 2016. Approximately Counting Approximately-Shortest Paths in Directed Acyclic Graphs. *Theory of Computing Systems* **58**:1, 45-59. [Crossref]
26. Bin Ma. Peptide De Novo Sequencing with MS/MS 1545-1547. [Crossref]
27. Bruce D. Pascal, Graham M. West, Catherina Scharager-Tapia, Ricardo Flefil, Tina Moroni, Pablo Martinez-Acedo, Patrick R. Griffin, Anthony C. Carvalloza. 2015. Software Analysis of Uncorrelated MS1 Peaks for Discovery of Post-Translational Modifications. *Journal of The American Society for Mass Spectrometry* **26**:12, 2133-2140. [Crossref]
28. Marco Blanco, Ralf Borndörfer, Michael Brückner, Nam Dũng Hoàng, Thomas Schlechte. 2015. On the Path Avoiding Forbidden Pairs Polytope. *Electronic Notes in Discrete Mathematics* **50**, 343-348. [Crossref]
29. T. Xu, S.K. Park, J.D. Venable, J.A. Wohlschlegel, J.K. Diedrich, D. Cociorva, B. Lu, L. Liao, J. Hewel, X. Han, C.C.L. Wong, B. Fonslow, C. Delahunty, Y. Gao, H. Shah, J.R. Yates. 2015. ProLuCID: An improved SEQUEST-like algorithm with enhanced sensitivity and specificity. *Journal of Proteomics* **129**, 16-24. [Crossref]
30. E. I. Berizovskaya, A. A. Ichalaynen, A. M. Antochin, V. F. Taranchenko, V. M. Goncharov, D. A. Mitrofanov, A. V. Udintsev, A. V. Aksenov, O. A. Shevlyakova, I. A. Rodin, O. A. Shpigun. 2015. Methods of processing mass spectrometry data to identify peptides and proteins. *Moscow University Chemistry Bulletin* **70**:5, 211-222. [Crossref]
31. Belal J. Muhialdin, Zaiton Hassan, Fatimah Abu Bakar, Hussein L. Algboory, Nazamid Saari. 2015. Novel Antifungal Peptides Produced by Leuconostoc mesenteroides DU15 Effectively Inhibit Growth of Aspergillus niger. *Journal of Food Science* **80**:5, M1026-M1030. [Crossref]

32. Yinglei Song, Albert Y. Chi. 2015. Peptide sequencing via graph path decomposition. *Information Sciences* **301**, 262-270. [Crossref]
33. Yinglei Song, Menghong Yu. 2015. On finding the longest antisymmetric path in directed acyclic graphs. *Information Processing Letters* **115**:2, 377-381. [Crossref]
34. Bin Ma. Peptide De Novo Sequencing with MS/MS 1-4. [Crossref]
35. Kyle K. Biggar, Kenneth B. Storey. 2014. New Approaches to Comparative and Animal Stress Biology Research in the Post-genomic Era: A Contextual Overview. *Computational and Structural Biotechnology Journal* **11**:19, 138-146. [Crossref]
36. Bjoern Titz, Ashraf Elamin, Florian Martin, Thomas Schneider, Sophie Dijon, Nikolai V. Ivanov, Julia Hoeng, Manuel C. Peitsch. 2014. Proteomics for systems toxicology. *Computational and Structural Biotechnology Journal* **11**:18, 73-90. [Crossref]
37. Yinglei Song. 2014. A New Parameterized Algorithm for Rapid Peptide Sequencing. *PLoS ONE* **9**:2, e87476. [Crossref]
38. Guangxu Jin, Stephen T.C. Wong. Proteomics-Based Theranostics 21-42. [Crossref]
39. Matúš Mihalák, Rastislav Šrámek, Peter Widmayer. Counting Approximately-Shortest Paths in Directed Acyclic Graphs 156-167. [Crossref]
40. Jingwen Yao, Shin-ichi Utsunomiya, Shigeki Kajihara, Tsuyoshi Tabata, Ken Aoshima, Yoshiya Oda, Koichi Tanaka. 2014. Peptide Peak Detection for Low Resolution MALDI-TOF Mass Spectrometry. *Mass Spectrometry* **3**:1, A0030-A0030. [Crossref]
41. Kyowon Jeong, Sangtae Kim, Pavel A. Pevzner. 2013. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* **29**:16, 1953-1962. [Crossref]
42. LIN HE, XI HAN, BIN MA. 2013. DE NOVO SEQUENCING WITH LIMITED NUMBER OF POST-TRANSLATIONAL MODIFICATIONS PER PEPTIDE. *Journal of Bioinformatics and Computational Biology* **11**:04, 1350007. [Crossref]
43. Jakub Kováč. 2013. Complexity of the path avoiding forbidden pairs problem revisited. *Discrete Applied Mathematics* **161**:10-11, 1506-1512. [Crossref]
44. Can Bruce, Kathryn Stone, Erol Gulcicek, Kenneth Williams. 2013. Proteomics and the Analysis of Proteomic Data: 2013 Overview of Current Protein-Profiling Technologies. *Current Protocols in Bioinformatics* **41**:1, 13.21.1-13.21.17. [Crossref]
45. Hao Chi, Haifeng Chen, Kun He, Long Wu, Bing Yang, Rui-Xiang Sun, Jianyun Liu, Wen-Feng Zeng, Chun-Qing Song, Si-Min He, Meng-Qiu Dong. 2013. pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra. *Journal of Proteome Research* **12**:2, 615-625. [Crossref]
46. Tobias Kind, Oliver Fiehn. Advances in structure elucidation of small molecules using mass spectrometry 129-166. [Crossref]
47. Susan K. Van Riper, Ebbing P. de Jong, John V. Carlis, Timothy J. Griffin. Mass Spectrometry-Based Proteomics: Basic Principles and Emerging Technologies and Directions 1-35. [Crossref]
48. Kyowon Jeong, Sangtae Kim, Pavel A. Pevzner. UniNovo : A Universal Tool for de Novo Peptide Sequencing 100-117. [Crossref]
49. KET FAH CHONG, HON WAI LEONG. 2012. TUTORIAL ON DE NOVO PEPTIDE SEQUENCING USING MS/MS MASS SPECTROMETRY. *Journal of Bioinformatics and Computational Biology* **10**:06, 1231002. [Crossref]
50. MohammadTaghi Hajiaghayi, Rohit Khandekar, Guy Kortsarz, Julián Mestre. 2012. The checkpoint problem. *Theoretical Computer Science* **452**, 88-99. [Crossref]

51. Swapnil Bhatia, Yong J. Kil, Beatrix Ueberheide, Brian T. Chait, Lemmuel Tayo, Lourdes Cruz, Bingwen Lu, John R. Yates, Marshall Bern. 2012. Constrained De Novo Sequencing of Conotoxins. *Journal of Proteome Research* **11**:8, 4191-4200. [Crossref]
52. Zhexue Wei, Daming Zhu. De Novo Peptide Sequencing Based on Vertex Contraction Algorithm 48-52. [Crossref]
53. S. Andreotti, G. W. Klau, K. Reinert. 2012. Antilope—A Lagrangian Relaxation Approach to the de novo Peptide Sequencing Problem. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9**:2, 385-394. [Crossref]
54. Bin Ma, Richard Johnson. 2012. De Novo Sequencing and Homology Searching. *Molecular & Cellular Proteomics* **11**:2, O111.014902. [Crossref]
55. Sylvain Bischof, Jonas Grossmann, Wilhelm Gruissem. Proteomics and its application in plant biotechnology 55-65. [Crossref]
56. Himani S. Ranasinghe, Arjan Scheepens, Ernest Sirimanne, Murray D. Mitchell, Christopher E. Williams, Mhoyra Fraser. 2012. Inhibition of MMP-9 Activity following Hypoxic Ischemia in the Developing Brain Using a Highly Specific Inhibitor. *Developmental Neuroscience* **34**:5, 417-427. [Crossref]
57. Jens Allmer. 2011. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Review of Proteomics* **8**:5, 645-657. [Crossref]
58. S Bocker, B Kehr, F Rasche. 2011. Determination of Glycan Structure from Tandem Mass Spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **8**:4, 976-986. [Crossref]
59. Jean-Charles Boisson, Laetitia Jourdan, El-Ghazali Talbi. 2011. Metaheuristics based de novo protein sequencing: A new approach. *Applied Soft Computing* **11**:2, 2271-2278. [Crossref]
60. Shenghui Zhang, Yaojun Wang, Dongbo Bu, Hong Zhang, Shiwei Sun. 2011. ProbPS: A new model for peak selection based on quantifying the dependence of the existence of derivative peaks on primary ion intensity. *BMC Bioinformatics* **12**:1, 346. [Crossref]
61. Yan Yan, Shenggui Zhang, Fang-Xiang Wu. 2011. Applications of graph theory in protein structure identification. *Proteome Science* **9**:Suppl 1, S17. [Crossref]
62. Nathan J. Edwards. Protein Identification from Tandem Mass Spectra by Database Searching 119-138. [Crossref]
63. Swapnil Bhatia, Yong J. Kil, Beatrix Ueberheide, Brian Chait, Lemmuel L. Tayo, Lourdes J. Cruz, Bingwen Lu, John R. Yates, Marshall Bern. Constrained De Novo Sequencing of Peptides with Application to Conotoxins 16-30. [Crossref]
64. Chongle Pan, Byung H Park, William H McDonald, Patricia A Carey, Jillian F Banfield, Nathan C VerBerkmoes, Robert L Hettich, Nagiza F Samatova. 2010. A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics* **11**:1. . [Crossref]
65. Tobias Kind, Oliver Fiehn. 2010. Advances in structure elucidation of small molecules using mass spectrometry. *Bioanalytical Reviews* **2**:1-4, 23-60. [Crossref]
66. Kyung-Hoon Kwon. 2010. Analytical methods for proteome data obtained from SDS-PAGE multi-dimensional separation and mass spectrometry. *Journal of Analytical Science and Technology* **1**:1, 1-14. [Crossref]
67. Christopher Hughes, Bin Ma, Gilles A. Lajoie. De Novo Sequencing Methods in Proteomics 105-121. [Crossref]
68. MohammadTaghi Hajiaghayi, Rohit Khandekar, Guy Kortsarz, Julián Mestre. The Checkpoint Problem 219-231. [Crossref]

69. Jingfen Zhang, Dong Xu, Wen Gao, Guohui Lin, Simin He. 2009. Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification. *Rapid Communications in Mass Spectrometry* **23**:21, 3448-3456. [Crossref]

70. Andreas Bertsch, Andreas Leinenbach, Anton Pervukhin, Markus Lubeck, Ralf Hartmer, Carsten Baessmann, Yasser Abbas Elnakady, Rolf Müller, Sebastian Böcker, Christian G. Huber, Oliver Kohlbacher. 2009. De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *ELECTROPHORESIS* **30**:21, 3736-3747. [Crossref]

71. Xiaowen Liu, Yonghua Han, Denis Yuen, Bin Ma. 2009. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics* **25**:17, 2174-2180. [Crossref]

72. Gaurav Kulkarni, Ananth Kalyanaraman, William R. Cannon, Douglas Baxter. A Scalable Parallel Approach for Peptide Identification from Large-Scale Mass Spectrometry Data 423-430. [Crossref]

73. Ritendra Datta, Marshall Bern. 2009. Spectrum Fusion: Using Multiple Mass Spectra for De Novo Peptide Sequencing. *Journal of Computational Biology* **16**:8, 1169-1182. [Abstract] [PDF] [PDF Plus]

74. Petr Kolman, Ondřej Pangrác. 2009. On the complexity of paths avoiding forbidden pairs. *Discrete Applied Mathematics* **157**:13, 2871-2876. [Crossref]

75. Changyong Yu, Guoren Wang, Yuhai Zhao, Keming Mao, Junjie Wu, Wendan Zhai. Generating Peptide Sequence Tags for Peptide Identification via Tandem Mass Spectrometry 200-207. [Crossref]

76. Enzhi Shen, Yan Lei, Qian Liu, Yanbo Zheng, Chunqing Song, Jan Marc, Yongchao Wang, Le Sun, Qianjin Liang. 2009. Identification and characterization of INMAP, a novel interphase nucleus and mitotic apparatus protein that is involved in spindle formation and cell cycle progression. *Experimental Cell Research* **315**:7, 1100-1116. [Crossref]

77. Sangtae Kim, Nitin Gupta, Nuno Bandeira, Pavel A. Pevzner. 2009. Spectral Dictionaries. *Molecular & Cellular Proteomics* **8**:1, 53-69. [Crossref]

78. Changyong Yu, Guoren Wang, Junjie Wu, Keming Mao. Classifying b and y Ions in Peptide Tandem Mass Spectra 37-41. [Crossref]

79. Sacha Baginsky. 2009. Plant proteomics: Concepts, applications, and novel strategies for data interpretation. *Mass Spectrometry Reviews* **28**:1, 93-120. [Crossref]

80. Jakub Kováč, Tomáš Vinař, Broňa Brejová. Predicting Gene Structures from Multiple RT-PCR Tests 181-193. [Crossref]

81. Guangxu Jin, Xiaobo Zhou, Honghui Wang, Stephen T. C. Wong. The Challenges in Blood Proteomic Biomarker Discovery 273-299. [Crossref]

82. Seungjin Na, Jaeho Jeong, Heejin Park, Kong-Joo Lee, Eunok Paek. 2008. Unrestrictive Identification of Multiple Post-translational Modifications from Tandem Mass Spectrometry Using an Error-tolerant Algorithm Based on an Extended Sequence Tag Approach. *Molecular & Cellular Proteomics* **7**:12, 2452-2463. [Crossref]

83. S. Bocker, F. Rasche. 2008. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* **24**:16, i49-i55. [Crossref]

84. N. Bandeira, J. V. Olsen, M. Mann, P. A. Pevzner. 2008. Multi-spectra peptide sequencing and its applications to multistage mass spectrometry. *Bioinformatics* **24**:13, i416-i423. [Crossref]

85. KANG NING, NAN YE, HON WAI LEONG. 2008. ON PREPROCESSING AND ANTISYMMETRY IN DE NOVO PEPTIDE SEQUENCING: IMPROVING EFFICIENCY AND ACCURACY. *Journal of Bioinformatics and Computational Biology* **06**:03, 467-492. [Crossref]

86. Yi Wei, Enzhi Shen, Na Zhao, Qian Liu, Jinling Fan, Jan Marc, Yongchao Wang, Le Sun, Qianjin Liang. 2008. Identification of a novel centrosomal protein CrpF46 involved in cell cycle progression and mitosis. *Experimental Cell Research* **314**:8, 1693-1707. [Crossref]
87. BAOZHEN SHAN, BIN MA, KAIZHONG ZHANG, GILLES LAJOIE. 2008. COMPLEXITIES AND ALGORITHMS FOR GLYCAN SEQUENCING USING TANDEM MASS SPECTROMETRY. *Journal of Bioinformatics and Computational Biology* **06**:01, 77-91. [Crossref]
88. MATTHEW T. OLSON, JONATHAN A. EPSTEIN, ALFRED L. YERGEY. De novo sequencing of peptides 195-201. [Crossref]
89. Jian Liu. Toward High-Throughput and Reliable Peptide Identification via MS/MS Spectra 333-344. [Crossref]
90. Bin Ma. Peptide De Novo Sequencing with MS/MS 640-642. [Crossref]
91. Jonas Grossmann, Bernd Fischer, Katja Baerenfaller, Judith Owiti, Joachim M. Buhmann, Wilhelm Gruissem, Sacha Baginsky. 2007. A workflow to increase the detection rate of proteins from unsequenced organisms in high-throughput proteomics experiments. *PROTEOMICS* **7**:23, 4245-4254. [Crossref]
92. Nuno Bandeira, Karl R. Clauser, Pavel A. Pevzner. 2007. Shotgun Protein Sequencing. *Molecular & Cellular Proteomics* **6**:7, 1123-1134. [Crossref]
93. N. Bandeira, D. Tsur, A. Frank, P. A. Pevzner. 2007. Protein identification by spectral networks analysis. *Proceedings of the National Academy of Sciences* **104**:15, 6140-6145. [Crossref]
94. CHUNGONG YU, YU LIN, SHIWEI SUN, JINJIN CAI, JINGFEN ZHANG, DONGBO BU, ZHUO ZHANG, RUNSHENG CHEN. 2007. AN ITERATIVE ALGORITHM TO QUANTIFY FACTORS INFLUENCING PEPTIDE FRAGMENTATION DURING TANDEM MASS SPECTROMETRY. *Journal of Bioinformatics and Computational Biology* **05**:02a, 297-311. [Crossref]
95. Peter A. DiMaggio, Christodoulos A. Floudas. De novo peptide identification via mixed-integer linear optimization and tandem mass spectrometry 989-994. [Crossref]
96. Peter A. DiMaggio,, Christodoulos A. Floudas. 2007. A mixed-integer optimization framework for de novo peptide identification. *AIChE Journal* **53**:1, 160-173. [Crossref]
97. Bianca Naumann, Michael Hippler. Insights into chloroplast proteomics: from basic principles to new horizons 371-407. [Crossref]
98. Jens Allmer, Bianca Naumann, Christine Markert, Monica Zhang, Michael Hippler. 2006. Mass spectrometric genomic data mining: Novel insights into bioenergetic pathways inChlamydomonas reinhardtii. *PROTEOMICS* **6**:23, 6207-6220. [Crossref]
99. Peter Schubert, Michael D. Hoffman, Matthew J. Sniatynski, Juergen Kast. 2006. Advances in the analysis of dynamic protein complexes by proteomics and data processing. *Analytical and Bioanalytical Chemistry* **386**:3, 482-493. [Crossref]
100. Matthew T. Olson, Jonathan A. Epstein, Alfred L. Yergey. 2006. De novo peptide sequencing using exhaustive enumeration of peptide composition. *Journal of the American Society for Mass Spectrometry* **17**:8, 1041-1049. [Crossref]
101. Changjiang Xu, Bin Ma. 2006. Software for computational peptide identification from MS–MS data. *Drug Discovery Today* **11**:13-14, 595-600. [Crossref]
102. Seung Yon Rhee, Julie Dickerson, Dong Xu. 2006. BIOINFORMATICS AND ITS APPLICATIONS IN PLANT BIOLOGY. *Annual Review of Plant Biology* **57**:1, 335-360. [Crossref]
103. Juris Meija. 2006. Mathematical tools in analytical mass spectrometry. *Analytical and Bioanalytical Chemistry* **385**:3, 486-499. [Crossref]

104. Alexey I. Nesvizhskii, Franz F. Roos, Jonas Grossmann, Mathijs Vogelzang, James S. Eddes, Wilhelm Gruissem, Sacha Baginsky, Ruedi Aebersold. 2006. Dynamic Spectrum Quality Assessment and Iterative Computational Analysis of Shotgun Proteomic Data. *Molecular & Cellular Proteomics* **5**:4, 652-670. [Crossref]

105. Marshall Bern, David Goldberg. 2006. De Novo Analysis of Peptide Tandem Mass Spectra by Spectral Graph Partitioning. *Journal of Computational Biology* **13**:2, 364-378. [Abstract] [PDF] [PDF Plus]

106. Pedro Larrañaga, Borja Calvo, Roberto Santana, Concha Bielza, Josu Galdiano, Iñaki Inza, José A. Lozano, Rubén Armañanzas, Guzmán Santafé, Aritz Pérez, Victor Robles. 2006. Machine learning in bioinformatics. *Briefings in Bioinformatics* **7**:1, 86-112. [Crossref]

107. Patricia Hernandez, Markus Müller, Ron D. Appel. 2006. Automated protein identification by tandem mass spectrometry: Issues and strategies. *Mass Spectrometry Reviews* **25**:2, 235-254. [Crossref]

108. JIAN LIU, BIN MA, MING LI. 2006. PRIMA: PEPTIDE ROBUST IDENTIFICATION FROM MS/MS SPECTRA. *Journal of Bioinformatics and Computational Biology* **04**:01, 125-138. [Crossref]

109. Xue Wu, Nathan Edwards, Chau-Wen Tseng. Peptide Identification via Tandem Mass Spectrometry 253-278. [Crossref]

110. Nuno Bandeira, Dekel Tsur, Ari Frank, Pavel Pevzner. A New Approach to Protein Identification 363-378. [Crossref]

111. Weichuan Yu, Baolin Wu, Tao Huang, Xiaoye Li, Kenneth Williams, Hongyu Zhao. Statistical Methods in Proteomics 623-638. [Crossref]

112. Ning Zhang, Xiao-jun Li, Mingliang Ye, Sheng Pan, Benno Schwikowski, Ruedi Aebersold. 2005. ProbIDtree: An automated software program capable of identifying multiple peptides from a single collision-induced dissociation spectrum collected by a tandem mass spectrometer. *PROTEOMICS* **5**:16, 4096-4106. [Crossref]

113. Christoph Borchers, Ting Chen, Nouri Neamati. Application of Proteomics in Basic Biological Sciences and Cancer 263-288. [Crossref]

114. Jingfen Zhang, Wen Gao, Jinjin Cai, Simin He, Rong Zeng, Runsheng Chen. 2005. Predicting Molecular Formulas of Fragment Ions with Isotope Patterns in Tandem Mass Spectra. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**:3, 217-230. [Crossref]

115. Christodoulos A. Floudas. 2005. Research challenges, opportunities and synergism in systems engineering and computational biology. *AIChE Journal* **51**:7, 1872-1884. [Crossref]

116. Bo Yan, You-Xing Qu, Feng-Lou Mao, Victor N. Olman, Ying Xu. 2005. PRIME: A Mass Spectrum Data Mining Tool for De Nova Sequencing and PTMs Identification. *Journal of Computer Science and Technology* **20**:4, 483-490. [Crossref]

117. D. Brent Weatherly, James A. Atwood, Todd A. Minning, Cameron Cavola, Rick L. Tarleton, Ron Orlando. 2005. A Heuristic Method for Assigning a False-discovery Rate for Protein Identifications from Mascot Database Search Results. *Molecular & Cellular Proteomics* **4**:6, 762-772. [Crossref]

118. YONGHUA HAN, BIN MA, KAIZHONG ZHANG. 2005. AN AUTOMATA APPROACH TO MATCH GAPPED SEQUENCE TAGS AGAINST PROTEIN DATABASE. *International Journal of Foundations of Computer Science* **16**:03, 487-497. [Crossref]

119. YONGHUA HAN, BIN MA, KAIZHONG ZHANG. 2005. SPIDER: SOFTWARE FOR PROTEIN IDENTIFICATION FROM SEQUENCE TAGS WITH DE NOVO SEQUENCING ERROR. *Journal of Bioinformatics and Computational Biology* **03**:03, 697-716. [Crossref]

120. Bin Ma, Kaizhong Zhang, Chengzhi Liang. 2005. An effective algorithm for peptide de novo sequencing from MS/MS spectra. *Journal of Computer and System Sciences* **70**:3, 418-430. [Crossref]
121. Hongying Zhong, Liang Li. 2005. An algorithm for interpretation of low-energy collision-induced dissociation product ion spectra forde novo sequencing of peptides. *Rapid Communications in Mass Spectrometry* **19**:8, 1084-1096. [Crossref]
122. Richard S. Johnson. Interpreting tandem mass spectra of peptides . [Crossref]
123. Christian Cole, Patrick J. Lester, Simon J. Hubbard. Mass spectrometric data mining for protein sequences . [Crossref]
124. B. Yan, C. Pan, V. N. Olman, R. L. Hettich, Y. Xu. 2005. A graph-theoretic approach for the separation of b and y ions in tandem mass spectra. *Bioinformatics* **21**:5, 563-574. [Crossref]
125. Marshall Bern, David Goldberg. EigenMS: De Novo Analysis of Peptide Tandem Mass Spectra by Spectral Graph Partitioning 357-372. [Crossref]
126. Yonghua Han, Bin Ma, Kaizhong Zhang. An Automata Approach to Match Gapped Sequence Tags Against Protein Database 167-177. [Crossref]
127. Baozhen Shan. Stochastic Context-Free Graph Grammars for Glycoprotein Modelling 247-258. [Crossref]
128. Ari Frank, Stephen Tanner, Pavel Pevzner. Peptide Sequence Tags for Fast Database Search in Mass-Spectrometry 326-341. [Crossref]
129. K. W. Lau, B. Stapley, S. Hubbard, H. Yin. Matching Peptide Sequences with Mass Spectra 390-397. [Crossref]
130. Jagath C Rajapakse, Kai-Bo Duan, Wee Kiang Yeo. 2005. Proteomic Cancer Classification with Mass Spectrometry Data. *American Journal of PharmacoGenomics* **5**:5, 281-292. [Crossref]
131. Yunhu Wan, Ting Chen. A Hidden Markov Model Based Scoring Function for Mass Spectrometry Database Search 342-356. [Crossref]
132. Simone Cristoni, Luigi Rossi Bernardi. 2004. Bioinformatics in mass spectrometry data analysis for proteomics studies. *Expert Review of Proteomics* **1**:4, 469-483. [Crossref]
133. Bingwen Lu, Ting Chen. 2004. Algorithms for de novo peptide sequencing using tandem mass spectrometry. *Drug Discovery Today: BIOSILICO* **2**:2, 85-90. [Crossref]
134. Alexey I Nesvizhskii, Ruedi Aebersold. 2004. Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug Discovery Today* **9**:4, 173-181. [Crossref]
135. Andreas Wilke, Christian Rückert, Daniela Bartels, Michael Dondrup, Alexander Goesmann, Andrea T. Hüser, Sebastian Kespohl, Burkhard Linke, Martina Mahne, Alice McHardy, Alfred Pühler, Folker Meyer. 2003. Bioinformatics support for high-throughput proteomics. *Journal of Biotechnology* **106**:2-3, 147-156. [Crossref]
136. Bin Ma, Kaizhong Zhang, Christopher Hendrie, Chengzhi Liang, Ming Li, Amanda Doherty-Kirby, Gilles Lajoie. 2003. PEAKS: powerful software for peptidede novo sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry* **17**:20, 2337-2342. [Crossref]
137. Wenzhu Zhang, Andrew N. Krutchinsky, Brian T. Chait. 2003. "De novo" peptide sequencing by MALDI-quadrupole-ion trap mass spectrometry: A preliminary study. *Journal of the American Society for Mass Spectrometry* **14**:9, 1012-1021. [Crossref]
138. Bingwen Lu, Ting Chen. 2003. A Suboptimal Algorithm for De Novo Peptide Sequencing via Tandem Mass Spectrometry. *Journal of Computational Biology* **10**:1, 1-12. [Abstract] [PDF] [PDF Plus]

139. Bin Ma, Kaizhong Zhang, Chengzhi Liang. An Effective Algorithm for the Peptide De Novo Sequencing from MS/MS Spectrum 266-277. [Crossref]

140. O. Lubeck, C. Sewell, Sheng Gu, Xian Chen, D.M. Cai. 2002. New computational approaches for de novo peptide sequencing from MS/MS experiments. *Proceedings of the IEEE* **90**:12, 1868-1874. [Crossref]

141. Ning Zhang, Ruedi Aebersold, Benno Schwikowski. 2002. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *PROTEOMICS* **2**:10, 1406-1412. [Crossref]

142. Ning Zhang, Ruedi Aebersold, Benno Schwikowski. 2002. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *PROTEOMICS* **2**:10, 1406-1412. [Crossref]

143. Ning Zhang, Ruedi Aebersold, Benno Schwikowski. 2002. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *PROTEOMICS* **2**:10, 1406-1412. [Crossref]

144. Alfred L. Yergey, Jens R. Coorssen, Peter S. Backlund, Paul S. Blank, Glen A. Humphrey, Joshua Zimmerberg, Jennifer M. Campbell, Marvin L. Vestal. 2002. De novo sequencing of peptides using MALDI/TOF-TOF. *Journal of the American Society for Mass Spectrometry* **13**:7, 784-791. [Crossref]

145. 2002. Current literature in mass spectrometry. *Journal of Mass Spectrometry* **37**:1, 119-132. [Crossref]

146. Ting Chen, Jacob D. Jaffe, George M. Church. 2001. Algorithms for Identifying Protein Cross-Links via Tandem Mass Spectrometry. *Journal of Computational Biology* **8**:6, 571-583. [Abstract] [PDF] [PDF Plus]

147. T. Fridman, R. Day, J. Razumovskaya, D. Xu, A. Gorin. Probability profiles-novel approach in tandem mass spectrometry De Novo sequencing 415-418. [Crossref]

148. Yonghua Han, Bin Ma, Kaizhong Zhang. SPIDER: software for protein identification from sequence tags with de novo sequencing error 198-207. [Crossref]

149. Bo Yan, Chongle Pan, V.N. Olman, R.L. Hettich, Ying Xu. Separation of ion types in tandem mass spectrometry data interpretation --a graph-theoretic approach 228-236. [Crossref]

150. A. Gorin, R.M. Day, A. Borziak, M.B. Strader, G.B. Hurst, T. Fridman. Probability profile method - new approach to data analysis in tandem mass spectrometry 479-482. [Crossref]

151. Nuno Bandeira, Julio Ng, Dario Meluzzi, Roger G. Linington, Pieter Dorrestein, Pavel A. Pevzner. De Novo Sequencing of Nonribosomal Peptides 181-195. [Crossref]

152. Ritendra Datta, Marshall Bern. Spectrum Fusion: Using Multiple Mass Spectra for De Novo Peptide Sequencing 140-153. [Crossref]

153. Yu Lin, Yantao Qiao, Shiwei Sun, Chungong Yu, Gongjin Dong, Dongbo Bu. A Fragmentation Event Model for Peptide Identification by Mass Spectrometry 154-166. [Crossref]

154. Matthew A. Goto, Eric J. Schwabe. A Dynamic Programming Algorithm for De Novo Peptide Sequencing with Variable Scoring 171-182. [Crossref]