

On *de novo* interpretation of tandem mass spectra for peptide identification

Vineet Bafna*

Nathan Edwards†

ABSTRACT

The correct interpretation of tandem mass spectra is a difficult problem, even when it is limited to scoring peptides against a database. *De novo* sequencing is considerably harder, but critical when sequence databases are incomplete or not available. In this paper we build upon earlier work due to Dancik et al., and Chen et al. to provide a dynamic programming algorithm for interpreting *de novo* spectra. Our method can handle most of the commonly occurring ions, including *a*, *b*, *y*, and their neutral losses. Additionally, we shift the emphasis away from sequencing to assigning ion types to peaks. In particular, we introduce the notion of *core interpretations*, which allow us to give confidence values to individual peak assignments, even in the absence of a strong interpretation. Finally, we introduce a systematic approach to evaluating *de novo* algorithms as a function of spectral quality. We show that our algorithm, in particular the core-interpretation, is robust in the presence of measurement error, and low fragmentation probability.

1. INTRODUCTION

Proteomics is often defined as the direct analysis of the expressed proteins in various cellular processes. It incorporates tools from cell biology (isolating proteins from specific pathways or cellular compartments), protein chemistry (fractionation/separation and/or digestion of complex protein mixtures), and mass spectrometry for further analysis. There are two aspects to this analysis: *identification* of the expressed proteins, and *quantification*, or measuring levels of expression of specific proteins.

Recent advances in mass spectrometry instrument technology have made it possible to detect proteins at very low (picomole) concentrations, at an accuracy of a few parts per

*Informatics Research, Celera Genomics. Current address: The Center for Advancement of Genomics, 1901 Research Blvd., 6th floor, Rockville, MD 20850, vbafna@tcag.org

†Informatics Research, Celera Genomics, 45 W. Gude Drive, Rockville MD, 20850, Nathan.Edwards@celera.com

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

RECOMB '03 April 10–13, 2003, Berlin, Germany
Copyright 2003 ACM 1-58113-635-8/03/0004 ...\$5.00.

million. Coupled with improvements in computer hardware and algorithms for analysis, mass spectrometry, particularly *tandem mass spectrometry*, is rapidly becoming the method of choice for the high-throughput identification of proteins.

1.1 Mass Spectrometry

Put simply, a mass spectrometer is a device that measures masses. In tandem mass spectrometry, a peptide is subjected to stress induced fragmentation, and the mass of the fragments are then used as a *fingerprint* for the peptide. An *interpretation* of a tandem spectrum is the identification of the peptide, given this fingerprint. In the following, we describe the fragmentation process in some detail. Much of this section is paraphrased from an earlier publication [1] and can be skipped.

All amino-acids, the building blocks of proteins, have the same basic structure, shown in Figure 1(a). Amino-acids are distinguished from each other by the secondary structure of the side chain R. Amino acids form *peptides* when joined together in sequence by *peptide bonds*. This sequence of amino-acids identifies the peptide.

In tandem mass spectrometry (MS/MS), many peptides are ionized with one or more units of charge, and one chosen for fragmentation by *collision-induced dissociation* (CID). Fragments retaining the ionizing charge after CID have their mass-charge ratio measured. Since peptides typically break at a peptide-bond when they fragment by CID, the resulting spectrum contains information about the constituent amino-acids of the peptide. The fragmentation of the peptide in CID is a stochastic process governed by the physicochemical properties of the peptide and the energy of collision. The charged fragment can be inferred by the position of the broken bond and the side retaining the charge. In figure 1(b), the peptide bonds that break to form N-terminal a_1, b_1, c_1 fragments, and C-terminal $x_{n-1}, y_{n-1}, z_{n-1}$ fragments are shown. While *a*, *b*, *y* represent the commonly occurring fragments, high energy instruments often generate other fragments, including *internal* fragments formed by breakage of two peptide bonds, and fragments formed by breaks in side-chains. One or more of these fragments retain the charge unit(s), and their mass-charge ratio is registered. Figure 1(c) shows the single charge being retained by y_{n-1} .

In a single experiment, many charged fragments are formed by CID of multiple copies of the same peptide. The aggregate of the mass-charge ratios detected is called the *MS/MS spectrum*. A cartoon MS/MS spectrum for the peptide SGFLEEDK is shown in Figure 2. It helps illustrate how the MS/MS spectrum can be used to determine the sequence of amino-acids of a peptide. Note that the dif-

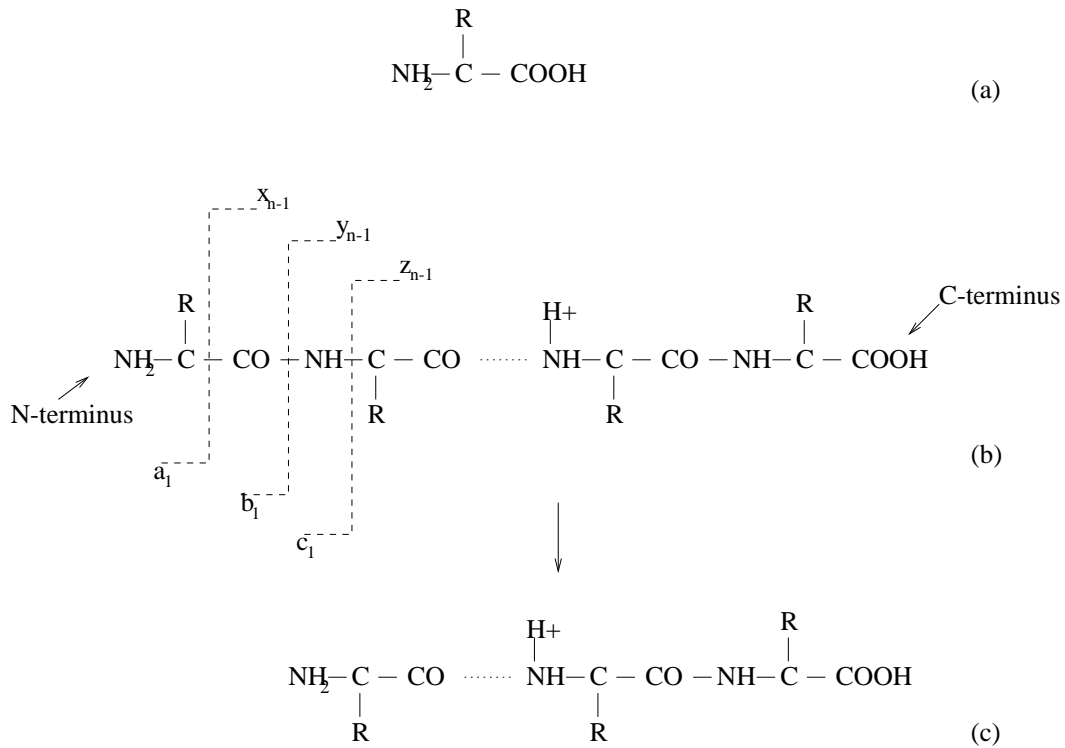


Figure 1: (a) The structure of an amino-acid. (b) An ionized peptide. (c) y_{n-1}^+ ion

88	145	292	405	534	663	778	924	b-ions
S	G	F	L	E	E	D	K	
924	837	780	633	520	391	262	141	y-ions

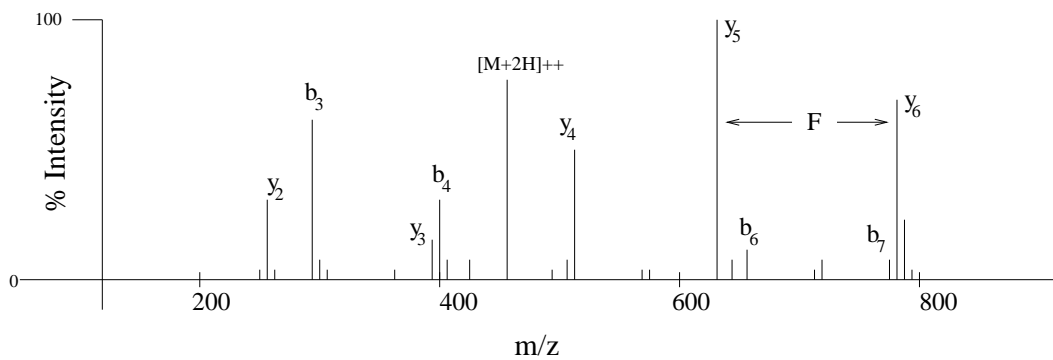


Figure 2: MS/MS spectrum for peptide SGFLEEDK.

ference in mass-charge ratio of the adjacent singly-charged y -ions, y_5 , and y_6 is exactly the mass of the residue F . If the fragmentation process produced every y -ion singly charged and no others, the difference between adjacent peaks in the *ladder* would indicate the amino-acid at each position along the peptide. In real MS/MS spectra however, there is no information on either type (b, y, \dots), position or charge of the fragment ion. Further, a complete ladder is usually not present. Additionally, some spectral peaks could be simply the result of contaminants, and finally, the measured mass-charge ratios are only as close to the actual mass-charge ratio as accuracy of the instrument guarantees.

Consequently, the unambiguous identification of peptide sequence using tandem mass spectrometry remains a challenge. While a comprehensive discussion of the different algorithms applied to this problem is beyond the scope of this paper, we describe the broad algorithmic approaches used to-date to put our work in context.

1.2 Earlier work

We argue that the core of most of the software programs for analyzing tandem MS data contain an implementation of the following three modules:

Interpretation: The *input* is a *MS/MS spectrum*, the *output* is *interpreted-MS/MS-data*. Interpreted-MS/MS-data is anything that can be reliably inferred from the MS/MS spectrum or the instrument. It may include parent peptide mass, partial or complete sequence tags, and combinations of sequence tags and molecular masses.

Filtering: The *input* is *interpreted-MS/MS-data* and a peptide sequence database. The *output* is a list of *candidate-peptides* that might have generated the MS/MS spectrum.

Scoring: The *input* is a list of *candidate-peptides* and the *MS/MS spectrum*. The *output* is a ranking of the *candidate-peptides* along with a score and possibly a p -value (probability that the score was achieved by random chance).

The flow of data through these modules is straightforward. The MS/MS spectrum is first processed, or *interpreted*, to find anything about the peptide that can be asserted with high confidence. Typically, this includes the parent peptide mass and possibly partial or complete sequence tags. This interpreted data is used to quickly *filter* a peptide sequence database to eliminate peptides that could not have generated the observed spectrum. For example, only peptides with mass approximately equal to the parent peptide mass of the MS/MS spectrum need be considered. The candidate peptides that pass the filter are then subject to a careful, and more expensive *scoring* procedure. The score ranks the candidate peptides and may even estimate the probability that the candidate peptide achieved a particular score entirely by chance.

1.3 Database Search vs. *de novo*

A class of algorithms (usually classified as *de-novo* sequencing algorithms) rely heavily on interpretation to identify the complete peptide sequence, and often do not use a database at all (see for example [2, 4, 5, 8, 13]). They

perform best when the spectra have relatively complete ladders and little noise. On the other hand, so-called *database-searching* algorithms [6, 7, 10, 12] rely primarily on good scoring. The peptide that scores the highest or has a low p -value is the one that best explains the spectrum. The success of these algorithms relies on the completeness of databases, and the availability of a good scoring mechanism. Regardless of their emphasis, most of the algorithms currently in use actually use elements of all three of these modules. The *de-novo* sequencing algorithms often output a list of possible peptides which need to be validated by database searching or additional experimentation. In-fact, Taylor and Johnson [13] propose combining their *de-novo* sequencing with Fasta [9] style database scoring. On the other hand, Mann and Wilm [10] report on the effectiveness of generating sequence tags for effective scoring. Other database searching programs like SEQUEST [6] do not generate sequence tags but filter the database on the basis of parent mass, and possibly immonium ions. It is clear that effective peptide identification software must make use of good algorithms for all of the three modules. Recently, Bafna and Edwards [1] argued that with the availability of whole genome, a pure *de novo* interpretation is not required. Interpretation and candidate peptide generation could be used mainly to eliminate peptides from being scored, and the focus should be on designing a good scoring module that explicitly models the chemistry of peptide fragmentation and ionization, spectral noise, and instrument error. They modeled the MS-MS spectrum generation as a two stage stochastic process, and presented an efficient dynamic programming algorithm for predicting the peptide most likely to have generated the spectrum according to the parameters of the model.

1.4 *De novo* interpretation

Clearly, scoring peptides against a database seems to be the method of choice for reliable interpretations. Nevertheless, *de novo* interpretation of MS/MS spectra is still a problem of great interest. Given the average size of an exon (150bp), and that of a peptide (30bp), there is a strong chance that a peptide of interest is split across two exons and cannot be identified in a straightforward translation. On the other hand, the available transcript and EST databases are still incomplete, and prone to sequencing errors. Often, the highest scoring peptide from a database search is not the obviously correct answer, and chemists often fall back to a manual *de novo* analysis of the spectrum. Finally, many model animals and plants are not likely to have complete databases for some time, and a *de novo* interpretation might be the only feasible solution.

Each peak in the MS/MS spectrum corresponds to cleavage between a pair of residues. If the assignment of the peak to an ion-type is known, the peak can give the mass of the residues in the prefix of a peptide. Consider for example, figure 2. If it is known that the peak around 600 is a y -ion, it corresponds to the prefix mass of 292 daltons which is the mass for *SGF*. Dancik et al. [4] introduced the notion of a spectral graph. Each spectral peak contributes several nodes, one for each assignment of an ion type. Each node has the corresponding prefix mass associated with it. There is an edge labeled with amino-acid a from a node u to node v if the difference in prefix masses equaled the residue mass of a . Finally, each node is weighted according to a "premium for present ions, penalty for missing ions" principle. The

interpretation of the spectrum is equivalent to the problem of finding the longest path in this graph. For good quality spectra, their interpretation often leads to the correct identification. As has been observed earlier [3, 11], a problem with this approach is that the longest path often include multiple nodes from the same mass peak.

The problem of finding the longest path while avoiding multiple assignments to the same mass peak is NP-complete in the general case. However, Chen et al. [3], and Dancik et al. [4] make the interesting observation that the forbidden pairs in MS-MS data are *non-interleaving*, and that allows a polynomial time approach to finding such paths. Chen et al. [3] exploit this abstraction. In their most general formulation, they allow each peak to be either a *b*, or a *y* ion or a noise peak, and find the highest scoring path, where each edge corresponds to a combination of residues, and every peak is assigned to at most one ion type.

We generalize and extend this approach in a few directions. First, we exploit a simple structural property of ion types, so as to be able to allow assignment to most of the commonly occurring ion types, including, but not limited to *a*, *b*, *y*, and their neutral losses (H_2O , and NH_3). This is a significant improvement in practice, because neutral losses etc. are fairly common in the presence of acidic and basic residues, and the *a*-ion is often seen in addition to the *b*-ion. Second, we reformulate the score function to correspond better to the chemists' intuition of penalizing high intensity noise peaks, and rewarding interpreted peaks according to intensity of the interpreted peak as well as the ion type assigned to each peak. Finally, we exploit the dynamic programming machinery to output sub-optimal interpretations, including a *core-interpretation*. A *core-interpretation* is an assignment of ions to peaks that is the same in every optimal interpretation of the spectrum.

2. DEFINITIONS

Peptide A *peptide* $p \in \mathcal{A}^n$ is a sequence of $n \geq 1$ residues from the set of amino-acids, $\mathcal{A} = \{A, C, \dots, Y\}$.

Residue-Mass The *residue-mass* $R(p)$ of a peptide is the sum of the masses of its amino-acid residues.

Peptide MS/MS Spectrum A *peptide MS/MS spectrum* is defined by the pair (M, S) , where M is the residue-mass of the (parent) peptide, and $S = \{s_1, s_2, \dots, s_k\}$ for $s_i \in \mathbf{R}^+$, which specifies the k observed mass-charge ratios of the spectrum.

Residue-Tag A *residue-tag* $r \in \mathbf{R}^+$ has the property that there exists a peptide p such that $R(p) = r$.

Residue-Tag-Sequence A *residue-tag-sequence* of a mass M is a sequence of residue-tags r_1, r_2, \dots, r_m , such that $M = \sum_i r_i$.

Prefix-Residue-Mass We define the *prefix-residue-mass* $R(s, \iota)$ of a mass-charge ratio $s \in S$ from a peptide MS/MS spectrum (M, S) and ion-type ι as

$$R(s, \iota) = \begin{cases} s - o(\iota) & \text{where } \iota \text{ is a} \\ & \text{N-terminal ion-type,} \\ M - (s - o(\iota)) & \text{where } \iota \text{ is a} \\ & \text{C-terminal ion-type.} \end{cases}$$

where $o(\iota)$ is the difference between the observed mass-charge ratio of a peptide fragment and the residue-mass of the amino-acid sequence of the fragment.

Interpreted-Mass-Sequence Let (M, S) be a peptide MS/MS spectrum and \mathcal{I} be a set of ion-types. Let A be a (possibly incomplete) assignment of ion-types from \mathcal{I} to the mass-charge ratios of S . Each ion-type assignment of A implies a prefix-residue-mass. Let R_1, R_2, \dots, R_m be the sorted list of prefix-residue-masses defined by the mass-charge ratios of S that have an ion-type assignment. The *interpreted-mass-sequence* $R(A, S, \mathcal{I}) = r_1, r_1, \dots, r_{m+1}$ is defined by

$$r_i = \begin{cases} R_1 & \text{for } i = 1, \\ R_i - R_{i-1} & \text{for } 2 \leq i \leq m, \\ M - R_m & \text{for } i = m + 1. \end{cases}$$

Interpretation An *interpretation* of a peptide MS/MS Spectrum (M, S) is a (possibly incomplete) assignment A of ion-types \mathcal{I} to the mass-charge ratios of S such that the interpreted-mass-sequence $R(A, S, \mathcal{I})$ forms a residue-tag-sequence of mass M .

Inner(Outer)-Interpretation Let R^L, R^R have the property that R^L, R^R , and $(R^R - R^L)$ are residue-tags and $R^L \leq R^R$. An (R^L, R^R) -*inner-interpretation* of a spectrum (M, S) is an interpretation of the form $R^L, r_1, r_2, \dots, (M - R^R)$. In other words, the assigned mass-charge ratios of S have prefix-residue-masses between R^L and $(M - R^R)$.

Similarly, we define the (R^L, R^R) -*outer-interpretation* to be an interpretation of the form $r_1, r_2, \dots, (R^R - R^L), \dots, r_m$. In other words, the assigned mass-charge ratios of S have prefix-residue-masses that are either at most R^L or at least $(M - R^R)$.

3. MS/MS INTERPRETATION SCORE

Obviously, all tandem mass spectra admit many interpretations, including the trivial one in which no peak is assigned an ion-type. We seek a score function that satisfies the chemists' intuition of a good interpretation, but can be computed efficiently. We could seek to maximize the number of residue-masses that correspond to single amino-acids, as the eventual goal is to identify the peptide. However, many peptides fragment incompletely, and there often isn't enough evidence for assigning individual amino acids. A second approach is to maximize the number of peaks that are interpreted, and penalize those that are not.

We generalize this slightly by defining a score $\delta(s, \iota)$, for a peak $s \in S$ assigned to a fragment ion $\iota \in \mathcal{I}$. Likewise, the function $\delta(s, \phi)$ also describes the penalty for not assigning s to any fragment ion. Define the *Sum of Assigned Peaks-score* (SAP-score) for an interpretation A as $\sum_{s \in S} \delta(s, A(s))$. The SAP-score abstracts much, but not all, of the chemists' intuition for a good score function. One of the popular measures for an interpretation is the *Total Ion Current* interpreted. This refers to the sum of the assigned peak intensities, with the intensity being described by spectral peak height, or area. Another popular approach is to score (and penalize) based on the ion-type, with *b, y* ion assignments being favored over others. It is easy to see that δ can be chosen appropriately to capture this intuition.

4. SIMPLE ION-TYPES AND SAP-SCORE COMPUTATION

Consider a peptide spectrum (M, S) , and a set of ion-types \mathcal{I} . Recall that every assignment of an ion ι to a peak $s \in S$ implies a prefix-residue-mass, $R(s, \iota)$. Let $R_s(\mathcal{I}) = \cup_{\iota \in \mathcal{I}} R(s, \iota)$ denote the set of prefix residue masses for s . Partition $R_s(\mathcal{I})$ into a left-set $R_s^L(\mathcal{I}) = \{r \in R_s(\mathcal{I}) \mid r \leq M/2\}$, and a right set $R_s^R(\mathcal{I}) = \{r \in R_s(\mathcal{I}) \mid r > M/2\}$. Let $r_s^L = \min\{r \mid r \in R_s^L(\mathcal{I})\}$ and $r_s^R = \max\{r \mid r \in R_s^R(\mathcal{I})\}$ be the extreme prefix residue masses for s .

DEFINITION 1.: *A set of ion-types \mathcal{I} is simple if*

1. For all $s \in S$ and all $r, r' \in R_s^L(\mathcal{I})$,
 $|r - r'| < \min_{a \in \mathcal{A}} R(a)$.
2. For all $s \in S$ and all $r, r' \in R_s^R(\mathcal{I})$,
 $|r - r'| < \min_{a \in \mathcal{A}} R(a)$.
3. For all $s, t \in S$, $r_s^L < r_t^L \Leftrightarrow r_s^R > r_t^R$.

The following lemma explains why simple ions-types are the key to designing an efficient algorithms for *de novo* interpretation of tandem mass spectra.

LEMMA 1.: *Given an interpretation A of a peptide MS/MS spectrum (M, S) using only simple ion-types \mathcal{I} , let $r_1, r_2, r_3, \dots, r_p$ be the interpreted mass sequence. Let m be such that $\sum_{i=1}^m r_i \leq M/2$ and $\sum_{i=m+1}^p r_i > M/2$. For each r_i of the interpreted mass sequence, define the spectral witness set S_i to be those $s \in S$ such that the prefix-residue-mass $R(s, A(s)) = \sum_{k=1}^i r_k$. Then,*

1. For all $1 \leq i < j \leq m$ and all $s \in S_i, t \in S_j$, $r_s^R > r_t^R$.
2. For all $m < j < i \leq p$ and all $s \in S_i, t \in S_j$, $r_s^R < r_t^R$.

PROOF. Let $1 \leq i < j \leq m$ have witness peaks $s \in S_i, t \in S_j$ such that $r_s^R < r_t^R$. Since A uses only simple ion-types, we must also have $r_s^L > r_t^L$. As $s \in S_i$, then $R(s, A(s)) = \sum_{k=1}^i r_k$; and similarly, since $t \in S_j$, then $R(t, A(t)) = \sum_{k=1}^j r_k$. The interpreted mass sequence r_1, r_2, \dots, r_p consists of residue-masses of at least $\min_{a \in \mathcal{A}} R(a)$. Therefore

$$R(t, A(t)) - R(s, A(s)) \geq \sum_{k=1}^j r_k - \sum_{k=1}^i r_k \geq \sum_{k=i+1}^j r_k \geq \min_{a \in \mathcal{A}} R(a).$$

On the other hand,

$$R(t, A(t)) - R(s, A(s)) \leq R(t, A(t)) - r_s^L \leq R(t, A(t)) - r_t^L.$$

Therefore

$$R(t, A(t)) - r_t^L \geq \min_{a \in \mathcal{A}} R(a). \quad (1)$$

Since $j \leq m$, it must be the case that $A(t) \in R^L(\mathcal{I})$. Therefore equation (1) contradicts the definition of simple ion-types. Case 2 follows similarly. \square

We fix additional notation for the remainder of this paper. For a peptide MS/MS spectrum (M, S) , we label the elements $s \in S$ according to increasing r_s^R . We pre-compute

two sets. The first is an array RM of size n containing all putative residue masses from $\cup_{s \in S} R_s(\mathcal{I})$ for a simple ion-type set \mathcal{I} . Also set $\text{RM}[0] = 0$, $\text{RM}[n] = M$. Second, we compute the set of all residue-tags less than or equal to M and denote it as \mathcal{V}_M . The trivial residue-tag 0 is explicitly added to \mathcal{V}_M .

Define $S[i][v][w]$ as the SAP-score of the highest scoring $(\text{RM}[v], \text{RM}[w])$ -inner-interpretation of peaks $1, 2, \dots, i$. The score of the complete interpretation is given by $S[k][0][n]$. The following recurrence holds

THEOREM 2.:

$$S[0][v][w] = \begin{cases} 0 & \text{if } \text{RM}[v] \leq \text{RM}[w], \\ & \text{RM}[v] \in \mathcal{V}_M, \\ & \text{RM}[w] - \text{RM}[v] \in \mathcal{V}_M, \\ & M - \text{RM}[w] \in \mathcal{V}_M; \\ -\infty & \text{otherwise.} \end{cases}$$

$$S[i][v][w] = \max \begin{cases} S[i-1][v][w] + \delta(i, \phi); \\ S[i-1][u][w] + \delta(i, \iota), \\ \quad \forall u \text{ s.t. } \text{RM}[u] = r_i(\iota) \in R_i^L, \\ \quad \text{RM}[u] \geq \text{RM}[v], \\ \quad \text{RM}[u] - \text{RM}[v] \in \mathcal{V}_M, \\ \quad \text{RM}[w] - \text{RM}[u] \in \mathcal{V}_M; \\ S[i-1][v][u] + \delta(i, \iota), \\ \quad \forall u \text{ s.t. } \text{RM}[u] = r_i(\iota) \in R_i^R, \\ \quad \text{RM}[u] \leq \text{RM}[w], \\ \quad \text{RM}[u] - \text{RM}[v] \in \mathcal{V}_M, \\ \quad \text{RM}[w] - \text{RM}[u] \in \mathcal{V}_M. \end{cases}$$

PROOF. In the $(\text{RM}[v], \text{RM}[w])$ -inner-interpretation of peaks $1, 2, \dots, i$, peak i either has an interpretation or not. If not, the score is given by $S[i-1][v][w]$ plus the penalty $\delta(i, \phi)$ for not using the peak. If peak i does have an interpretation $\iota \in \mathcal{I}$, we get a score of $\delta(i, \iota)$ for using that peak. Assume w.l.o.g, that the prefix residue mass assigned to peak i is given by $r_i(\iota) = \text{RM}[u] \in R_i^L$. Then, any interpretation of peaks $1, \dots, i-1$ that results in a prefix mass $r < \text{RM}[u]$ is a violation of lemma 1. Therefore, an optimal $(\text{RM}[v], \text{RM}[w])$ -inner-interpretation of peaks $1, \dots, i-1$ is given by an optimal $(\text{RM}[u], \text{RM}[w])$ -inner-interpretation. Thus $S[i][v][w] = S[i-1][u][w] + \delta(i, \iota)$. A similar argument holds when peak i is assigned to a prefix residue mass in R_i^R . \square

THEOREM 3.: *An optimum interpretation of a spectrum (M, S) utilizing any set of simple ion-types \mathcal{I} can be computed in time $O(|S|^3 |I|^3 \log(M))$.*

PROOF. The maximum size of the array RM is $|S||I|$. The number of iterations is given by $|S|(|S||I|)^2$. In each iteration, we consider each of the possible $|I|$ assignments of the peak. Finally, we need to ensure that any new mass differences form valid residue tags. This can be done by searching the precomputed and sorted set \mathcal{V}_M of size $O(M)$. \square

5. EXTENSIONS TO THE BASIC SCORING SCHEME

5.1 Suboptimal interpretations

The structure allows us to maintain the L best interpretations. For $1 \leq l \leq L$, let $S[l][i][v][w]$ denote the score of the l -th best interpretation. Let $\text{rank}_l(C)$ be the l -th largest value in a set C . Then

THEOREM 4.:

$$S[l][0][v][w] = \begin{cases} 0 & \text{if } \text{RM}[v] \leq \text{RM}[w], \\ & \text{RM}[v] \in \mathcal{V}_M, \\ & \text{RM}[w] - \text{RM}[v] \in \mathcal{V}_M, \\ & M - \text{RM}[w] \in \mathcal{V}_M, \\ & 1 \leq l \leq L; \\ -\infty & \text{otherwise.} \end{cases}$$

$$S[l][i][v][w] = \text{rank}_i \left(\begin{array}{l} S[l_1][i-1][v][w] + \delta(i, \phi), \\ 1 \leq l_1 \leq L; \\ S[l_1][i-1][u][w] + \delta(i, \iota), \\ 1 \leq l_1 \leq L, \\ \forall u \text{ s.t.} \\ \text{RM}[u] = r_i(\iota) \in R_i^L, \\ \text{RM}[u] \geq \text{RM}[v], \\ \text{RM}[u] - \text{RM}[v] \in \mathcal{V}_M, \\ \text{RM}[w] - \text{RM}[u] \in \mathcal{V}_M; \\ S[l_1][i-1][v][u] + \delta(i, \iota), \\ 1 \leq l_1 \leq L, \\ \forall u \text{ s.t.} \\ \text{RM}[u] = r_i(\iota) \in R_i^R, \\ \text{RM}[u] \leq \text{RM}[w], \\ \text{RM}[u] - \text{RM}[v] \in \mathcal{V}_M, \\ \text{RM}[w] - \text{RM}[u] \in \mathcal{V}_M. \end{array} \right)$$

A naive implementation of this requires computation on $L|S|(|S||I|)^2$ cells. Each computation requires iterating through all of $|I|$ ions in the list, and sorting and searching a list of $L|I|$ elements, as well as searching a list of $O(M)$ elements. Thus a simple implementation for the L best interpretations takes $O(L^2|S|^3|I|^4 \log(L|I|) \log M)$ time, and $O(L|S|^3|I|^2)$ space.

5.2 Forward Backward Scoring and Core Interpretations

Let the score function $S[i][v][w]$ be the *forward* score. We can define an analogous *backward* score $T[i][v][w]$ as the highest scoring $(\text{RM}[v], \text{RM}[w])$ -outer-interpretation of peaks $i+1, i+2, \dots, k$. It is not hard to see that

LEMMA 5.:

$$T[k+1][v][w] = \begin{cases} 0 & \text{if } \text{RM}[v] \leq \text{RM}[w], \\ & \text{RM}[v] \in \mathcal{V}_M, \\ & \text{RM}[w] - \text{RM}[v] \in \mathcal{V}_M, \\ & M - \text{RM}[w] \in \mathcal{V}_M; \\ -\infty & \text{otherwise.} \end{cases}$$

$$T[i][v][w] = \max \left\{ \begin{array}{l} T[i+1][v][w] + \delta(i, \phi); \\ T[i+1][u][w] + \delta(i, \iota), \\ \forall u \text{ s.t.} \\ \text{RM}[u] = r_i(\iota) \in R_i^L, \\ \text{RM}[u] \leq \text{RM}[v], \\ \text{RM}[u] \in \mathcal{V}_M, \\ \text{RM}[v] - \text{RM}[u] \in \mathcal{V}_M; \\ T[i+1][v][u] + \delta(i, \iota), \\ \forall u \text{ s.t.} \\ \text{RM}[u] = r_i(\iota) \in R_i^R, \\ \text{RM}[u] \geq \text{RM}[w], \\ \text{RM}[u] - \text{RM}[v] \in \mathcal{V}_M, \\ M - \text{RM}[u] \in \mathcal{V}_M. \end{array} \right.$$

The forward and backward computation allows us to explore the space of other suboptimal solutions to the problem. Also, as described below, it allows us to quantify the effect of assigning a peak to an ion.

LEMMA 6.: Consider i, u such that $\text{RM}[u] = r_i(\iota)$ for peak i , and some $\iota \in \mathcal{I}$. Define $H[i][u]$, as the highest scoring interpretation in which peak i is assigned to ι .

$$H[i][u] = \begin{cases} \delta(i, \iota) + \max_v (S[i-1][v][u] + T[i+1][v][u]) \\ \text{if } u \geq v; \\ \delta(i, \iota) + \max_w (S[i-1][u][w] + T[i+1][u][w]) \\ \text{if } u \leq w. \end{cases}$$

For peaks i in which there is exactly one u_m s.t. $H[i][u_m] \simeq S[m][0][n]$, and for all other u , $H[i][u] \ll S[m][0][n]$ are peaks that should have a *fixed interpretation*. Finding and assigning these peaks gives a core interpretation that is consistent with all optimum interpretations.

6. COMPUTATIONAL EXPERIMENTS

The ability of any *de novo* interpretation to successfully identify a peptide from its tandem mass spectrum depends critically on the spectrum. In order to resolve the peptide sequence down to individual amino-acids, the spectrum must contain at least one ion representing the cleavage of a chemical bond between each pair of adjacent amino-acids. The observation of such a complete ladder of ions is rare in practice. In the absence of additional information, such as a protein sequence database, a *de novo* interpretation cannot determine the position of peptide backbone bonds for which no ion is observed in the spectrum. However, tandem mass spectra often contain short ion ladders that represent some portion of the peptide sequence, termed sequence tags by Mann and Wilm [10]. *De novo* interpretation algorithms must be able to extract all the available information from a spectrum, including sequence tags if they are present, even when a complete ion ladder of ions is not.

The accuracy of the mass spectrometry instrument's measurement of ion mass-to-charge ratio is also very important for a successful *de novo* interpretation. When a protein sequence database is used to identify peptides from tandem mass spectra, each ion m/z measurement can be compared against its theoretical value. In a *de novo* interpretation, however, we must consider differences between two m/z measurements. If we are unlucky, it is possible for the difference between two putative prefix residual masses from the array RM to be as much as $(2+z)\epsilon$ from the true value, where z is the charge state of the tandem mass spectrum, and ϵ is the m/z measurement error of the instrument. The number of possible explanations for a m/z difference grows very quickly as this tolerance increases and significantly reduces the specificity of a *de novo* interpretation.

Finally, the presence of ions that are unexplained by the peptide fragmentation model can mislead *de novo* interpretation algorithms, particularly when ions that confirm the peptide identification are missing. The richness of the peptide fragmentation model supported by the *de novo* interpretation is a key factor here, since including additional ion-types in the interpretation can turn unexplained and misleading ions into ions that confirm a peptide identification. Our *de novo* interpretation model can accommodate any set of simple ion-types.

6.1 Random Artificial Spectrum Model

In order to evaluate how the performance of our *de novo* interpretation algorithm deteriorates as the spectrum becomes harder to interpret, we have constructed a random tandem mass spectrum model. The intention here is *not*

to generate random spectra that look like real spectra, but rather to generate random spectra that are potentially difficult for *de novo* interpretation algorithms to perform well on. We will study the performance of our algorithm as the (randomly generated) properties of the spectra deteriorate in the three directions outlined above.

We will base our random spectrum model on the default parameters used by the SEQUEST [6] sequence database search software to identify peptides from tandem mass spectra. SEQUEST measures the correlation between the observed spectrum and an artificial spectrum formed by applying a simple fragmentation model to the putative peptide sequence from the sequence database.

Given a peptide, we construct a synthetic tandem mass spectrum with parent mass corresponding to the peptide; charge state 2; b and y ions with intensity 1.0; a ions with intensity 0.5; neutral water loss b and y ions, when the resulting fragment contains an acidic residue (STDE) with intensity 0.2; neutral ammonia loss b and y ions, when the resulting fragment contains a basic residue (KQR) with intensity 0.2; a parent ion in charge state 2 with intensity 0.1; neutral water loss parent ion in charge state 2 if the peptide contains an acidic residue with intensity 0.05; neutral ammonia loss parent ion in charge state 2 if the peptide contains a basic residue with intensity 0.05; and doubly charged y ions with intensity 0.1.

Our random tandem mass spectra are generated from this synthetic spectrum \hat{S} according to two parameters, ε , the m/z measurement error, and γ , the chance that we observe an ion with intensity 1.0. For each ion of \hat{S} , we independently sample a m/z measurement error uniformly at random in the range $[-\varepsilon, \varepsilon]$. Further, we independently at random decide to keep each ion with intensity i from \hat{S} with probability $\min\{\gamma i, 1\}$. The parent mass m/z measurement is also perturbed by a random m/z measurement error sampled uniformly at random from the range $[-\varepsilon, \varepsilon]$.

6.2 Implementation

We have implemented the forward and backward scoring *de novo* interpretation dynamic programs and the core interpretation algorithm. We have used a very unsophisticated ion-type assignment score function, with $\delta(i, \iota)$ equal to the intensity of ion i . When an ion is not assigned to any ion-type, $\delta(i, \phi) = 0$. We construct our set of valid residue tags \mathcal{V}_M for residue tags up to 300 Daltons. Residue tags greater than 300 Daltons are considered to be valid outright.

6.3 Experiment Specifics

We will use the human serum albumin peptide LVNEVTEFAK in charge state 2 as the basis for our random tandem mass spectra. The measurement error parameter ε is set to either 0.1 or 0.2 Daltons. When $\varepsilon = 0.1$, we use a worst case tolerance of 0.4 Daltons when comparing m/z differences against theoretical values. When $\varepsilon = 0.2$, we use a tolerance of 0.5 Daltons, which is smaller than the worst case tolerance of 0.8 Daltons. In this case, the measurement error and our choice of tolerance will hinder our interpretation of the random spectrum.

The chance of observing an ion with intensity i , is $\min\{\gamma i, 1\}$, with γ taking on the values $\{2, 1, \frac{2}{3}, \frac{1}{2}, \frac{2}{5}, \frac{1}{3}, \frac{2}{7}\}$. We run our *de novo* interpretation algorithm on 100 random spectra generated with each pair of (γ, ε) parameter values.

We will not show the effect of γ directly, but instead show

the performance of our *de novo* interpretation against the average number of ions of the random spectrum that define each inter-amino-acid position along the peptide. For example, under our random spectrum model above, with γ set to 1, we expect at least 2 and potentially as many as 5 ions representing each inter-amino-acid position along the peptide. With γ set to $\frac{2}{5}$, we expect approximately one ion to represent each inter-amino-acid position.

Figure 3(a) shows the percentage of random spectra, restricted to those for which *all* inter-amino-acid positions were represented by at least one ion, in which the *de novo* interpretation algorithm was able to return the interpretation (I|L)V(N|[2G])EVTEFA(Q|[AG]|K). Note that this interpretation is as precise as the amino-acid residue masses permit, for this peptide. Each point in Figure 3(a) represents at least 10 random spectra.

Figure 3(b) shows the average percentage of the inter-amino-acid positions defined in each random spectrum that the *de novo* interpretation successfully identifies. Again, this is plotted against the average number of ions defining each inter-amino-acid position. Each point in Figure 3(b) represents the average proportion of positions identified over at least 10 random spectra.

Figure 3(c) shows the average percentage of the total intensity in the random spectrum that is assigned an explanation by the *de novo* interpretation algorithm. Note that we do not require that the ion-types assigned be consistent with the peptide from which the spectrum was generated. Each point in Figure 3(c) represents the average proportion of the total intensity explained over at least 10 random spectra.

Careful consideration of Figures 3(a), 3(b), and 3(c) suggest that as the number of ions defining each inter-amino-acid position decreases, many different ion-type assignments explain a similar proportion of the total intensity of the spectrum. This reinforces the notion that a *de novo* interpretation can recover only as much information as is present in the spectrum, and no more. However, core interpretations can help us understand when the spectrum supports multiple, equally plausible, explanations for a peak.

Figures 4 and 5 shows how core interpretations can give us confidence in the ion-type assignments of our *de novo* algorithm. In this figure we plot the histogram of the difference between the score of the optimal ion-type assignment and the next best sub-optimal ion-type assignment for a peak. Score differences are normalized by the score of the optimal assignment. Figure 4 and 5 clearly demonstrates that when the *de novo* interpretation assigns the correct ion-type to an ion, the difference between the optimal score and the next best sub-optimal score will usually indicate that this is the case. On the other hand, when there are many equally good ways to interpret the spectrum, in which case the assigned ion-type may well be incorrect, the difference between the optimal score and the next best sub-optimal score is small.

7. CONCLUSIONS AND FUTURE WORK

The correct interpretation of tandem mass spectra is a difficult problem, even when it is limited to scoring peptides against a database. *De novo* interpretation is a much harder problem, and available commercial software is usually not considered adequate for practical applications. Indeed many peptides fragment only partially so that the resulting spectrum is no longer a unique fingerprint for the peptide. Thus, even the definition of interpretation as the identifica-

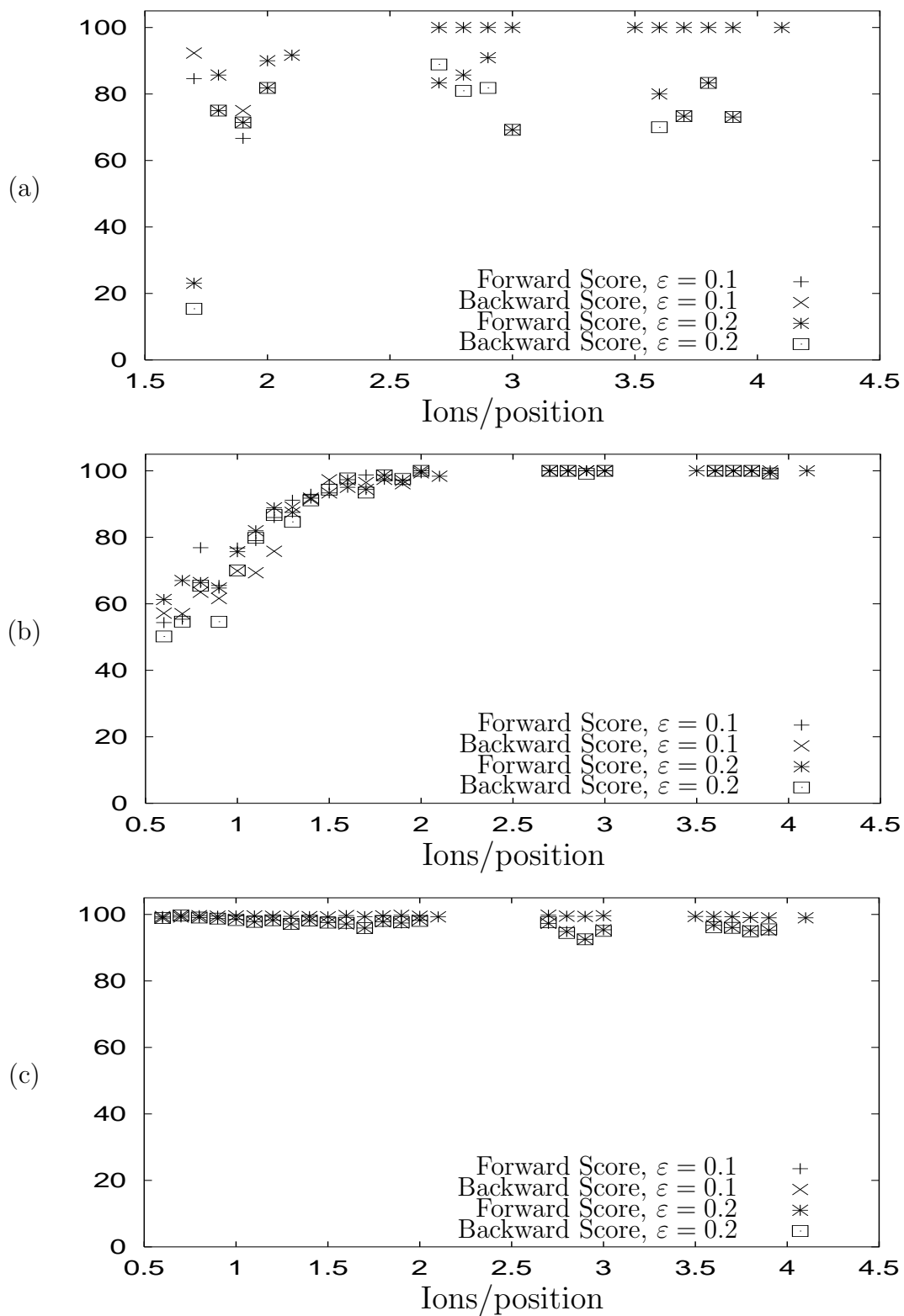


Figure 3: (a) Percentage of random spectra with correct peptide interpretation, as a function of the average number of ions defining each inter-amino-acid position; (b) Average percentage of positions correctly identified, as a function of the average number of ions defining each inter-amino-acid position; (c) Average percentage of total intensity explained, as a function of the average number of ions defining each inter-amino-acid position.

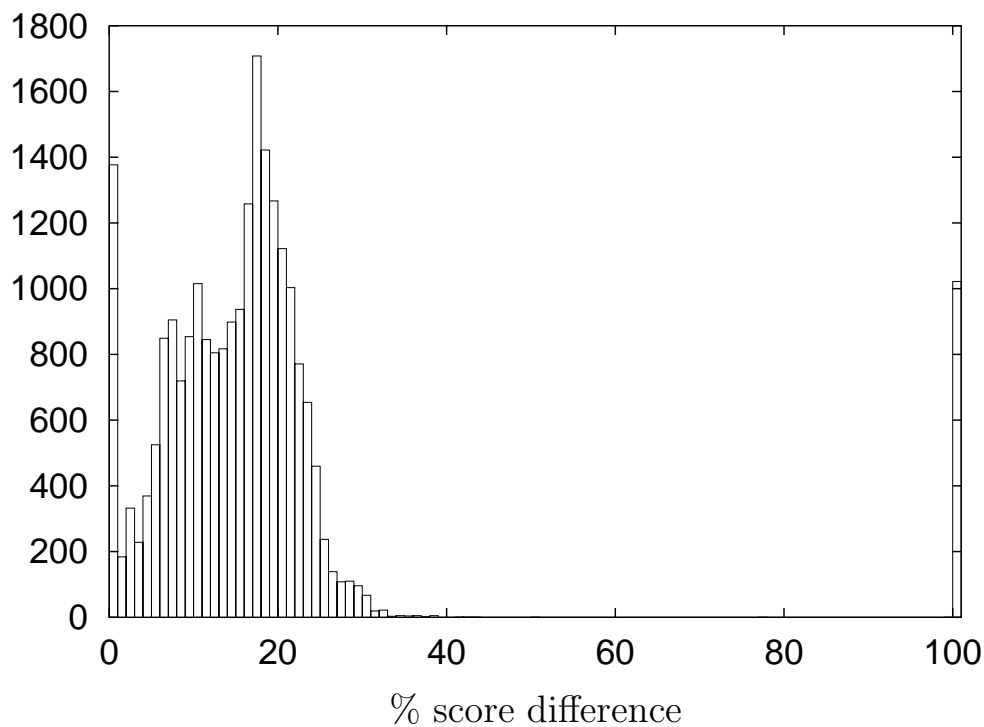


Figure 4: Histogram of percent relative difference between the score of the core ion-type assignment and next best sub-optimal assignment for correct ion-type assignments.

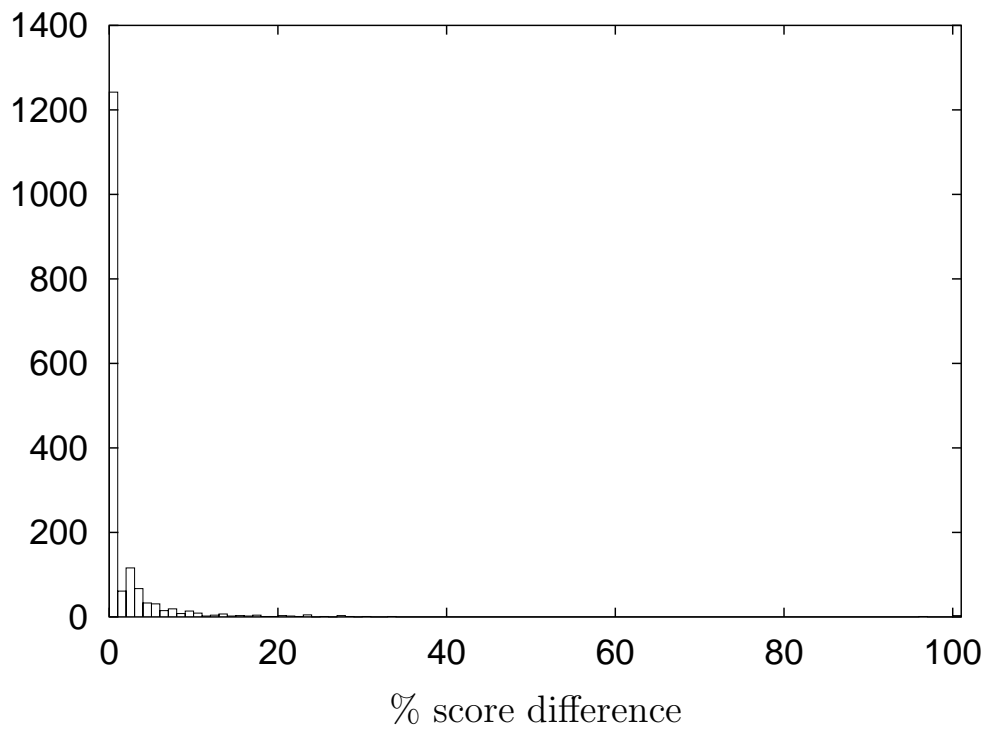


Figure 5: Histogram of percent relative difference between the score of the core ion-type assignment and next best sub-optimal assignment for incorrect ion-type assignments.

tion of the correct peptide is misguided, or not universally applicable. We shift the emphasis to scoring an interpretation that assigns ion-types to peaks. Our method, building upon earlier work of Dancik et al. [4], and Chen et al. [3], provides a dynamic programming technique to find the optimum interpretations under many reasonable score functions. We handle most of the possible ion types, including *a* ions, and neutral losses, but not internal fragments that might be seen in high energy CID spectra. Extending our approach to allow assignment of internal ions is an important open problem.

Another issue is the charge on the ions. In the case of an ESI source, the parent ion and many fragments may have multiple charge units assigned to them. Our approach can only deal with singly charged fragments. However, for many peaks of significant intensity, the charge can be deduced by looking at the spacing between isotopic neighbors, and an appropriately de-charged peak can be used for *de novo* interpretation.

The dynamic programming machinery allows us to explore other, sub-optimal solutions. Following standard techniques not described here, we can sample from the space of almost optimal solutions to get a probability of correct interpretation. We also have the notion of core-interpretations, which allows us to quantify the correctness of specific peak assignments. In particular, if the global score obtained from an optimal peak assignment is significantly greater than the score from alternative peak assignments, we have confidence in the specific peak assignment, even when a complete interpretation is of dubious quality. Thus, a core-interpretation allows us to gain information from poorly fragmented spectra.

As an additional note, the mass spectrometry community does not yet have a database of curated spectra of varying quality, that one can use to test algorithms. In the absence of such data sets, we take the first step towards *simulating* data sets by theoretically fragmenting known peptides. While this approach is not novel, and our simulator is admittedly naive, this is the first approach to studying the performance of *de novo* algorithms as a function of spectral quality (fragmentation probability and measurement error). The development of a good simulator with agreed-to parameters will go a long way in aiding the comparison and development of *de novo* techniques for interpreting tandem mass spectra.

8. REFERENCES

- [1] V. Bafna and N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17 Suppl 1:S13–21, June 2001. Appeared in Intl. Conference on Intelligent Systems for Molecular Biology.
- [2] C. Bartels. Fast algorithm for peptide sequencing by mass spectrometry. *Biomedical and Environmental Mass Spectrometry*, 19:363–368, 1990.
- [3] T. Chen, M. Y. Kao, M. Tepel, J. Rush, and G.M. Church. A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 8(6):571–83, 2001.
- [4] V. Dancik, T. Addona, K. Clauser, J. Vath, and P.A. Pevzner. *De novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6:327–342, 1999.
- [5] J. Fernández de Cossio, J. Gonzales, and V. Besada. Protein identification using mass spectrometric information. *Comput. Appl. Biosci.*, 11:427–434, 1995.
- [6] J. Eng, A. McCormack, and J. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of American Society of Mass Spectrometry*, 5:976–989, 1994.
- [7] D. Fenyo, J. Qin, and B.T. Chait. Protein identification using mass spectrometric information. *Electrophoresis*, 19(6):998–1005, 1998.
- [8] R.J. Johnson and K. Biemann. Computer program (seqpep) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. *Biomedical and Environmental Mass Spectrometry*, 18:945–957, 1989.
- [9] D.J. Lipman and W.R. Pearson. Rapid and sensitive protein similarity searches. *Science*, 227:1435–1441, 1985.
- [10] M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66:4390–4399, 1994.
- [11] P. A. Pevzner. *Computational Molecular Biology: An Algorithmic Approach*. MIT Press, 2000.
- [12] P.A. Pevzner, V. Dancik, and C.L. Tang. Mutation-tolerant protein identification by mass-spectrometry. In R. Shamir, S. Miyano, S. Istrail, P.A. Pevzner, and M.S. Waterman, editors, *International Conference on Computational Molecular Biology (RECOMB)*, pages 231–236. ACM Press, 2000.
- [13] J.A. Taylor and R.S. Johnson. Sequence database searches via *de novo* peptide sequencing by mass spectrometry. *Rapid Communications in Mass Spectrometry*, 11:1067–1075, 1997.