JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 6, Numbers 3/4, 1999
Mary Ann Liebert, Inc.
Pp. 327–342

De Novo Peptide Sequencing via Tandem Mass Spectrometry

VLADO DANČÍK,[1,2] THERESA A. ADDONA,[1] KARL R. CLAUSER,[1]
JAMES E. VATH,[1] and PAVEL A. PEVZNER[3]

ABSTRACT

Peptide sequencing via tandem mass spectrometry (MS/MS) is one of the most powerful tools in proteomics for identifying proteins. Because complete genome sequences are accumulating rapidly, the recent trend in interpretation of MS/MS spectra has been database search. However, *de novo* MS/MS spectral interpretation remains an open problem typically involving manual interpretation by expert mass spectrometrists. We have developed a new algorithm, SHERENGA, for *de novo* interpretation that automatically learns fragment ion types and intensity thresholds from a collection of test spectra generated from any type of mass spectrometer. The test data are used to construct optimal path scoring in the graph representations of MS/MS spectra. A ranked list of high scoring paths corresponds to potential peptide sequences. SHERENGA is most useful for interpreting sequences of peptides resulting from unknown proteins and for validating the results of database search algorithms in fully automated, high-throughput peptide sequencing.

Key words: protein sequencing, mass-spectrometry.

1. INTRODUCTION

IN A FEW SECONDS, a tandem mass spectrometer is capable of ionizing a mixture of peptides with different sequences and measuring their respective parent mass/charge ratios, selectively fragmenting each peptide into pieces and measuring the mass/charge ratios of the fragment ions (MS/MS spectra of peptides). The peptide sequencing problem is then to derive the sequence of the peptides given their MS/MS spectra. For an ideal fragmentation process and an ideal mass spectrometer the sequence of a peptide could be simply determined by converting the mass differences of consecutive ions in a spectrum to the corresponding amino acids. This ideal situation would occur if the fragmentation process could be controlled so that each peptide was cleaved between every two consecutive amino acids and a single charge was retained on only the N-terminal piece. In practice, the fragmentation processes in mass spectrometers are far from ideal. As a result, *de novo* peptide sequencing remains an open problem and even a simple spectrum may require tens of minutes for a trained expert to interpret.

[1]Millennium Pharmaceuticals, 640 Memorial Drive, Cambridge, Massachusetts.
[2]Mathematical Institute, Slovak Academy of Sciences, Grešákova 6, Košice, Slovakia.
[3]Departments of Mathematics, Computer Science and Molecular Biology, University of Southern California, Los Angeles, California.

Previous attempts to develop automated *de novo* peptide sequencing algorithms have followed either global or local search paradigms. The former approach (Sakurai *et al.*, 1984) involves the generation of all amino acid sequences and corresponding *theoretical spectra*, i.e., calculation of all theoretically possible fragment masses for each sequence. The goal is to find a sequence with the best match between the experimental and theoretical spectrum. Since the number of sequences grows exponentially with the length of peptide, different pruning techniques were designed to limit the combinatorial explosion in global methods. Prefix pruning (Hamm *et al.*, 1986; Ishikawa and Niva, 1986; Johnson and Biemann, 1989; Siegel and Bauman, 1988; Yates *et al.*, 1991; Zidarov *et al.*, 1990) restricts the computational space to sequences whose prefixes match the experimental spectrum well. The difficulty with the prefix approach is that pruning frequently discards the correct sequence if its prefixes are poorly represented in the spectrum. Another problem is that the spectrum information is used only *after* the potential peptide sequences are generated.

Local approaches tend to be more efficient for *de novo* peptide sequencing because they use the spectral information *before* any candidate sequence is evaluated. In different modifications of the local approach the peaks in a spectrum are transformed to a *spectrum graph* representation (Bartels, 1990; Fernández de Cossío *et al.*, 1995; Hines *et al.*, 1992; Taylor and Johnson, 1997). The peaks in the spectrum serve as *vertices* in the spectrum graph while the edges of the graph correspond to linking vertices differing by the mass of an amino acid. Each peak in an experimental spectrum is transformed into several vertices in a spectrum graph, and each vertex represents a possible fragment ion type assignment for the peak. Since a spectral peak could be derived from either the N- or C-terminus of the peptide, allowing both is accommodated by converting all vertices to N-terminal equivalents. The *de novo* peptide sequencing problem is thus cast as finding the longest path in the resulting directed acyclic graph. Since efficient algorithms for finding the longest paths are known (Cormen *et al.*, 1991), such approaches have the potential to efficiently prune the set of all peptides to the set of high-scoring paths in the spectrum graph.

Though some groups developed *de novo* sequencing programs beginning in the late 1980s, none is in widespread use today. The more widely used database search programs (Clauser *et al.*, 1996; Eng *et al.*, 1994; Fenyo, 1997; Fenyo *et al.*, 1998; Mann and Wilm, 1994) rely on the ability "to look the answer up in the back of the book" when studying genomes of extensively sequenced organisms. Although database matching is very useful, a biologist who attempts to clone a *new* gene based on MS/MS data (Lingner *et al.*, 1997) needs *de novo* rather than database matching algorithms. However, the development of *de novo* algorithms falls behind the experimental practice due to the following unsolved computational problems:

- **Parameter Learning.** Existing algorithms tend to be instrument dependent, i.e., they are designed for the kind of fragment ions that are most likely for the authors' particular type of mass spectrometer. No rigorous approach to defining ion types and intensity thresholds in an instrument-independent fashion has yet been described.
- **Spectrum Graph.** When the peptide fragmentation is incomplete the spectrum graph may break into disconnected components. Random noise in the spectrum may generate many false vertices and edges in the spectrum graph that promote false-positive interpretations in the absence of a good scoring schema. Errors in the parent mass/charge assignment lead to misalignment between N-terminal and C-terminal vertices in the spectrum graph. No computational approach to adjust inappropriate parent mass/charge assignment has yet been described.
- **Scoring Schema.** No rigorous approach to scoring paths in the spectrum graph has yet been described such that it is difficult to recognize the correct answer among the possible solutions.
- **Sequencing Algorithm.** The best scoring path in the spectrum graph may correspond to unrealistic solutions because it uses multiple vertices associated with the same spectral peak. No approach to analyze ions of unknown charge state has yet been described.

We have developed an algorithm and software SHERENGA that addresses the above problems.

2. PEPTIDE SEQUENCING PROBLEM AND SPECTRUM GRAPH

Let A be the set of amino acids with molecular masses $m(a), a \in A$. A *peptide* $P = p_1, \ldots, p_n$ is a sequence of amino acids, and the (parent) mass of peptide P is $m(P) = \sum m(p_i)$. A *partial peptide* $P' \subset P$ is a substring $p_i \cdots p_j$ of P of mass $\sum_{i \le t \le j} m(p_t)$.

Peptide fragmentation in a *tandem mass spectrometer* can be characterized by a set of numbers $\Delta = \{\delta_1, \ldots, \delta_k\}$ representing *ion types*. A *δ-ion* of a partial peptide $P' \subset P$ is such a modification of P' that has mass $m(P') - \delta$. For tandem mass spectrometry, the *theoretical* spectrum of peptide P can be calculated by subtracting all possible ion types $\delta_1, \ldots, \delta_k$ from the masses of all partial peptides of P (i.e., every partial peptide generates k masses in the theoretical spectrum). An (experimental) spectrum $S = \{s_1, \ldots, s_m\}$ is a set of masses of (fragment) ions. A *match* between spectrum S and peptide P is the number of masses that experimental and theoretical spectra have in common.

Tandem mass spectrometry peptide sequencing problem. Given spectrum S, the set of ion types Δ, and the mass m find a peptide of mass m with the maximal match to spectrum S.

Denote the partial *N-terminal* peptide $p_1, \ldots, p_i$ as $P_i, i = 1, \ldots, n-1$ and the partial *C-terminal* peptide $p_j, \ldots, p_n$ as $P_j^-, j = 2, \ldots, n$. In practice MS/MS spectrum S consists mainly of some of δ ions of partial N-terminal and C-terminal peptides. For example, the most frequent N-terminal ions (in Biemann, 1990 notation) are b, a, b–H_2O, b–NH_3 ($\Delta = \{1, -27, -17, 16\}$) for an ion-trap mass spectrometer (Fig. 1).

Assume, for the sake of simplicity, that a spectrum from a tandem mass spectrometer consists mainly of N-terminal ions. We capture the relationship between peptide P and ion types $\Delta = \{\delta_1, \ldots, \delta_k\}$ by transforming spectrum S to a *spectrum graph* $G_\Delta(S)$ (Bartels, 1990). Vertices of the graph are integers representing potential masses of partial peptides. Every peak of spectrum $s \in S$ generates k vertices $V(s) = \{s + \delta_1, \ldots, s + \delta_k\}$. The set of vertices of spectrum graph then is $\{s_{\text{initial}}\} \cup V(s_1) \cup \cdots \cup V(s_m) \cup \{s_{\text{final}}\}$, where $s_{\text{initial}} = 0$ and $s_{\text{final}} = m(P)$. Two vertices u and v are connected by a directed edge from u to v if $v - u$ is the mass of some amino acid and the edge is *labeled* by this amino acid. If we look on vertices as potential N-terminal peptides, the edge from u to v implies that the sequence at v may be obtained by extending the sequence at u by one amino acid (Fig. 2).

A spectrum S of a peptide P is called *complete* if S contains at least one ion type corresponding to P_i for every $1 \le i \le n$. The use of a spectrum graph is based on the observation that for a complete spectrum there exists a path of length n from s_{initial} to s_{final} in $G_\Delta(S)$ that is labeled by P. This observation casts the tandem mass spectrometry peptide sequencing problem as one of the finding the correct path in the set of all paths (Fig. 3).

Unfortunately, experimental spectra frequently contain noise and are not complete. Another problem is that MS/MS experiments performed with the same peptide but a different type of mass spectrometer will produce significantly different spectra. Different ionization methods (electroscopy, MALDI, fast-atom bombardment) combined with different mass analyzers (ion trap, quadrupole, time of flight, magnetic sector) have dramatic impact on the propensities for producing particular fragment ion types. Therefore every algorithm for *de novo* peptide sequencing should be adjusted for a particular type of a mass spectrometer. To address this problem we first describe an algorithm for an *automatic* learning of ion types and intensity thresholds from a sample of experimental spectra of known sequence and a new approach to finding the correct path in a spectrum graph that addresses incomplete and noisy spectra. We introduce the *offset frequency function* that represents an important new tool for defining the ion-type tendencies for particular mass spectrometers. The offset frequency function allows different types of mass spectrometers to be evaluated based on their propensity to generate different ion types, thus making our algorithm *instrument independent*.

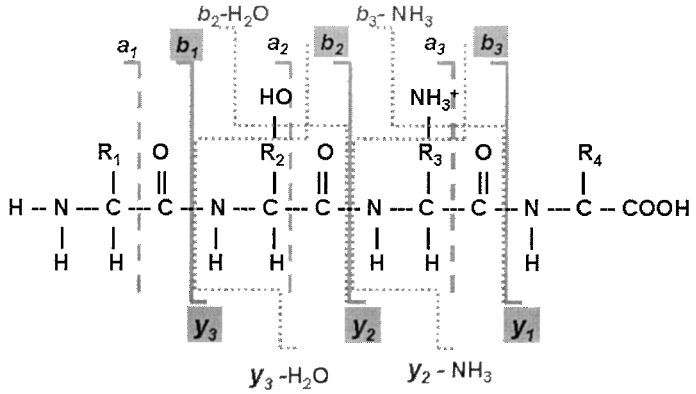

FIG. 1. Typical fragmentation patterns in tandem mass spectrometry.

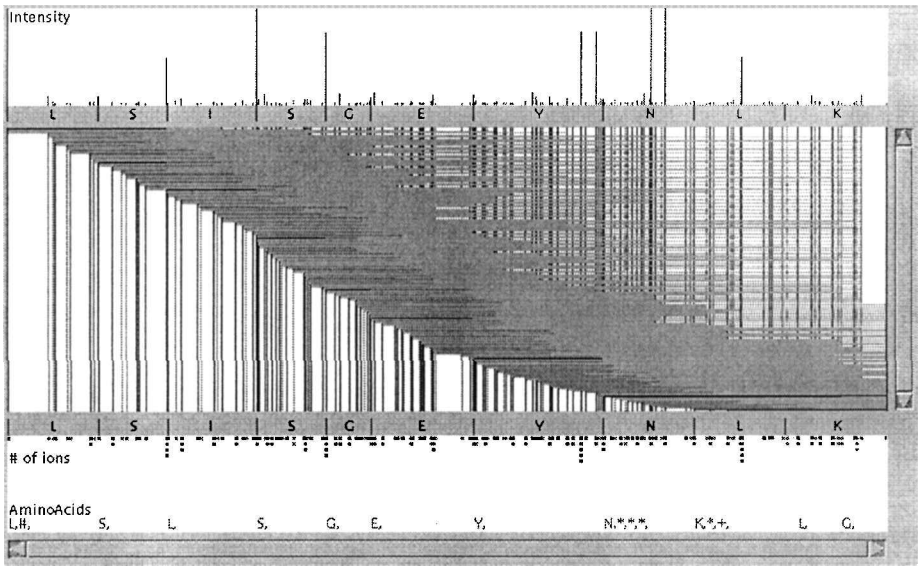

FIG. 2. Noise in spectrum generates many "false" edges in the spectrum graph and disguises edges corresponding to real peptide sequences. Sequence reconstructions correspond to finding an optimal path in the spectrum graph.

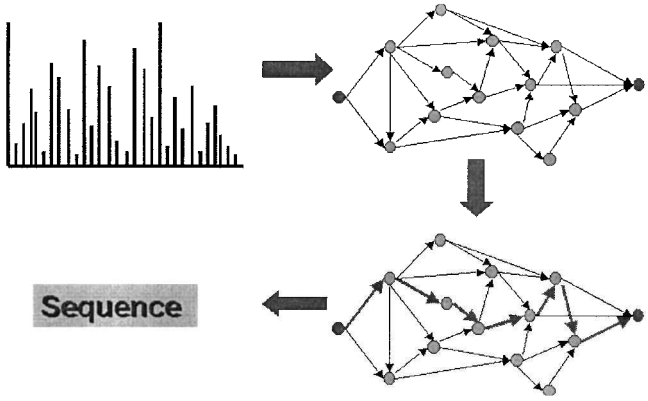

FIG. 3. Sequence reconstructions correspond to paths in the spectrum graph.

3. LEARNING ION TYPES AND INTENSITY THRESHOLDS

If the ion types $\Delta = \{\delta_1, \ldots, \delta_k\}$ produced by a given mass spectrometer are not known the spectrum cannot be interpreted. Below we show how to learn the set Δ and ion propensities from a sample of experimental spectra of known sequences *without* any prior knowledge of the fragmentation patterns.

Let $S = \{s_1, \ldots, s_m\}$ be a spectrum corresponding to the peptide P. A partial peptide P_i and a peak s_j have an offset $x_{ij} = m(P_i) - s_j$; we can treat x_{ij} as a discrete random variable. Since the probability of offsets corresponding to "real" fragment ions is much larger than the probability of random offsets, the peaks in the empirical distribution of the offsets reveal fragment ions. The statistics of offsets over all ions and all partial peptides provides a reliable learning algorithm for ion types.

Given spectrum S, offset x, and precision ε let $H(x, S)$ be the number of pairs (P_i, s_j), $i = 1, \ldots, n-1$, $j = 1, \ldots, m$ that have offset $m(P_i) - s_j$ within distance ε from x. The *offset frequency function* is defined as $H(x) = \sum_S H(x, S)$, where the sum is taken over all spectra from the learning sample (Fig. 4). To learn about C-terminal ions we do the same for pairs (P_i^-, s_j). Offsets $\Delta = \{\delta_1, \ldots, \delta_k\}$ corresponding to peaks of $H(x)$ represent the ion types produced by a given mass spectrometer. All the significant offsets we found correspond to known ion types (Table 1).

A peak in an MS/MS spectrum actually represents a mass/charge (m/z) ratio of the corresponding fragment ion. Mass spectrometers are capable of producing ions with charge 2 or even more; in such a case the observed

TABLE 1. INFORMATION ABOUT TERMINAL ION TYPES LEARNED FROM EXPERIMENTAL SPECTRA

Offset	*Integer offset*	*Count*	*Filtered count*	*Probability*	*Average intensity*	*Term*	*Ion*
18.85	19	604	604	0.6895	4.5457	C	y
0.85	1	568	565	0.6484	2.2788	N	b
−17.05	−17	338	338	0.3858	1.1966	N	b–H_2O
0.90	1	248	231	0.2831	0.5716	C	y–H_2O
−27.15	−27	204	200	0.2329	0.7537	N	a
20.05	20	183	180	0.2089	3.4699	C	y^2
−16.15	−16	159	152	0.1815	1.2766	N	b–NH_3
1.90	2	131	114	0.1495	0.6680	C	y–NH_3
−35.20	−35	151	141	0.1724	0.5253	N	b–H_2O–H_2O
−34.20	−34	134	131	0.1530	0.4736	N	b–H_2O–NH_3
−44.25	−44	129	126	0.1473	0.5516	N	a–NH_3
−45.15	−45	107	98	0.1221	0.4820	N	a–H_2O
2.30	2	102	95	0.1164	1.7460	C	y^2–H_2O
−16.10	−16	97	84	0.1107	0.4913	C	y–H_2O–NH_3
−17.15	−17	91	71	0.1039	0.4935	C	y–H_2O–H_2O

The remaining offsets have average count 45 and average intensity 0.431024.

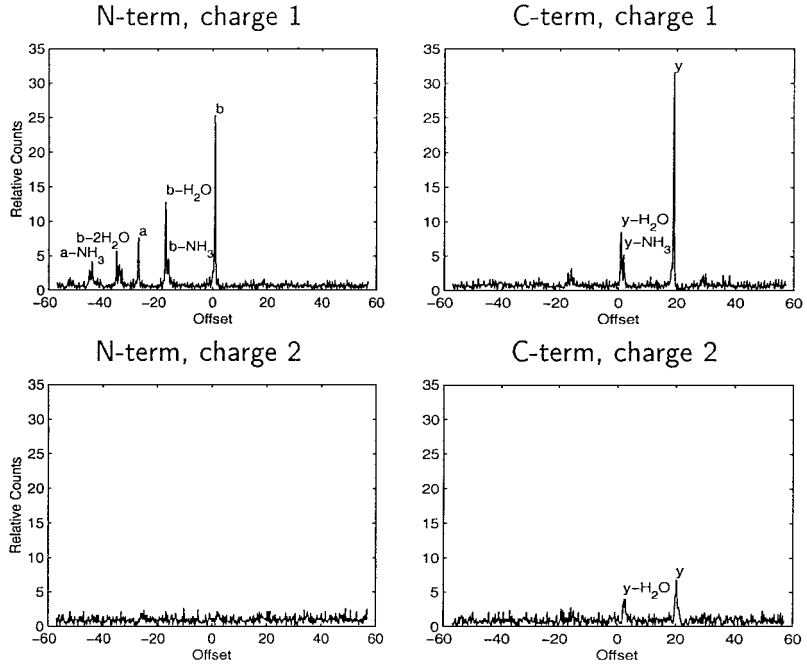

FIG. 4. Plots of the offset frequency functions. Horizontal axes represent offsets between peaks in spectra and masses of partial peptide molecules. Vertical axes represent normalized offset counts with 1 being the average count. The only significant offsets for doubly charged ions correspond to y and y–H_2O ions.

m/z is half (third, ...) of the ion's actual mass. We analyze doubly charged ions by investigating an offset frequency function $H^{+2}(x, S)$ where offsets are given by $m(P_i) - 2s_j$.

Peaks in a spectrum differ in *intensity* and it is necessary to address the question of setting a threshold for distinguishing the signal from noise in a spectrum prior to transforming it to a spectrum graph. Low thresholds lead to excessive growth of the spectrum graph while high thresholds lead to fragmentation of the spectrum graph. Earlier *de novo* sequencing algorithms set up the intensity thresholds for experimental spectra in a largely heuristic manner and have not addressed the fact that the intensity thresholds are ion type dependent. The offset frequency function allows intensity thresholds to be set up in a rigorous way.

Given a spectrum, we group intensities into *bins* of size K and *rank* K peaks with largest intensity by 1, next K peaks are ranked by 2 and so on. Figure 5 illustrates that the offset frequency function falls rapidly with rank increasing. The change of $H(x)$ depending on the intensity rank (Fig. 6) guides us in selecting

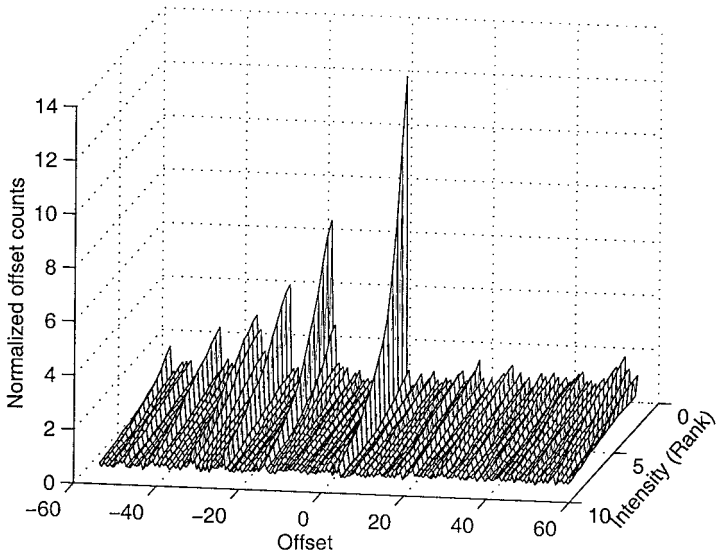

FIG. 5. Offset frequency function depending on intensity and rank.

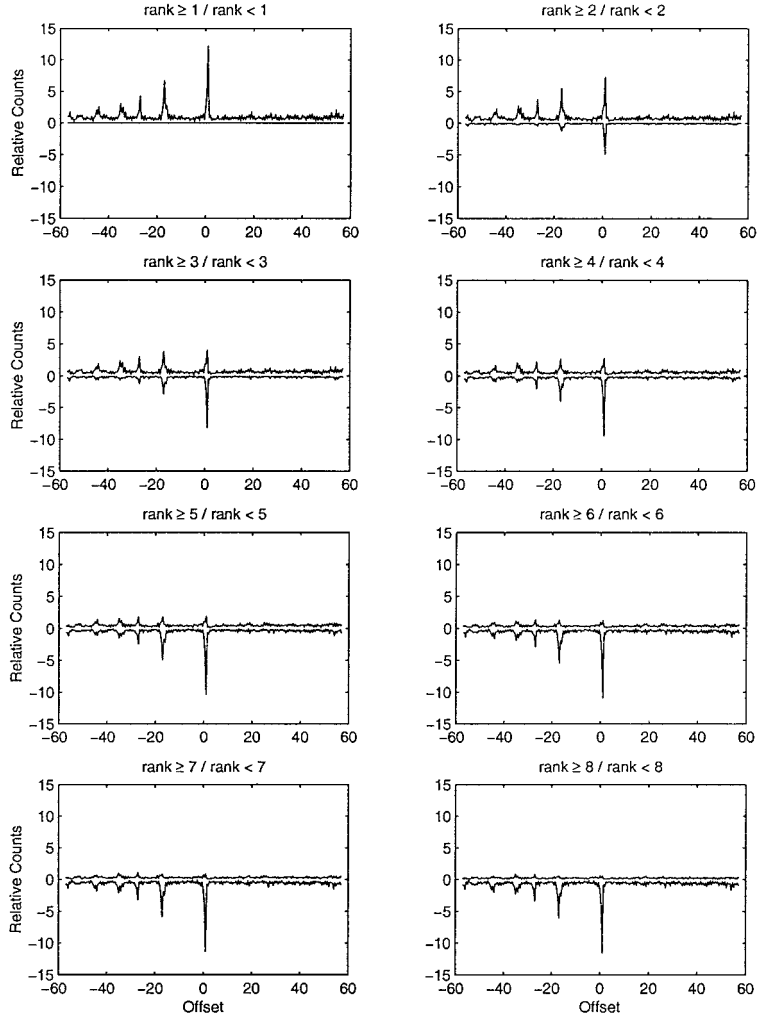

FIG. 6. Offset frequency function $H(x)$ for N-terminal ion types depending on rank [the size of the bins for computing ranks is (parent mass)/100]. Plots show offset frequency function for ions with rank at least i (upper parts) and with rank less than i (lower parts).

intensity thresholds. Instead of applying uniform intensity threshold for the whole input we apply thresholds selectively depending on ion types and a parent mass. Figure 6 convincingly demonstrates that the intensities ranked below 5 represent nothing but random noise since the offset frequency function has no pronounced peaks in this region. This conclusion suggests a limit for the number of peaks that should be used in MS/MS database search programs. Moreover, Fig. 6 demonstrates that intensity thresholds are ion type dependent. For example, the analysis of b-ions can be limited to intensity ranks 1, 2, and 3, while the analysis of b–H_2O ions can be limited to intensity ranks 3, 4, and 5. A similar analysis implies that only intensities ranked 1 and 2 (i.e., 20–30 high-intensity peaks) should be considered for y-ions while intensities ranked 2, 3, and 4 represent potential y–H_2O ions.

4. REDEFINING SPECTRUM GRAPH

A naive approach to construction of the spectrum graph described above does not take into account inaccuracies in experimental mass measurements. Above we assumed that if a partial peptide P_i produces peaks $s_1, \ldots, s_k$ corresponding to the ion types $\delta_1, \ldots, \delta_k$ then $s_1 + \delta_1 = s_2 + \delta_2 = \cdots = s_k + \delta_k = m(P_i)$ and all k-ion types generate the same vertex in the spectrum graph. Of course, this is not the case for real spectra. Due to inaccuracies of mass measurements the peaks $s_1, \ldots, s_k$ correspond to different vertices $s_j + \delta_j$, $1 \le j \le k$ within mass tolerance.

The *merging algorithm* decides what vertices in the spectrum graph are to be merged into one vertex. It is important to merge appropriate vertices; if we do not merge vertices that correspond to the same partial peptide, we will interpret meaningful peaks of spectra as noise. On the other hand, if we merge vertices that do not correspond to the same peptide, we may interpret noise as meaningful peaks.

We use a *greedy* algorithm to merge vertices. At every step we find the closest vertices, u (generated from peak s) and v (generated from t) and merge them. The weight of new vertex will be the weighted average $[i(s)u + i(t)v]/[i(s) + i(t)]$ of weights of u and v. We repeat merging until all vertices are at least ε apart for a given precision ε. The greedy algorithm for merging provides satisfying results for most spectra. Analysis of histograms of offsets between most frequent ion types can be used to select appropriate ε (0.5 in our case).

We connect vertices u and v by an edge (labeled a) if the mass of an amino acid a is *approximately* equal to the distance between the two vertices, i.e., $-\varepsilon < v - u - m(a) < \varepsilon$ for error range ε. Analysis of the histograms of offsets for all pairs of peak implies that $\varepsilon = 0.5$ is an appropriate choice for error range in defining edges of spectrum graph (data are not shown).

If a peptide undergoes incomplete fragmentation, the spectrum graph does not contain a vertex corresponding to an underrepresented position in that peptide. This can lead to a fragmented graph or, more frequently, a graph with paths that do not correspond to feasible solutions. To overcome this problem we modify the spectrum graph by introducing *gap* edges that model di- and tripeptides spanning underrepresented positions.

Sometimes in the course of merging, the weights of appropriate vertices become off more than $\varepsilon = 0.5$ even when there are corresponding peaks with difference within 0.5 of the amino acid mass. Since such vertices are not connected by an edge, we are at risk of losing important edges in the spectrum graph. To avoid it we introduce *bridge* edges in the spectrum graph. We connect two vertices u and v either by a (regular) edge with label a if $-\varepsilon < |v - u| - m(a) < \varepsilon$ or by a bridge edge if there are peaks $s, t \in S$ and ion type $\delta \in \Delta$ such that $-\varepsilon < |s - t| - m(a) < \varepsilon$ and vertex $s + \delta$ was merged into u and vertex $t + \delta$ was merged into v.

5. PARENT MASS

Accurate determination of the parent mass/charge is extremely important in *de novo* peptide sequencing. An error in parent mass measurement leads to systematic errors in the masses of vertices for C-terminal ions thus making peptide reconstruction difficult. In practice, the difference between the real parent mass (given by the sum of amino acids of a peptide) and the experimentally observed parent mass is frequently so large that the errors in sequence interpretation become almost unavoidable. To address this problem we have designed a combinatorial algorithm for parent mass correction (driven by the relationship between corresponding b- and y-ions and the parent mass) that can provide a more accurate determination of the parent mass than the experiments.

We use simple *alignment of spectra* to compute parent masses. If $S = \{s_1, \ldots, s_m\}$ is the spectrum of a peptide P then the *reflection* of S is a spectrum $\bar{S} = \{\bar{s}_1, \ldots, \bar{s}_m\}$ such that $\bar{s}_i = m(P) - s_i - d$, where

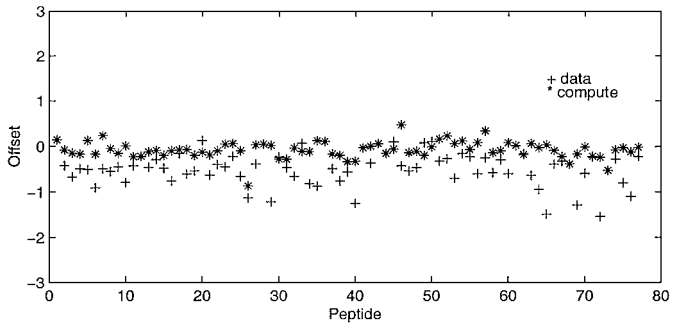

FIG. 7. The offsets between experimentally observed parent masses and $m(P)$ are marked by +. The offset between combinatorially computed parent masses and $m(P)$ are marked by *. The average error for parent mass computed by our algorithm is 0.0766 (standard deviation 0.1844) while for observed parent mass it is 0.4743 (standard deviation 0.3732).

$d = m(\text{y-ion}) - m(\text{b-ion})$ is the difference of offsets of y-ions and b-ions. Note that if a spectrum S contains a peak s that corresponds to a b-ion of a partial peptide P_i and peak t that corresponds to a y-ion of P_i^- then $\bar{s} = t$ and therefore spectra S and $\bar{S}$ have a common element. For correct $m(P)$ we should see good alignment between peaks corresponding to b-ions in S and peaks corresponding to y-ions in $\bar{S}$.

We use this observation to devise an algorithm for computing the parent mass. For a spectrum $S = \{s_1, \ldots, s_m\}$ and a number x we define $\bar{S}(x) = \{\bar{s}_1, \ldots, \bar{s}_m\}$ where $\bar{s}_i = x - s_i - d$. Spectra S and $\bar{S}$ may have some peaks in common just by chance, for a "random" mass x the number of peaks in common is approximately $[m(P)/d^2(S)] \approx 0.5$. However, for $x = m(P)$ spectra S and $\bar{S}$ tend to have more peaks in common due to the alignment between b-ions and y-ions. Since the condition that both P_i and P_i^- ions are present in the spectra is satisfied in 45% of cases (average number of aligned peaks is 6.4) we are able to devise the following combinatorial approach to estimate $m(P)$.

Let $c[S, \bar{S}(x)]$ be the number of peaks $s_i \in S$ and $\bar{s}_j \in \bar{S}(x)$ such that $|s_i - \bar{s}_j| < \varepsilon$, where ε is given precision. The value of x that maximizes $c[S, \bar{S}(x)]$ then would be an appropriate choice for parent mass. Should there be many choices for x, we can select one that minimizes the sum of distances $|s_i - \bar{s}_j|$ of the aligned peaks $s_i \in S$ and $\bar{s}_j \in \bar{S}$. Figure 7 demonstrates that this approach significantly improves the accuracy of the parent mass determination. This approach can similarly be used to correct a misassignment of the parent mass/charge value resulting from an incorrect charge assignment (data not shown).

6. SCORING PATHS IN SPECTRUM GRAPH

The goal of scoring is to quantify how well a candidate peptide "explains" a spectrum and to choose the peptide that explains the spectrum the best. If $p(P, S)$ is the probability that a peptide P produces spectrum S then the goal is to find a peptide P maximizing $p(P, S)$ for a given spectrum S. Below we describe a probabilistic model, evaluate $p(P, S)$, and derive a rigorous scoring schema for paths in the spectrum graph (versus largely heuristic previous approaches). The longest path in the weighted spectrum graph corresponds to the peptide P that "explains" spectrum S the best.

In a probabilistic approach tandem mass spectrometry is characterized by a set of ion types $\Delta = \{\delta_1, \ldots, \delta_k\}$ and their probabilities $\{p(\delta_1), \ldots, p(\delta_k)\}$ such that δ_i-ions of a partial peptide are produced independently with probabilities $p(\delta_i)$. A mass spectrometer also produces a "random noise" that in any position may generate a peak with probability q_R. Therefore, a peak at position corresponding to a δ_i-ion is generated with probability $q_i = p(\delta_i) + [1 - p(\delta_i)]q_R$ that can be estimated from the observed empirical distributions (Table 1). A partial peptide may theoretically have up to k corresponding peaks in the spectra. It has all k peaks with probability $\prod_{i=1}^{k} q_i$ and it has no peaks with probability $\prod_{i=1}^{k}(1 - q_i)$. The probabilistic model defines the probability $p(P, S)$ that a peptide P produces spectrum S. Below we describe how to compute $p(P, S)$ and derive scoring that leads to finding a peptide maximizing $p(P, S)$ for a given spectrum P.

Suppose that a candidate partial peptide P_i produces ions $\delta_1, \ldots, \delta_l$ ("present" ions) and does not produce the ions $\delta_{l+1}, \ldots, \delta_k$ ("missing" ions) in the spectrum S. These l "present" ions will result in a vertex in the spectrum graph corresponding to P_i. How should we score this vertex? The existing database search algorithms use "a premium for explained ions" and/or "penalty for unexplained ions" approach suggesting that the score for this vertex should be proportional to $q_1 \cdots q_l$ or maybe $q_1/q_R \cdots q_l/q_R$ to normalize the

probabilities against the chemical and electronic noise. (The ratios q_i/q_R can be taken from the offset frequency function.) Below we show that significant improvement results from penalizing for the nonpresence of ions in the experimental spectrum, which is possible from fragmentation of a candidate sequence. The (probability) score of the vertex is then given by

$$\frac{q_1}{q_R}\cdots\frac{q_l}{q_R}\frac{(1-q_{l+1})}{(1-q_R)}\cdots\frac{(1-q_k)}{(1-q_R)}$$

("premium for present ions, penalty for missing ions"). This important observation was overlooked in scoring the MS/MS database hits. Although the "premium for present ions, penalty for missing ions" approach may sound counterintuitive, it is confirmed both by our theoretical analysis and improvements in SHERENGA performance as compared to not penalizing for missing ions.

We explain the role of this principle for a resolution of a simple alternative between dipeptide GG and amino acid N of the same mass. In the absence of "penalty for missing ions" GG is selected over N in the presence of *any* (even very weak random noise) peak supporting the position of the first G. Our results imply that such a rule leads to many wrong GG-abundant interpretations and indicate that a better rule is to vote for GG if it is supported by a *b*- or *y*-type ion with sufficient intensities, which is automatically enforced by our "premium for present ions, penalty for missing ions" scoring. The same concepts extend to ambiguities between AG/GA vs. K or Q.

For the sake of simplicity we assume that all partial peptides are equally likely and ignore the intensities of peaks for now. We discretize the space of all masses in the interval from 0 to the parent mass $m(P)=M$, denote $T=\{0,\ldots,M\}$, and represent the spectrum as an M-mer vector $S=\{s_1,\ldots,s_M\}$ such that s_t is the indicator of presence/absence of peaks at position t ($s_t=1$ if there is a peak at position t and $s_t=0$ otherwise). For a given peptide P and position t, s_t is a 0–1 random variable with probability distribution $p(P,s_t)$.

Let $T_i=\{t_{i1},\ldots,t_{ik}\}$ be the set of positions that represents Δ-ions of a partial peptide P_i where $\Delta=\{\delta_1,\ldots,\delta_k\}$. Let $R=T\setminus\cup_i T_i$ be the set of positions that is not associated with any partial peptides. The probability distribution $p(P,s_t)$ depends on whether $t\in T_i$ or $t\in R$. For a position $t=t_{ij}\in T_i$ the probability $p(P,s_t)$ is given by

$$p(P,s_t)=\begin{cases} q_j, & \text{if } s_t=1 \text{ (i.e., a peak is generated at position } t)\\ 1-q_j, & \text{otherwise.}\end{cases} \tag{1}$$

Similarly for $t\in R$ the probability $p(P,s_t)$ is given by

$$p_R(P,s_t)=\begin{cases} q_R, & \text{if } s_t=1 \text{ (i.e., there is a random noise at position } t),\\ 1-q_R, & \text{otherwise.}\end{cases} \tag{2}$$

and the overall probability of "noisy" peaks in the spectrum can be estimated as $\prod_{t\in R} p_R(P,s_t)$.

Let $p(P_i,S)=\prod_{t\in T_i} p(P,s_t)$ be the probability that a peptide P_i produces a given spectrum at positions from the set T_i (all other positions ignored). For the sake of simplicity, assume that each peak of the spectrum belongs only to one set T_i and that all positions are independent. Then

$$p(P,S)=\prod_{t=1}^{M} p(P,s_t)=\left[\prod_{i=1}^{n} p(P_i,S)\right]\prod_{t\in R} p_R(P,s_t)$$

For a given spectrum S the value $\prod_{t\in T} p_R(P,s_t)$ does not depend on P and the maximization of $p(P,S)$ is the same as the maximization of

$$\frac{p(P,S)}{p_R(S)}=\frac{\prod_{i=1}^{n}\prod_{j=1}^{k} p\left(P,s_{t_{ij}}\right)\prod_{t\in R} p_R(P,s_t)}{\prod_{t\in T} p_R(P,s_t)}$$

$$=\prod_{i=1}^{n}\prod_{j=1}^{k}\frac{p\left(P,s_{t_{ij}}\right)}{p_R\left(P,s_{t_{ij}}\right)}$$

where $p_R(S)=\prod_{t\in T} p_R(P,s_t)$.

In logarithmic scale the above formula together with (1) and (2) implies the additive "premium for present ions, penalty for missing ions" scoring of vertices in the spectrum graph. It is worth mentioning that this scoring and "premium for present ions" scoring can be converted into each other by a parametric transformation. To incorporate the intensities into scoring we assume that intensity for ion type δ_j is distributed according to empirical distribution $I_{\delta_j}(x)$ and modify formulas (1) and (2) accordingly.

7. SEQUENCING ALGORITHM: ANTISYMMETRIC PATHS

After the weighted spectrum graph is constructed we cast the peptide sequencing problem as the *longest path problem in directed acyclic graph*. This problem is solved by a fast linear time dynamic programming algorithm, thus giving the spectrum graph approach an advantage over the global approaches.

Unfortunately, this simple algorithm does not quite work in practice. The problem is that every peak in the spectrum may be interpreted either as an N-terminal ion or C-terminal ion. Therefore, every "real" vertex (corresponding to a mass m) has a "fake" *twin* vertex [corresponding to a mass $m(P) - m - offset$]. Moreover, if the real vertex has a high score then its fake twin also has a high score. The longest path in the spectrum graph then tends to include *both* real vertex and its fake twin since they both have high scores. Such paths do not correspond to feasible sequence interpretations and should be avoided. However, the known longest path algorithms do not allow for avoiding such paths. This problem was overlooked in the previous work on *de novo* peptide sequencing.

Therefore, the reduction of the tandem mass spectrometry peptide sequencing to the longest path problem described earlier is inadequate. Below we formulate the *antisymmetric longest path* problem in a way that adequately models the peptide sequence interpretation.

Let G be a graph and let T be a set of *forbidden pairs* of vertices of G (twins). A path in G is called antisymmetric if it contains at most one vertex from every forbidden pair. The *antisymmetric longest path problem* is to find a longest antisymmetric path in G with a set of forbidden pairs T.

Unfortunately, the antisymmetric longest path problem is NP-hard (Garey and Johnson, 1979), thus indicating that efficient algorithms for solving this problem are unlikely. However, this negative result does not imply that it is hopeless to find an efficient algorithm for tandem mass spectrometry peptide sequencing since the problem has a *special structure* of forbidden pairs.

Vertices in the spectrum graph are numbers that correspond to masses of potential partial peptides. Two forbidden pairs of vertices (x_1, y_1) and (x_2, y_2) are *noninterleaving* if the intervals (x_1, y_1) and (x_2, y_2) do not interleave. A graph G with a set of forbidden pairs is called *proper* if every two forbidden pairs of vertices are noninterleaving.

Thus the tandem mass spectrometry peptide sequencing problem corresponds to an antisymmetric longest path problem in a proper graph. We prove that there exists an efficient algorithm for the antisymmetric longest path problem in a proper graph (to be described elsewhere).

8. RESULTS

Figure 8 shows the SHERENGA performance depending on the quality of spectra. A dot inside a rectangle in column i indicates that SHERENGA correctly computes the mass of partial peptide P_i in the corresponding row. Two consecutive dots in columns i and $i+1$ usually indicate that SHERENGA correctly interprets the ith amino acid in the corresponding row. The major conclusion is that SHERENGA almost always correctly interprets the representative positions of the spectra and misinterprets mainly nonrepresentative positions (corresponding to white and some light gray rectangles). Since any *de novo* algorithm would likely misinterpret nonrepresentative positions we feel that SHERENGA is close to the best possible performance of *de novo* algorithms. Moreover, SHERENGA misinterpretations are usually local, i.e., they are limited to substitutions of short pieces with the same mass. Figure 9 illustrates the first 20 predictions in more detail.

Figure 10 shows examples of interpretations with different quality. In the case of *perfect* interpretations SHERENGA annotates all the peaks correctly resulting in the correct peptide sequence. In the case of *good* interpretations SHERENGA annotates all or most of the peaks correctly and interprets the middle portion of the sequence. Good interpretations usually contain ambiguities at initial and/or terminal 1–3 amino acids. The perfect and good interpretations account for 75% of cases. Figure 10 also presents an example of a bad interpretation with misinterpreted peaks in the spectra and usually only a short piece of the peptide sequence

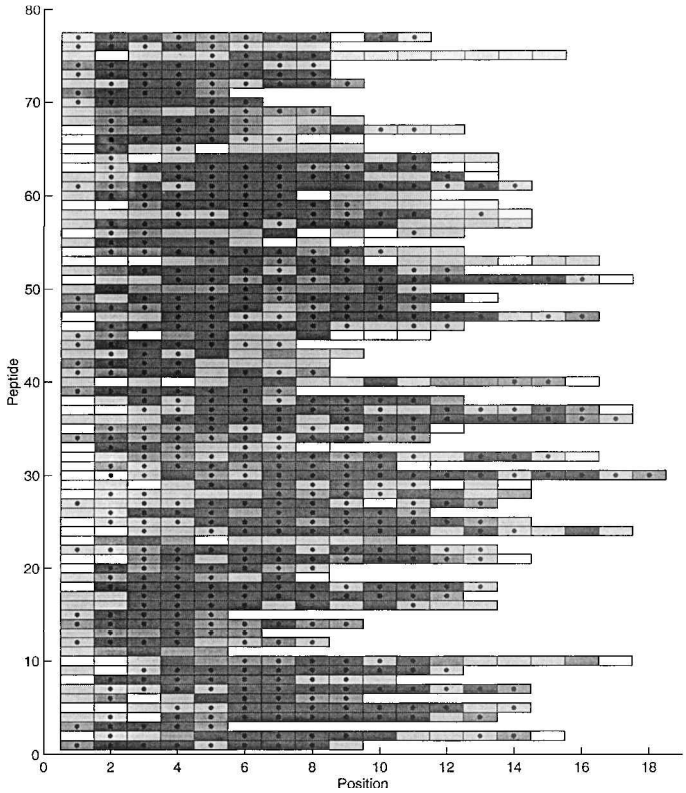

FIG. 8. The presence of ion types depending on position for spectra from our learning sample. Every rectangle corresponds to a partial peptide and the color of the rectangles indicates the presence/absence of different fragment ions. The intensity thresholds are defined based on the offset frequency function. The dark gray rectangle shows the presence of both b- and y-ions where at least one of them exceeds the intensity threshold. The medium dark gray rectangle shows the presence of a b- or y-ion (but not both) exceeding the intensity threshold. The light gray rectangle shows the presence of a low-intensity b and/or y-ion or the presence of another ion type. The white rectangle shows the absence of any peak in the spectra that may be associated with that position. The dots in the middle of rectangles indicate that the mass of the actual partial peptide is correctly assigned with the top scoring sequence interpreted by SHERENGA for the corresponding MS/MS spectrum. Dots in two consecutive columns indicate that SHERENGA correctly interprets the amino acid between the two positions.

interpreted correctly. In most such examples the correct answer contains a secondary (minor) fragment ion type with atypically strong intensity. Other causes of bad interpretation are spectra with very limited peptide fragmentation. Of course, in such cases any scoring cannot help since many sequence permutations are consistent with the spectral peaks. It is in such cases that database matching approaches are highly effective, as only one of the sequence permutations will be present in the database.

To evaluate the performance of *de novo* algorithms we introduce the *ladder difference* between the predicted and real peptide. Every peptide $P = p_1 \cdots p_n$ is associated with the ladder $L(P)$ of $n-1$ masses of partial peptides $m(P_i)$, $i = 1, n-1$. The ladder difference between peptides is a set difference between their ladders. For example, for (real) peptide TPVSEHVTK and (predicted) peptide TPVSCYVTK (Fig. 9) there are seven coinciding masses in the ladders (after T, P, V, S and before V, T, K) and two noncoinciding ones (before H and Y). The ladder difference $D(\text{TPVSEHVTK, TPVSCYVTK}) = 2$. More precisely, given the predicted peptide P_{pred} and real peptide P_{real} we define $|L(P_{\text{pred}}) \ominus L(P_{\text{real}})|$ as the false-positive error (equals 1 for our example) and $|L(P_{\text{real}}) \ominus L(P_{\text{pred}})|$ as the false-negative error (equals 1 for our example). The ladder distance between peptides is the sum of their false-positive and false-negative errors.

The initial/final positions of the peptides frequently contain sequence ambiguities. To eliminate the influence of these positions we adjust the definition of the ladder distance by excluding the initial/final positions that do not match. More precisely, we find the first and last matching elements in the ladders of real and predicted peptides and compare the ladders between these positions. If more than three initial/final elements of these ladders do not match, we compare the ladders starting/ending from/at three initial/final positions. Table 2 presents the cumulative results of SHERENGA predictions in the ladder distance measure.

TABLE 2. THE PERFORMANCE OF SHERENGA ALGORITHM IN THE LADDER DISTANCE

	0	*1*	*2*	*3*	*4*	*>5*
0	42.8	7.8	0.0	0.0	0.0	0.0
1	11.7	10.4	3.9	0.0	0.0	0.0
2	0.0	5.2	2.6	0.0	1.3	0.0
3	0.0	1.3	0.0	1.3	1.3	0.0
4	0.0	0.0	0.0	0.0	0.0	3.9
>5	0.0	0.0	0.0	0.0	1.3	5.2

The number in row *i* and column *j* shows the percentage of predictions with *i* false-positive errors and *j* false-negatives errors.

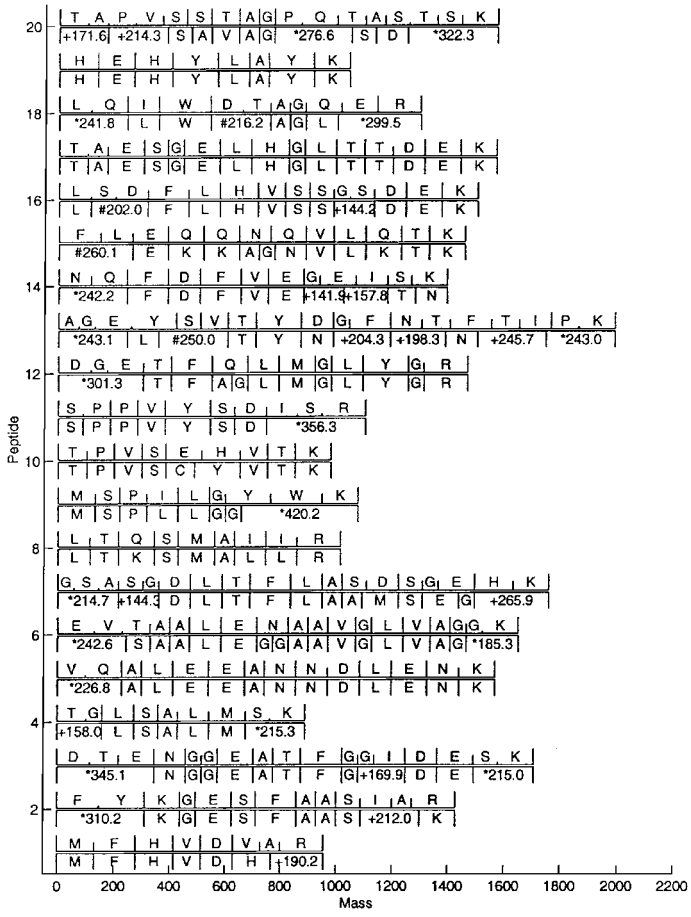

FIG. 9. The real (upper ladder) and predicted (lower ladder) peptides for the first 20 spectra in the sample. The peptides are shown by the ladders of bands reflecting the masses of partial peptides. Ambiguities are reflected by symbols $^{+}$ (less than 5 variants), $^{\#}$ (5–10 variants) and * (more than 10 variants) followed by the mass of the corresponding dipeptide or tripeptide. The height of the band in the upper ladder reflects the ion types supporting this band. A tall band indicates a support by a b- or y-ion, a medium band indicates a support by any other ion, a and short band indicates no support by ions. Ambiguities usually correspond to nonrepresentative positions (medium and short bands).

9. EXPERIMENTAL CONDITIONS AND SPECTRAL PREPROCESSING

Peptides were obtained from in-gel or in-solution tryptic digestion of proteins isolated from yeast lysates, mouse plasma, and urine. All MS/MS spectra were acquired using electrospray ionization on an LCQ ion trap mass spectrometer Finnigan-MAT (San Jose, CA) operated in automated LC/MS/MS mode. Reversed-phase high-performance liquid chromatography separation was performed with a 75-μm-id capillary columns

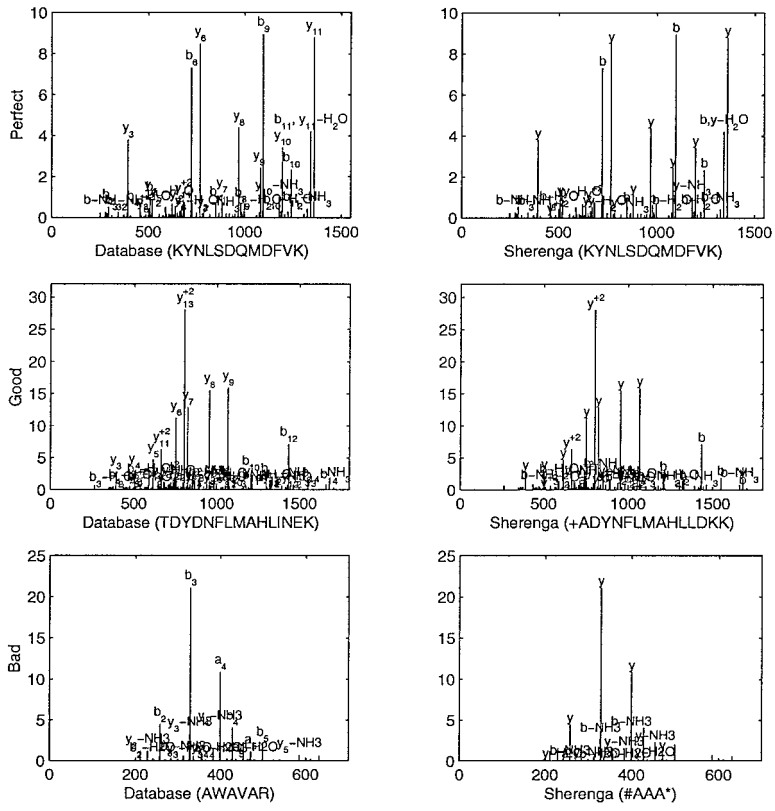

FIG. 10. Examples of SHERENGA predictions.

flowing at a rate of 0.5 μl/min with typical injection of 100 fmol–1 pmol peptide. Centroided MS/MS spectra were acquired in normal scan mode yielding unit resolution with 30 V relative collision energy, 3 Da isolation width, 5 microscans, and 60 msec maximum inject time. Prior to interpretation with SHERENGA, spectra were preprocessed to remove peaks in a 20-Da window below the precursor ion to eliminate the precursor ion and ions representing neutral losses of H_2O and NH_3 if present. Fragment ion peaks were deisotoped to retain only monoisotopic peaks.

10. DISCUSSION

Tandem mass spectrometry is very successful in identification of proteins present in genome databases (Patterson and Aebersold, 1995). Several database search algorithms use either partially interpreted (Mann and Wilm, 1994) or uninterpreted (Clauser *et al.*, 1996; Eng *et al.*, 1994; Fenyo *et al.*, 1998) data. These techniques require, of course, that the database contains the identical sequence or one very similar to the peptide studied. Furthermore, database search strategies tend to succeed in spite of sequence ambiguities arising from incomplete fragmentation where the MS/MS experiment does not fragment the peptide between each amino acid. Because they need discriminate only among the limited subset of sequences encoded in the genome of a living organism, database search strategies enable assignment of a sequence to a spectrum containing only partial fragmentation, while a *de novo* interpretation might yield only a few consecutive amino acids of sequence (Clauser *et al.*, 1999b).

Meanwhile, *de novo* peptide sequencing by MS/MS has not been widely practiced because of the tendency toward incomplete fragmentation with the most common instrumentation and the expertise required for spectral interpretation. As a result, MS/MS has had a very limited practical impact on discovery of *new* proteins. There are precious few examples describing cloning of a gene on the basis of MS/MS-derived sequence information alone (Wen *et al.*, 1992; Lingner *et al.*, 1997; Jimémez *et al.*, 1998). Subsequent design of degenerate PCR primers for gene amplification is typically hindered by the ambiguities and the short lengths of the *de novo* sequenced peptides. In both the cloning of GalB1,3(4)GlcNAca2,3-sialyltransferase (Wen *et al.*, 1992) and the

catalytic subunit of telomerase (Lingner *et al.*, 1997) 14 peptides were sequenced *de novo* following isolation of the protein. By contrast the small cardioactive peptide preprohormone (Jimémez *et al.*, 1998) was cloned using the single sequence derived *de novo* following isolation of the mature neuropeptide. In each case both data acquisition and interpretation were performed manually by expert mass spectrometrists.

The primary hindrance to generation of a complete peptide sequence has been the extent of fragmentation a peptide undergoes during an MS/MS experiment. Low-energy collision-induced dissociation (CID) instruments, such as triple quadrupole (Hunt *et al.*, 1986), hybrid sector quadrupole (Bean *et al.*, 1991), hybrid quadrupole time of flight (Shevchenko *et al.*, 1997), and ion traps (Jonscher and Yates, 1997), typically produce MS/MS spectra yielding a partial sequence of varying levels of completeness, and a complete sequence with peptides fortuitously amenable to fragmentation. In contrast near-complete peptide sequences typically result from MS/MS with instruments employing high-energy CID. Distinguishing between the amino acids leucine and isoleucine, which have identical mass, is also possible only with high-energy CID instruments. However, four-sector, (Medzihradszky and Burlingame, 1994) and hybrid sector/time of flight (Medzihradszky *et al.*, 1996) are available only in a few laboratories since they are expensive and need highly skilled operators. Fortunately for cloning purposes, after proteolytic digestion of a protein and sequencing its peptides, there are many chances to generate a peptide sequence while only two PCR primers of suitable degeneracy are needed for gene amplification.

The first discoveries of new proteins based on *de novo* MS/MS data and recent development in fully automated data acquisition for LC/MS/MS experiments confirm the need for development of a reliable *de novo* peptide sequencing algorithm. Furthermore, it is foreseeable that the development of an automated LC/MS/MS system integrated with a *de novo* interpretation algorithm could open a door toward "proteome sequencing." Long stretches of a polypeptide sequence could be assembled following generation and sequencing of overlapping sets of peptides from separate treatment of complex protein mixtures with proteolytic enzymes of differing specificity. Complete protein sequence determination has already been demonstrated with such a strategy on a single protein (Hopper *et al.*, 1989).

Mainstream gene discovery projects frequently start in model organisms with partially sequenced genomes where MS/MS-based database searches are susceptible to failure when the protein under study is not in the database. Frequent errors and uncertainties in ESTs provide another motivation for *de novo* MS/MS interpretation. To consistently succeed in such an environment, database search strategies employed with MS/MS spectral interpretation must therefore be mismatch tolerant. Consequently, recent efforts in algorithm development have focused on *sensitive* database searches to achieve that goal (Clauser *et al.*, 1999b; Taylor and Johnson, 1997). SHERENGA as a prelude to sequence similarity search programs such as BLAST provides a remedy in such a situation that can accommodate multiple substitutions/insertions/deletions in the target sequence as well as errors in ESTs. In the case of perfect or good SHERENGA interpretations (accounting for 75% of cases) PCR primer design is enabled. Moreover, ambiguities in peptide sequence interpretation can be accommodated by repeated similarity searches using the set of highest scoring sequences in the SHERENGA ranked list of candidate sequences.

Another application of SHERENGA exists in validation of conventional MS/MS database search. In fully automated sequencing efforts driven by database searching, minimizing the need for human data review depends on the ability to recognize cases when the peptide in question is not in a database or the database search is yielding a false-positive match. Correlation of the database search result with *de novo* interpretation can thus significantly improve the overall reliability of the process.

ACKNOWLEDGMENTS

We are grateful to Roland Annan, Klaus Biemann, David Clemens, Rich Ferrante, and Andrej Shevchenko for valuable discussions. This work was partially supported by Grant VEGA 2/4034/97.

REFERENCES

Bartels, C. 1990. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomed. Environ. Mass Spectrom.* 19, 363–368.

Bean, M.F., Carr, S.A., Thorne, G.C., Reilly, M.H., and Gaskell, S.J. 1991. Tandem mass spectrometry of peptides using hybrid and four-sector instruments: A comparative study. *Anal. Chem.* 63, 1473–1481.

Biemann, K. 1990. Appendix 5. nomenclature for peptide fragment ions (positive ions). *Methods. Enzymol*. 193, 886–887.

Clauser, K.R., Baker, P.R., and Burlingame, A.L. 1996. Peptide fragment-ion tags from maldi/psd for error-tolerant searching of genomic databases. *44th ASMS Conf. Mass Spectrom. Allied Topics*, *Portland, Oregon, May 12–16*, 365.

Clauser, K.R., Baker, P.R., and Burlingame, A.L. 1999a. Role of accurate mass measurement (10 ppm) in protein identification strategies employing ms or ms/ms and database searching. *Anal. Chem.* 71, 2871–2882.

Clauser, K.R., Baker, P.R., Foulk, R.A., Fisher, S.J., and Burlingame, A.L. 1999b. Peptide sequencing and homology-tolerant database searching using fragment-ion tags from maldi post-source decay mass spectra. In press.

Cormen, T.H., Leiserson, C.E., and Rivest, R.L. 1991. *Introduction to Algorithms*. MIT Press, Cambridge, MA.

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

Fenyo, D. 1997. A software tool for analysis of mass spectrometric disulfide mapping experiments. *CABIOS* 13, 617–618.

Fenyo, D., Qin, J., and Chait, B.T. 1998. Protein identification using mass spectrometric information. *Electrophoresis* 19, 998–1005.

Fernández de Cossío, J., Gonzales, J., and Besada, V. 1995. A computer program to aid the sequencing of peptides in collision-activated decomposition experiments. *CABIOS* 11(4), 427–434.

Garey, M.R., and Johnson, D.S. 1979. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, New York.

Hamm, C.W., Wilson, W.E., and Harvan, D.J. 1986. Peptide sequencing program. *CABIOS* 2, 365.

Hines, W.M., Falick, A.M., Burlingame, A.L., and Gibson, B.W. 1992. Pattern-based algorithm for peptide sequencing from tandem high energy collision-induced dissociation mass spectra. *J. Am. Soc. Mass. Spectrom.* 3, 326–336.

Hopper, S., Johnson, R.S., Vath, J.E., and Biemann, K. 1989. Glutaredoxin from rabbit bone marrow. *J. Biol. Chem.* 264, 20438–20447.

Hunt, D.F., Yates, J.R.D., Shabanowitz, J., Winston, S., and Hauer, C.R. 1986. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* 83, 6233–6237.

Ishikawa, K., and Niva, Y. 1986. Computer-aided peptide sequencing by fast atom bombardment mass spectrometry. *Biomed. Environ. Mass Spectrom*. 13, 373–380.

Jimémez, C.R., Li, K.W., Dreisewerd, K., Spijker, S., Kingston, R., Bateman, R.H., Burlingame, A.L., Smit, A.B., van Minnen, J., and Geraerts, W.P.M. 1998. Direct mass spectrometric peptide profiling and sequencing of single neurons reveals differential peptide patterns in a small neuronal network. *Biochemistry* 37, 2070–2076.

Johnson, R.J., and Biemann, K. 1989. Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomed. Environ. Mass Spectrom*. 18, 945–957.

Jonscher, K.R., and Yates, J.R. 1997. The quadrupole ion trap mass spectrometer—a small solution to a big challenge. *Anal. Biochem.* 244, 1–15.

Lingner, J., Hughes, T.R., Shevchenko, A., Mann, M., Lundblad, V., and Cech, T.R. 1997. Reverse transcriptase motifs in the catalytic subunit of telomerase. *Science* 276, 561–567.

Mann, M., and Wilm, M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399.

Medzihradszky, K.F., and Burlingame, A.L. 1994. The advantages and versatility of high-energy collision-induced dissociation based strategy for the sequence and structural determination of proteins. *Methods: Companion Methods Enzymol*. 6, 284–303.

Medzihradszky, K.F., Adams, G.W., Bateman, M.R., Green, R.H., and Burlingame, A.L. 1996. Peptide sequence determination by matrix-assisted laser ionization employing a tandem double focusing magnetic- acceleration time-of-flight mass spectrometer. *J. Am. Soc. Mass. Spectrom*. 7, 1–10.

Patterson, S.D., and Aebersold, R. 1995. Mass spectrometric approaches for the identification of gel-separated proteins. *Electrophoresis* 16(10), 1791–1814.

Sakurai, T., Matsuo, T., Matsuda, H., and Katakuse, I. 1984. Paas 3: A computer program to determine probable sequence of peptides from mass spectrometric data. *Biomed. Mass Spectrom.* 11(8), 396–399.

Shevchenko, A., Chernushevich, I., Ens, W., Standing, K.G., Thomson, B., Wilm, M., and Mann, M. 1997. Rapid 'de novo' peptide sequencing by a combination of nanoelectrospray, isotopic labeling and a quadrupole/time-of-flight mass spectrometer. *Rapid Commun. Mass Spectrom.* 11, 1015–1024.

Siegel, M.M., and Bauman, N. 1988. An efficient algorithm for sequencing peptides using fast atom bombardment mass spectral data. *Biomed. Environ. Mass Spectrom.* 15, 333–343.

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

Wen, D.X., Livingston, B.D., Medzihradszky, K.F., Kelm, S., Burlingame, A.L., and Paulson, J.C. 1992. Primary structure of gal beta 1,3(4)glcnac alpha 2,3-sialyltransferase determined by mass spectrometry sequence analysis and molecular cloning. Evidence for a protein motif in the sialyltransferase gene family. *J. Biol. Chem.* 267, 21011–21019.

Yates, J.R., Griffin, P.R., Hood, L.E., Zhou, J.X. 1991. Computer aided interpretation of low energy ms/ms mass spectra of peptides, 477–485. In Villafranca, J.J. ed., *Techniques in Protein Chemistry II*. Academic Press, San Diego.

Zidarov, D., Thibault, P., Evans, M.J., and Bertrand, M.J. 1990. Determination of the primary structure of peptides using fast atom bombardment mass spectrometry. *Biomed. Environ. Mass Spectrom*. 19, 13–16. 1990.

Address reprint requests to:
Vlado Dančík
Millennium Pharmaceuticals
640 Memorial Drive
Cambridge, MA 02139

E-mail: dancik@mpi.com