JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 10, Number 1, 2003
© Mary Ann Liebert, Inc.
Pp. 1–12

A Suboptimal Algorithm for *De Novo* Peptide Sequencing via Tandem Mass Spectrometry

BINGWEN LU and TING CHEN

ABSTRACT

Tandem mass spectrometry has emerged to be one of the most powerful high-throughput techniques for protein identification. Tandem mass spectrometry selects and fragments peptides of interest into N-terminal ions and C-terminal ions, and it measures the mass/charge ratios of these ions. The *de novo* peptide sequencing problem is to derive the peptide sequences from given tandem mass spectral data of k ion peaks without searching against protein databases. By transforming the spectral data into a *matrix spectrum graph* $G = (V, E)$, where $|V| = O(k^2)$ and $|E| = O(k^3)$, we give the first polynomial time suboptimal algorithm that finds all the suboptimal solutions (peptides) in $O(p|E|)$ time, where p is the number of solutions. The algorithm has been implemented and tested on experimental data. The program is available at *http://hto-c.usc.edu:8000/msms/menu/denovo.htm.*

Key words: proteomics, mass spectrometry, suboptimal algorithms, dynamic programming, de novo peptide sequencing.

INTRODUCTION

PROTEIN IDENTIFICATION IS CENTRAL to many proteomics projects. Tandem mass spectrometry combined with high-performance liquid chromatography (HPLC) has been one of the most powerful techniques in protein analysis. A mixture of proteins is first digested into peptides by enzymes such as trypsin. Peptides of interest are then separated by HPLCs, ionized, and measured for mass/charge ratios by a mass analyzer such as a Finnigan LCQ ESI-MS/MS mass spectrometer. Peptides with a specific mass/charge ratio are selected and further fragmented by methods such as collision-induced dissociation (CID), and all of the resulting ions are measured again by a mass spectrometer for mass/charge ratios. In the CID process, one peptide bond for each peptide molecule is broken, and the peptide is fragmented into two ions, typically an N-terminal ion called *b-ion* and a C-terminal ion called *y-ion*.

For example, a doubly charged peptide, ($\mathrm{NH_2CHR_1CO} - \cdots - \mathrm{NHCHR_iCO} - \cdots - \mathrm{NHCHR_nCOOH}$), is selected for the fragmentation. If the ith peptide bond is broken, the resulting b-ion and y-ion are ($\mathrm{NH_2CHR_1CO} - \cdots - \mathrm{NHCHR_iCO^+}$) and ($\mathrm{NH_2CHR_{i+1}CO} - \cdots - \mathrm{NHCHR_n^+COOH}$), respectively. These two ions are *complementary* because the original peptide sequence can be determined by joining them. Ideally, this dissociation process may break any peptide bond of a molecule, so, given many molecules with the same peptide sequence, the resulting b-ions and y-ions contain ions of all possible prefix and suffix

Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089.

subsequences. These ions display a characteristic pattern in the mass spectrometry, called a tandem mass spectrum. A hypothetical tandem mass spectrum is shown in Fig. 1. The reader is referred to Chen *et al.* (2001) for a graphical illustration of the peptide fragmentation process. The goal of the current paper is to derive the original peptide sequence(s) for a given tandem mass spectrum.

The interpretation of a tandem mass spectrum has to deal with two main factors. First, it is unknown whether a mass peak corresponds to an N-terminal ion or a C-terminal ion. Second, some ions may not appear in the spectrum. In practice, noise and other factors have to be considered in the interpretation of a tandem mass spectrum. For example, an ion may correspond to two or three different mass peaks because of the distribution of isotopic carbons in the molecules; an ion may lose a water or an ammonia molecule and display at a mass/charge ratio other than its normal one; the fragmentation method may result in some other ion types such as a-ions and z-ions (a-ions and z-ions are less abundant types of N-terminal ions and C-terminal ions, respectively); or every mass peak may display at a different height that is proportional to the amount of molecules for the particular ion types.

Several computer programs, such as SEQUEST (Eng *et al.*, 1994), Mascot (Perkins *et al.*, 1999), and ProteinProspector (Clauser *et al.*, 1999), have been developed to interpret tandem mass spectral data by searching a protein database. A typical program like SEQUEST correlates peptide sequences in a protein database with a tandem mass spectrum. Every peptide sequence in the database is first converted into a hypothetical tandem mass spectrum, and then it is matched against the experimental spectrum using a correlation function. Sequences with top correlation scores are reported. Other scoring functions based on probabilistic models (Qin *et al.*, 1997; Dancik *et al.*, 2000; Bafna and Edwards, 2001) have also been proposed for comparing a tandem mass spectrum with a peptide sequence. In order to derive a good scoring function, environmental parameters, such as the probability of random noise, have to be identified. The database approach gives an accurate identification of peptides if they are in the database, but it cannot handle peptides that are not in the database.

When high-quality spectral data are given, the *de novo* peptide sequencing is another approach (Dancik *et al.*, 1999; Taylor and Johnson, 1997; Chen *et al.*, 2001). The *de novo* peptide sequencing problem is as follows: given a spectrum S and a defined scoring function $f()$, find a peptide sequence q to maximize $f(S|q)$. The *de novo* peptide sequencing is the only solution for applications such as finding novel proteins, studying proteome before genome, and amino acid mutations. A *de novo* peptide sequencing program can also be used as a pre-processing tool for tandem mass spectral data. Combining such a program with a database search program like BLAST and FASTA, we can extract partial sequences directly from the spectral data and then validate them in the databases for complete protein sequences.

A dynamic programming algorithm for *de novo* peptide sequencing was proposed by Chen *et al.* (2001). The algorithm first transforms a spectrum into a *spectrum graph*, in which (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; and (2) a mass peak is transformed into a pair of N-terminal nodes in the graph, representing two possible but mutually exclusive assumptions of this mass peak: an N-terminal ion

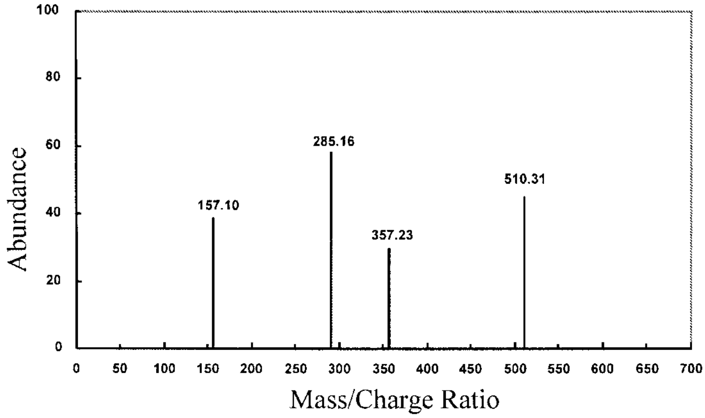

FIG. 1. A tandem mass spectrum of the peptide RQPKL(622.39 Daltons).

or a C-terminal ion. Then, the dynamic programming algorithm is called to find the highest-scoring path in the graph that goes through every pair of nodes corresponding to the same mass peak at most once. In its solution, the algorithm interprets every mass peak mutual-exclusively to an N-terminal ion, a C-terminal ion, or an uninterpretable ion. The concatenation of the amino acids of the edge labels in the path gives one or multiple candidate peptide sequences.

The focus of this paper is on finding suboptimal solutions (peptides) for the ***de novo*** peptide sequencing problem. The dynamic programming algorithm finds the optimal solution only, but the optimal solution may not be the real sequence that produces the spectrum in the experiment. Even database search programs sometimes report several sequences with similar scores. One reason is that the scoring function could misinterpret the spectral data because the physical and chemical processes that occur during the peptide fragmentation are not completely known. Noise and unknown ions may be interpreted as real ions by the programs. For these reasons, the suboptimal solutions are of great interest, especially when the real sequence cannot be clearly determined by the optimal solution.

In this paper, we introduce the notion of a ***matrix spectrum graph*** $G = (V, E)$ for a given tandem mass spectrum of k mass peaks, where $V = O(k^2)$ and $|E| = O(k^3)$. In conjunction with this graph, we develop a tree search algorithm to find all suboptimal solutions in $O(p|E|)$ time, where p is the number of solutions.

METHODS

Spectrum graph and dynamic programming algorithm

Details of the construction of a spectrum graph and the dynamic programming algorithms are given by Chen ***et al.*** (2001). We will now summarize the ideas of that paper in the following.

Construction of a spectrum graph. Tandem mass spectrometry measures mass/charge ratios of selected peptides and then measures their fragmented ions (see Fig. 1). Assume that the charges are known and the masses can be derived. Assume that an unknown peptide q has molecular weight W (uncharged) and k fragmented ions $I_1, \ldots, I_k$ with masses $w_1, \ldots, w_k$, respectively. A ***spectrum graph*** $G_s = (V_s, E_s)$ is created as follows.

Let $m = 2k + 1$. We first create two nodes, z_0 and z_m, on a line to represent the zero mass and the total residue mass, $W - 18$, of q, respectively. The 18 daltons are for the two extra hydrogens and one extra oxygen in q, besides the residues. All other nodes are created on the line between z_0 and z_m such that their distances to z_0 correspond to the associated masses. For each I_j, because it is unknown whether it is a b-ion or a y-ion, we create a pair of nodes, z_j and z_{m-j}, placed at the mass of $w_j - 1$ and $W - (w_j - 2)$, respectively, to represent two mutually exclusive assumptions: (1) I_j is a b-ion, and z_j represents the node with the residue mass of this b-ion; and (2) I_j is a y-ion, and z_{m-j} represents the node with the residue mass of its complementary b-ion. If this ion is real, either z_j or z_{m-j}, but not both, represents the real b-ion.

The edges of the spectrum graph G_s always point from the lower mass nodes to the higher mass nodes. If the mass difference between two nodes z_i and z_j equals the total mass of some amino acid residues, we draw a directed edge between z_i and z_j, pointing from the low-mass node to the high-mass node. Thus, the spectrum graph G_s is a directed acyclic graph along a line, and all edges point to the right on the real line. See Fig. 2 for an example of a spectrum graph constructed based on the spectrum shown in Fig. 1.

Let $f(\cdot)$ be a pre-defined edge (and node) scoring function. Then the peptide sequencing problem can be defined as follows: given a spectrum graph $G_s = (V_s, E_s)$ and an edge scoring function $f(\cdot)$, how can we find a maximum score path from z_0 to z_m, such that at most one of z_j and z_{m-j} for every $1 \leq j \leq k$ is on the path.

Dynamic programming algorithm. The dynamic programming algorithm, Algorithm Compute-Q, in Chen ***et al.*** (2001) solves the peptide sequencing problem. In summary, the nodes of G_s are first renamed in an order from left to right as $x_0, x_1, \ldots, x_k, y_k, \ldots, y_1, y_0$, where every pair, x_i and y_i, $1 \leq i \leq k$, corresponds to two mutually exclusive assumptions of the same mass peak. A matrix $Q(i, j)$ is used to store the maximum path score from x_0 to y_0 that contains the edge (x_i, y_j), $i \neq j$. The dynamic programming

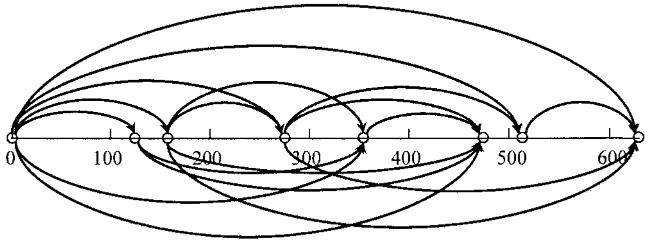

FIG. 2. An NC-spectrum graph, constructed from the spectrum shown in Fig. 1.

algorithm computes Q from lower indexes to higher indexes. Without loss of generality, assume $i < j$, then $Q(i, j) = \max\{Q(i, l) + f(y_j, y_l),\ l = 1, \cdots, j-1\}$. All elements of Q can be calculated by this recursion. The maximum score path can be found by tracing the elements of Q. In the following section, we develop an algorithm to find all suboptimal solutions, rather than just the optimal solution.

Matrix spectrum graph

Construction of a matrix spectrum graph. Let $x_0, x_1, \ldots, x_k, y_k, \ldots, y_1, y_0$ be the nodes of the spectrum graph $G_s = (V_s, E_s)$ listed in the order from left to right. Let $f(\cdot) > 0$ be the edge weight function of G_s. Define $X = \{x_0, x_1, \ldots, x_k\}$ and $Y = \{y_0, y_1, \ldots, y_k\}$. Define a directed weighted *matrix spectrum graph* as $G = (V, E)$, where $V \subseteq X \times Y$ and $E \subseteq V \times V$. Define a node $< x_i, y_j > \in V$, $i \neq j$, as v_{ij}. For any i, $v_{ii} \notin V$. There are two types of edges in E:

1. $(v_{ij}, v_{im}) \in E$ if $m > i$ and $m > j$ and $(y_m, y_j) \in E_s$. The edge function $f(v_{ij}, v_{im}) = f(y_m, y_j)$.
2. $(v_{ij}, v_{mj}) \in E$ if $m > i$ and $m > j$ and $(x_i, x_m) \in E_s$. The edge function $f(v_{ij}, v_{mj}) = f(x_i, x_m)$.

Obviously, G is a *directed acyclic graph* with $|V| = O(k^2)$ nodes and $|E| = O(k^3)$ edges because each node has at most $O(k)$ outgoing edges. An example of matrix spectrum for the spectrum graph in Fig. 2 is shown in Fig. 3.

Feasible path. Let v_{00} be the starting node of G. Let T be the set of *terminal nodes.* Each terminal node $v_{ij} \in T$ satisfies that $(x_i, y_j) \in E_s$ and v_{ij} has no outgoing edges in E. The terminal nodes in the example matrix spectrum graph are colored grey in Fig. 3. A *feasible* path for G is a path that starts from v_{00} and ends at a terminal node. Two feasible paths of the example of the matrix spectrum graph are

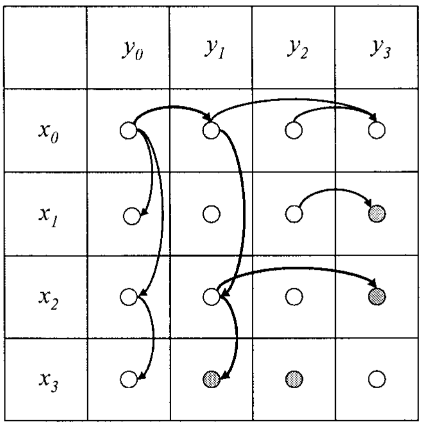

FIG. 3. An example of a matrix spectrum graph, constructed from the spectrum graph shown in Fig. 2. The terminal nodes are colored gray. The two feasible paths are bolded.

shown by bold lines. A ***feasible*** path (or solution) for G_s is defined as a path from x_0 to y_0 that goes through every pair of x_l and y_l, $1 \leq l \leq k$, at most once (Chen ***et al.***, 2001). Obviously, every feasible path for G_s can be mapped to a unique feasible path in G, and vice versa. Therefore, there is a one-to-one mapping between the feasible paths for G and the feasible paths for G_s. Since every feasible path for G_s corresponds to a unique suboptimal solution to the ***de novo*** peptide sequencing problem, we can find every feasible path for G to obtain all suboptimal solutions.

Suboptimal solution. Assume that P is the maximum score path for G, which can be found by a depth-first-search (DFS) algorithm. Let $f_{max} = f(P)$. Given a ratio α, $0 < \alpha \leq 1$, if a feasible path Q satisfies $f(Q) \geq \alpha \cdot f_{max}$, then Q is a ***α-suboptimal feasible path*** for G—or we say that Q is a ***suboptimal solution***. Therefore, the suboptimal ***de novo*** peptide sequencing problem is this: given a matrix spectrum graph G, find all α-suboptimal feasible paths.

Suboptimal algorithm

Define $l(v_{ij})$ to be the maximum path score among all paths between v_{00} and v_{ij}. Similarly, let $r(v_{ij})$ be the maximum path score among all paths between v_{ij} and the terminal nodes. If no path exists, set both $l-$ and $r-$ scores to be negative infinite. Let O_{ij} be the set of the outgoing edges of v_{ij}. The suboptimal algorithm consists of the following steps.

Step 1: Constructing the matrix spectrum graph. Given a tandem mass spectrum of k mass peaks, the spectrum graph $G_s = (V_s, E_s)$ can be constructed in $O(k^2)$ time, and the matrix spectrum graph $G = (V, E)$ can be directly constructed from G_s in $O(k^3)$ time because $|E| = O(k^3)$.

Step 2: Computing $l(\cdot)$ and $r(\cdot)$. The computation of $l(\cdot)$ is similar to the algorithm of finding the ***single-source shortest paths*** in a directed acyclic graph that can be solved in $O(k^3)$ time (Cormen ***et al.***, 2000). First we topologically sort all the nodes of G and compute $l(u) = f(v_{00}, u)$ for every edge $(v_{00}, u) \in E$. Then, for each vertex $u \in V$ taken in the topologically sorted order, we update $l(w) = \max\{l(w), l(u) + f(u, w)\}$ for all $(u, w) \in E$. Thus, all l-scores can be computed in $O(k^3)$ time. The computation of $r(\cdot)$ is similar to this algorithm except that the scores are computed in the reversed topologically sorted order. The total time is $O(k^3)$. The maximum path score f_{max} of G equals the maximum l-score. The $l(\cdot)$ and $r(\cdot)$ values of the example matrix spectrum graph (Fig. 3) are shown in Tables 1 and 2, respectively.

TABLE 1. THE $l(\cdot)$ VALUES COMPUTED ACCORDING TO THE MATRIX SPECTRUM GRAPH SHOWN IN FIG. 3

l()	*0*	*1*	*2*	*3*
0	0	1	$-\infty$	2
1	1	$-\infty$	$-\infty$	$-\infty$
2	1	2	$-\infty$	3
3	2	3	$-\infty$	$-\infty$

TABLE 2. THE $r(\cdot)$ VALUES COMPUTED ACCORDING TO THE MATRIX SPECTRUM GRAPH SHOWN IN FIG. 3

r()	*0*	*1*	*2*	*3*
0	3	2	$-\infty$	$-\infty$
1	$-\infty$	$-\infty$	1	0
2	$-\infty$	1	$-\infty$	0
3	$-\infty$	0	0	$-\infty$

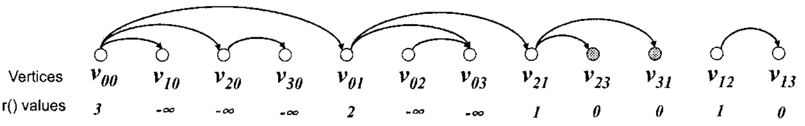

FIG. 4. A topological sort of the matrix spectrum graph shown in Fig. 3. Again, the terminal nodes are colored gray. The r() values are shown here for the audience's convenience.

Step 3: Depth-first-search and backtracking for suboptimal solutions. We use the depth-first-search (DFS) algorithm to find all suboptimal solutions. The nodes are sorted topologically, starting with v_{00}. Let the current path be $Q = (v_{00}, \ldots, u)$ and the path score of Q be $f(Q)$. There are two possible operations for u:

- Backtracking: if $f(Q) + r(u) < \alpha \cdot f_{max}$, there does not exist an α-suboptimal path with Q as the prefix subpath, so u is deleted from Q. Let s be the next-to-last node of Q and $(s, u) \in E$. The next node t after u in the topological order is selected and added to Q if $(s, t) \in E$. For example, in Fig. 4, $f_{max} = 3$. Suppose that the current path is $Q = (v_{00}, v_{01}, v_{03})$, where the last node $u = v_{03}$ with $r(u) = -\infty$, $f(Q) + r(u) < \alpha \cdot f_{max}$ for any positive α-value. In this case, $u = v_{03}$ is deleted from the path Q and the next node $t = v_{21}$ is selected and added to Q.
- Exploring: if $f(Q) + r(u) \geq \alpha \cdot f_{max}$, there exists an α-suboptimal path with Q as its prefix subpath, so the first node w in the topological order is selected and added to Q if $(u, w) \in E$. Again, see Fig. 4; Q is now (v_{00}, v_{01}, v_{21}), $f(Q) + r(u) = 2 + 1 = 3 = f_{max} \geq \alpha \cdot f_{max}$ for any α less than 1. Thus, v_{21} is included in Q, and the next node v_{23} is added to Q.

The algorithm iterates this process until Q is empty. In each iteration, if the last node u of Q is in the set of the terminal nodes T and $f(Q) \geq \alpha \cdot f_{max}$, Q is reported as a suboptimal solution and the algorithm continues.

The algorithm wastes no time in exploring any feasible path that is not α-suboptimal. The first solution is found in $O(|E|)$ time by the DFS algorithm because each edge is explored at most once and backtracked at most once. From one solution to the next solution, each edge is also backtracked at most once and explored at most once. Therefore, if there are p suboptimal solutions, this step takes $O(p|E|)$ time.

The complexity for this suboptimal algorithm is $O(p|E|)$ time and $O(|V| + |E|)$ space.

Correctness of the algorithm

We show in the following that the above suboptimal algorithm correctly finds all α-suboptimal feasible paths for G. The algorithm trims the total search space by the following heuristics corresponding to the two operations, Backtracking and Exploring, respectively, in Step 3:

- If $Q = (v_{00}, \ldots, v_{ij})$ is the current path in the algorithm and $f(Q) + r(v_{ij}) \geq \alpha \cdot f_{max}$, then there exists at least one α-suboptimal path with Q as its prefix. So, the algorithm should further explore the outgoing edges of v_{ij}.
- If $f(Q)+r(v_{ij}) < \alpha \cdot f_{max}$, then any feasible path P with Q as its prefix satisfies $f(P) \leq f(Q)+r(v_{ij}) < \alpha \cdot f_{max}$. So, the algorithm should not further explore Q but rather should backtrack v_{ij} from Q.

The correctness proof is as follows. Obviously, any output path of the algorithm is a feasible path. On the other hand, any α-suboptimal feasible path P, satisfying $f(P) \geq \alpha \cdot f_{max}$, can be found by the algorithm. For every edge $(v_{ij}, v_{mn}) \in P$, let $Q \subseteq P$ be the prefix subpath from v_{00} to v_{ij}, and let $R \subseteq P$ be the suffix subpath from v_{mn} to the terminal node of P. Then,

$$f(Q) + f(v_{ij}, v_{mn}) + r(v_{mn}) \geq f(Q) + f(v_{ij}, v_{mn}) + f(R) = f(P) \geq \alpha \cdot f_{max},$$

which means that if the prefix subpath Q is explored by the algorithm, the next edge $(v_{ij}, v_{mn}) \in E$ will be explored later. By inference, eventually P will be found by the algorithm. Meanwhile, no two outputs are the same, because all of the nodes are added in the topologically sorted order. Once a path Q is explored, all solutions containing Q as the prefix subpath will be found. If the last node of Q is deleted, Q will never appear in the algorithm again.

Ranking the suboptimal solutions

The suboptimal solutions are ranked using the following procedure: (1) a hypothetical mass spectrum is generated for each candidate peptide; (2) each hypothetical spectrum is then scored against the real mass spectrum using a simplified scoring function described below and (3) the candidate peptides are then ranked according to the scores.

The hypothetical spectra are generated as follows. For each suboptimal solution represented by a sequence of edges in the matrix spectrum graph, we first generate all of the possible candidate peptides, and then we construct a hypothetical spectrum for each of them. Only three types of ions are considered in the hypothetical spectrum: b-ions, y-ions, and b–H_2O ions. The abundance levels of the b-ions and the y-ions are set at 50 while the abundance levels of the b–H_2O ions are set at 25.

Each hypothetical spectrum is then compared against the experimental spectrum using the following scoring function. Let S1 be the sum of the abundance levels of all of the ions in the hypothetical spectrum and let S2 be the sum of the abundance levels of the ions (in the hypothetical spectrum) that match with some mass peaks in the experimental spectrum. We then compute the ratio S2/S1 for each hypothetical spectrum. The ratio S2/S1 shows how good each hypothetical spectrum fits the experimental spectrum. The candidate peptides are ranked according to the S2/S1 ratios.

Algorithm considering b–H_2O *ions*

The CID fragmentation will produce mainly b-ions and y-ions. However, other ions, such as b–H_2O ions (the type of ion formed when a b-ion loses one molecule of water) or y–H_2O ions, will also be present, usually with lower probability. We present here a suboptimal algorithm to consider the occurrence of b–H_2O in the spectra. The same idea can be applied to take other ions into account.

When we consider b–H_2O ions, each peak in the spectrum will be interpreted as a b-ion, a y-ion, or a b–H_2O ion. As a consequence, in the NC-spectrum graph, every simple N-node will become a supernode, consisting of one subnode where we interpret the peak as a b-ion and another subnode where we interpret the peak as a b–H_2O ion. Thus, without loss of generality, assuming $i < j$, the recursion for dynamic programming will become (recall that the original recursion for the Q matrix is $Q(i, j) = \max\{Q(i, l) + f(y_l, y_j),\ l = 1, \cdots, j - 1\}$):

$$S(x_i, y_j) = \max \begin{cases} S(x_i, y_l) + 1, \text{if} E(y_j, y_l) = 1, l = 1, \cdots, i-1 \\ S(x_i, z_l) + 1, \text{if} E(y_j, z_l) = 1, l = 1, \cdots, i-1 \end{cases}$$

$$S(z_i, y_j) = \max \begin{cases} S(z_i, y_l) + 1, \text{if} E(y_j, y_l) = 1, l = 1, \cdots, i-1 \\ S(z_i, z_l) + 1, \text{if} E(y_j, z_l) = 1, l = 1, \cdots, i-1 \end{cases}$$

$$S(z_i, z_j) = \max \begin{cases} S(z_i, y_l) + 1, \text{if} E(z_j, y_l) = 1, l = 1, \cdots, i-1 \\ S(z_i, z_l) + 1, \text{if} E(z_j, z_l) = 1, l = 1, \cdots, i-1 \end{cases}$$

$$S(x_i, z_j) = \max \begin{cases} S(x_i, y_l) + 1, \text{if} E(z_j, y_l) = 1, l = 1, \cdots, i-1 \\ S(x_i, z_l) + 1, \text{if} E(z_j, z_l) = 1, l = 1, \cdots, i-1 \end{cases}$$

Here, we use matrix S instead of matrix Q. Because each node can have two subnodes, we need to consider four possible connections between the two nodes. One thing that needs to be pointed out here is that for N-terminal nodes, each node has a subnode—a z-node representing a b–H_2O ion; while for C-terminal nodes, although each node still contains two subnodes, the coordinate of the second node (z-node) is set to 0. Confusion is avoided by this straightforward method because there will be no edge to end in a subnode with a 0 coordinate. We can then build an NC-spectrum graph based on matrix S. From the NC-spectrum graph, we can further build a matrix spectrum graph and perform our suboptimal algorithm on the matrix spectrum graph to find all suboptimal solutions. The suboptimal solutions are then ranked by the scoring function to find the candidate peptides.

TABLE 3. SOME FEATURES OF THE EXPERIMENTAL SPECTRA USED FOR OUR PROGRAM[a]

	Real sequence	*#Ions in real spectrum*	*#Hypo. b+y Ions*	*#b+y Ions presented*	*%b+y Ions presented*
1	DLGEEHFK	21	14	7	50.00
2	AEFVEVTK	40	14	13	92.86
3	YLYEIAR	17	12	6	50.00
4	LVNELTEFAK	40	18	11	61.11
5	KQTALVELLK	33	18	11	61.11
6	LSQKFPK	40	12	7	58.33
7	LGEYGFQNALIVR	40	24	14	58.33
8	ATEEQLK	40	12	9	75.00
9	DAFLGSFLYEYSR	40	24	15	62.50
10	FKDLGEEHFK	40	18	14	77.78
11	KVPQVSTPTLVEVSR	40	28	11	39.29
12	HLVDEPQNLIK	40	20	15	75.00
13	RHPEYAVSVLLR	40	22	13	59.09
14	TVMENFVAFVDK	23	22	8	36.36

[a] “#Hypo. b+y Ions” is the number of hypothetical b-ions and y-ions for the real peptide; “#b+y Ions Presented” is the number of hypothetical b-ions and y-ions that are presented by the real spectrum; “%b+y Ions Presented” is the percentage of hypothetical b-ions and y-ions that are presented by the real spectrum.

TABLE 4. THE TEST RESULTS OF THE SUOPTIMAL ALGORITHM ON 14 EXPERIMENTAL TANDEM MASS SPECTRA OF BSA PROTEINS[a]

	Real sequence	*Top reported seq.*	*Correct seq. by the program*	*Ranking*	*#Solutions*
1	DLGEEHFK	[228.72]GE[266.35][275.73]	[228.72]GE[266.35][275.73]	1	370
2	AEFVEVTK	AEFVEVT[K/Q]	AEFVEVT[K/Q]	1	1,927
3	YLYEIAR	Y[I/L]Y[240.61][230.02]	Y[I/L]Y[240.61][230.03]	1	147
4	LVNELTEFAK	[212.78]NE[I/L]TEFA[K/Q]	[212.78]NE[I/L]TEFA[K/Q]	1	11,154
5	KQTALVELLK	[K/Q][K/Q][171.78][I/L]VE[226.32][K/Q]	[K/Q][K/Q][171.78][I/L]VE[226.32][K/Q]	1	3,880
6	LSQKFPK	[171.90]REFP[K/Q]	[I/L]S[K/Q][K/Q]FP[K/Q]	2	3,962
7	LGEYGFQNALIVR	[173.6]VMNEFW[I/L][I/L]VR	[298.7]YGF[K/Q]N[184.7][I/L]VR	2	41,470
8	ATEEQLK	ATS[K/Q]ER[K/Q]	ATEE[K/Q][I/L][K/Q]	5	2,830
9	DAFLGSFLYEYSR	[261.46]AN[143.9]F[I/L]Y[K/Q]H[I/L][157.18]	[334.2][168.9]SF[I/L]YEYS[157.18]	7	27,464
10	FKDLGEEHFK	[293.78]HM[K/Q][170.08]PF[K/Q]	F[K/Q]D[I/L]GEEHF[K/Q]	9	12,613
11	KVPQVSTPTLVEVSR	[227.8]P[K/Q]VST[160.1]TS[290.27]SR	[227.8]P[K/Q]VSTP[224.03]V[228.4]SR	9	30,234
12	HLVDEPQNLIK	H[I/L]VDFT[K/Q]D[K/Q][204.48]	H[I/L]VDE[225.434]N[I/L][I/L][K/Q]	11	15,835
13	RHPEYAVSVLLR	RHV[K/Q][I/L]Y[142.74]T[I/L][I/L]R	RHPEY[170.39]SV[I/L][I/L]R	11	36,741
14	TVMENFVAFVDK	[199.78]MEHVE[I/L]TV[243.19]	[199.789]MENFVA[247.07][243.19]	16	27,785

[a] The notation [I/L] means that the interpretation can be either the amino acid I or L; similarly, the notation [K/Q] means that the interpretation can be either K or Q. We employ this notation because neither I/L nor K/Q can be distinguished in the mass spectrometry. "Top Reported Seq." is the top sequence reported by the *de novo* algorithm. "Correct Seq. by the Program" is the correct sequence reported by the *de novo* algorithm. The heading #Solutions is the number of b-ions and y-ions that appear in the real spectrum. The suboptimal value is set at $\alpha = 0.8$. For those mass spectra that have more than 40 mass peaks, only the top 40 peaks were selected for the suboptimal program.

TABLE 5. THE TEST RESULTS OF THE MODIFIED SUBOPTIMAL ALGORITHM WHERE b–H_2O IONS WERE ACCOUNTED FOR[a]

	Real sequence	*Top reported seq.*	*Correct seq. by the program*	*Ranking*	*#Solutions*
1	DLGEEHFK	[228.72]GSGYS[275.726]	not found	N/A	308
2	AEFVEVTK	AEFVEVT[K/Q]	AEFVEVT[K/Q]	1	6,122
3	YLYEIAR	Y[I/L]TRFTE	not found	N/A	12,554
4	LVNELTEFAK	[212.781]NE[I/L]TEFA[K/Q]	[212.78]NE[I/L]TEFA[K/Q]	1	9,023
5	KQTALVELLK	E[K/Q]DEVYTM[K/Q]	not found	N/A	14,181
6	LSQKFPK	[I/L]S[K/Q]GAFP[K/Q]	[I/L]S[K/Q][K/Q]FP[K/Q]	2	18,802
7	LGEYGFQNALIVR	[173.644]VMGGA[I/L][K/Q]MM[I/L]VR	not found	N/A	6,697
8	ATEEQLK	ATEEAG[I/L][K/Q]	not found	N/A	1,138
9	DAFLGSFLYEYSR	[332.988]N[144.264]F[I/L]YEH[269.979]	not found	N/A	956
10	FKDLGEEHFK	[274.821]R[K/Q]GENPF[K/Q]	not found	N/A	20,707
11	KVPQVSTPTLVEVSR	[227.8]P[K/Q]VST[160.6]VAG[251.6][242.7]	not found	N/A	7,865
12	HLVDEPQNLIK	H[I/L]VDE[225.18]N[I/L][I/L][K/Q]	H[I/L]VDE[225.18]N[I/L][I/L][K/Q]	1	6,927
13	RHPEYAVSVLLR	RHPE[I/L]P[209.7]V[I/L][I/L]R	not found	N/A	22,030
14	TVMENFVAFVDK	[199.8]MEHVEN[K/Q]A[243.196]	[199.8]MENFVAFV[243.2]	5	21,223

[a] See Table 4 for the explanation of "[I/L]" and "[K/Q]." The suboptimal value is again set at $\alpha = 0.8$. Only the top 40 peaks were selected for the suboptimal program for each spectrum.

RESULTS AND DISCUSSION

We implemented the suboptimal algorithm and tested it on the experimental tandem mass spectra of BSA proteins. The BSA proteins were first digested with trypsin and then injected into a reverse phase HPLC interfaced with a Finnigan LCQ ESI-MS/MS mass spectrometer.

The BSA protein was then digested ***in silico***, i.e., by the computer, to produce all possible trypsin-digested peptides. Each peptide was compared with the experimental tandem mass spectral data pertaining to the mass of the parent ion of each spectrum. The parent ion masses of the following peptides do not match with those of the experimental spectral data: CCTKPESER, YICDNQDTISSK, LCVLHEK, LVTDLTK, EYEATLEECCAK, YNGVFQECCQAEDK, CCAADDKEACFAVEGPK, LKPDPNTLCDEFK, DDPHACYSTVFDKLK, RPCFSALTPDETYVPK, LFTFHADICTLPDTEK, MPCTEDYLSLILNR, GLVLIAFSQYLQQCPFDEHVK, GACLLPK. Some other peptides whose parent ion masses match with the mass spectral data do not match well with the mass peaks in the mass spectral data. For example, the parent ion mass of the peptide SLHTLFGDELCK matches with the parent ion mass of some mass spectrum. However, none of the hypothetical y-ions of the peptide match with any peak in the real mass spectrum. The real mass spectrum might come from a contaminating protein in the BSA proteins. It is known that a good ladder of b-ions and/or y-ions is important for ***de novo*** peptide sequencing. Thus, these peptides are considered to have very poor matching spectra. They are excluded from our test: TCVADESHAGCEK, SHCIAEVEK, ETYGDMADCCEK, ECCHGDLLECADDR, CCTESLVNR, SLHTLFGDELCK, QTALVELLK.

For the rest of the peptides, the best matching mass spectrum for each peptide was chosen for out study. Some features of the spectra against the corresponding peptide were first studied and summarized in Table 3. On average, the spectra contain 61% of hypothetical b-ions and y-ions. These spectra are used for the test of our suboptimal program. The test results are summarized in Table 4. Among the 14 tandem mass spectra chosen for the test, the program successfully identified 5 peptides with scores that ranked them at the top among all suboptimal solutions. For the remaining 9 mass spectra, the program also successfully identified the correct peptides. The rankings of these peptides varied from 2 to 16, but all are in the top 0.1% among all the suboptimal solutions.

We also explored the possibility of considering b–H_2O ions in our suboptimal algorithm. The results are shown in Table 5. We can see that when b–H_2O ions are considered, the results are not as good. One reason is that when the program considered b–H_2O ions, it misinterpreted noise peaks as some kinds of b–H_2O ions. This suggests that the more ion types that are considered, the larger is the number of suboptimal solutions that exist, and the more likely it is that the real peptides hide among the many possible solutions. For the ***de novo*** sequencing to be successful, it is important to experimentally improve the predictability of the peptide fragmentation process.

Here, we chose a simplified scoring function to test the correctness of our algorithm. Obviously, there exist better and more sophisticated scoring functions. We will consider them in a later version of our program.

ACKNOWLEDGMENTS

We thank Dr. George M. Church at the Harvard Medical School for providing the tandem mass spectral data of the BSA protein. This research was partially supported by grants NSF-ITR EIA—0112934, NIH NIGMS 1-R01-RR16522-01, and the University of Southern California.

REFERENCES

Bafna, V., and Edwards, N. 2001. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 17, Suppl. 1, S13–21.

Chen, T., Kao, M.Y., Tepel, M., Rush, J., and Church, G.M. 2001. A dynamic programing approach for *de novo* peptide sequencing via tandem mass spectrometry. *J. Comp. Biol.* 8, 325–337.

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. *Anal. Chem.* 71, 2871–2882.

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

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. Mass Spectrometry* 5, 976–989.

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.

Qin, J., Fenyo, D., Zhao, Y., Hall, W.M., Chao, D.M., Wilson, C.J., Young, R.A., and Chait, B.T. 1997. A strategy for rapid, high-confidence protein identification. *Anal. Chem.* 69, 3995–4001.

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.

Address correspondence to:
Ting Chen
Department of Biological Sciences
University of Southern California
1042 West 36th Place, DRB 290
Los Angeles, CA 90089

E-mail: tingchen@hto.usc.edu