



SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database

Vineet Bafna¹ and Nathan Edwards¹

¹Informatics Research, Celera Genomics, 45 W. Gude Drive, Rockville, MD, 20850, USA

ABSTRACT

Proteomics, or the direct analysis of the expressed protein components of a cell, is critical to our understanding of cellular biological processes in normal and diseased tissue. A key requirement for its success is the ability to identify proteins in complex mixtures. Recent technological advances in tandem mass spectrometry has made it the method of choice for high-throughput identification of proteins. Unfortunately, the software for unambiguously identifying peptide sequences has not kept pace with the recent hardware improvements in mass spectrometry instruments. Critical for reliable high-throughput protein identification, *scoring* functions evaluate the quality of a match between experimental spectra and a database peptide. Current scoring function technology relies heavily on ad-hoc parameterization and manual curation by experienced mass spectrometrists. In this work, we propose a two-stage stochastic model for the observed MS/MS spectrum, given a peptide. Our model explicitly incorporates fragment ion probabilities, noisy spectra, and instrument measurement error. We describe how to compute this probability based score efficiently, using a dynamic programming technique. A prototype implementation demonstrates the effectiveness of the model.

Contact: Vineet.Bafna@Celera.Com

INTRODUCTION

Proteomics, or the direct analysis of the expressed protein components of a cell, is critical to our understanding of cellular biological processes. A comparison of the expressed proteins in normal versus diseased tissue can provide key insights into the action and effects of a disease. The proteins that are differentially expressed are very likely to contain diagnostic markers and protein targets for therapeutic intervention by drugs. In addition, the identification of all the components of a protein complex can help elucidate biochemical processes such as transcription and translation.

A key requirement for the success of proteomics is the ability to identify proteins in complex mixtures. Consequently, mass spectrometry (MS), particularly *tandem*

mass spectrometry, is rapidly becoming the method of choice for the high-throughput identification of proteins.

Mass Spectrometry

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 series 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 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 physiochemical 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 N-terminal a_1, b_1, c_1 fragments, and the C-terminal x_{n-1}, y_{n-1} , and z_{n-1} fragments are shown. While a, b, y represent the commonly occurring fragments, a high energy collision often results in other fragments, including *internal* fragments formed by breakage at two points, 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 difference 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

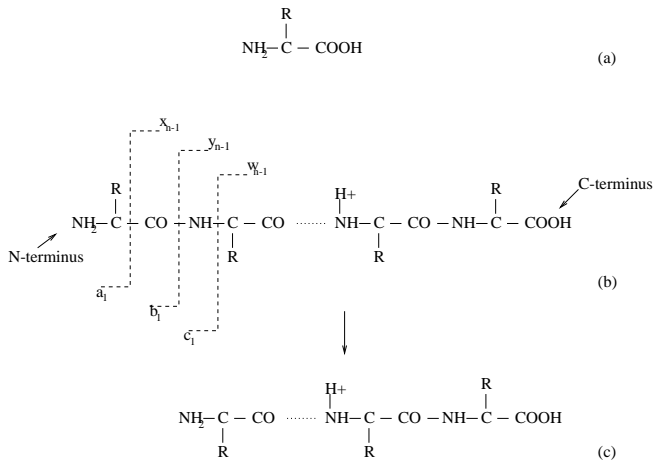


Fig. 1. (a) The structure of an amino-acid. (b) An ionized peptide. (c) y_{n-1}^+ ion

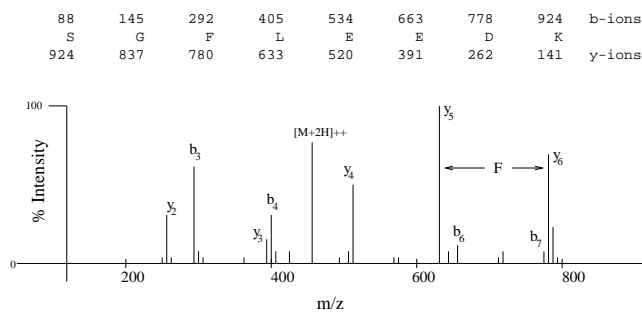


Fig. 2. MS/MS spectrum for peptide SGFLEEDK.

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 , ...), 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 out work in context.

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.

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 Bartels (1990); Dancik et al. (1999); de Cossio et al. (1995); Johnson and Biemann (1989); Taylor and Johnson (1997)). They perform best when the spectra have relatively complete ladders and little noise. On the other hand, so-called *database-searching* algorithms Eng et al. (1994); Fenyo et al. (1998); Mann and Wilm (1994); Pevzner et al. (2000) 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 (1997) propose combining their *de-novo* sequencing with Fasta (Lipman and Pearson (1985)) style database scoring. On the other hand, Mann and Wilm (1994) report on the effectiveness of generating sequence tags for effective scoring. Other database searching programs like Sequest (Eng et al. (1994)) 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. With the onset of the genomic sequence, and improvement in coverage and quality of databases, a pure *de-novo* interpretation is not required. Good scoring is essential for eliminating false positives, while aggressive interpretation and filtering help reduce the number of candidate peptides that must be scored making the search more efficient.

This paper focuses on scoring. A good scoring module is the mainstay of all database searching algorithms. Earlier work on this problem primarily involved the notion of shared peak count, or a count of the number of spectral peaks that could be assigned to fragments of the candidate peptide. Eng et al. (1994) introduce the notion of discrete correlation, which corrects this score by subtracting the mean of a cross-correlation function. Perkins et al. (1997), and Qin et al. (1997) introduce the notion of a p -value for the score based on theoretical and empirical considerations. Pevzner et al. (2000) establish that the shared peak count is not a reliable measure in the presence of mutations or chemical modifications. They introduce the notion of spectral alignment and use it for mutation and modification tolerant scoring.

Despite these efforts, scoring remains an inexact science, even for the case of no mutations. Published accounts of the success rate in identifying peptides from

MS/MS spectra remain quite low, and often require confirmation by a human operator. A human operator as the final step is the only way these scoring functions can incorporate detailed knowledge of the physiochemical properties of a peptide, such as which fragments are likely to form for a peptide in a particular type of instrument. Consider for example, *neutral losses*. Experienced mass spectrometers know that the peptide must contain one of the acidic amino-acids S , T , D , or E for an H_2O loss to be observed, whereas the loss of NH_3 occurs mainly in the presence of a basic amino-acid residue R , K and Q . Thus, a spectral peak identified as a $y - H_2O$ should be down-weighted if there are no acidic residues in the candidate peptide. Currently available score functions do not use these rules in a structured and quantitative manner. Further, existing score functions make no attempt to explicitly model instrument measurement error. A spectral peak is typically attributed to the fragment with mass-charge ratio that is 'nearest' within some predefined tolerance. Finally, real MS/MS spectra typically have a not insignificant number of 'noise' peaks, peaks for which no peptide fragmentation explanation exists. Current available score functions make no attempt to explicitly model noise peaks.

Our work seeks to address these issues. We model the process of MS/MS spectrum generation by a two-step stochastic process. The first step involves generation of fragments from a peptide, according to a probability distribution estimated from many training samples. The second step involves the generation of a spectrum from the fragments according to the distribution of the instrument measurement error. We present an algorithm based on dynamic programming that efficiently computes the probability of a specific spectrum being generated by a candidate peptide, assuming our stochastic process.

A STOCHASTIC MODEL FOR SPECTRUM GENERATION

In order to describe our model for the stochastic process of MS/MS spectrum generation by a peptide, we define some terms.

MS/MS Spectrum: A MS/MS spectrum $S \in \mathbf{R}_+^k$ is a vector of positive real numbers specifying the k observed mass-charge ratios of the spectral peaks.

Peptide: A peptide $p \in \mathcal{A}^n$ is a sequence of n amino-acid residues from the alphabet of amino-acid symbols, $\mathcal{A} = A, C, \dots, Y$.

Fragment Space: An enumeration $\mathcal{F}(p)$ of all fragment mass-charge ratios that a peptide p might produce. For example, for low energy MS/MS ionization technology, we typically consider single cleavages

with a, b and y ion types; and one, two or three units of charge to enumerate all possible mass-charge ratios that might be generated. Each element of $\mathcal{F}(p)$, then, is a fragment-charge state pair. Thus,

$$\mathcal{F}(p) = \{(a_1, i), (b_1, i), (y_1, i), \dots, (a_n, i), (b_n, i), (y_n, i), i = 1, 2, 3\}$$

For high energy ionization technologies, we must consider double cleavage and no cleavage events and more ion types. Fortunately, for this case, it is usually safe to consider only one unit of charge for each fragment. Denote the mass-charge ratio of a fragment $f \in \mathcal{F}(p)$ by $m/z(f)$.

Fragmentation Pattern: Given a peptide p , a fragmentation pattern $F \subseteq \mathcal{F}(p)$ is a subset of the fragment space representing the observed fragments generated from a peptide p .

Fragmentation Space: The fragmentation space $\phi(p)$ of a peptide p is the set of all fragmentation patterns of p . That is,

$$\phi(p) = \{F : F \subseteq \mathcal{F}(p)\}$$

Noise: We consider any peak of S for which $F(p)$ provides no explanation to be a noise peak.

The tandem mass spectrum is generated by the following two step random process.

Fragmentation: Each of the many copies of a peptide that pass into the secondary collision chamber fragments according to the experimental conditions and the nature of the peptide. Some of the generated fragments retain the charge carried by the parent peptide, “fly” to the particle detector, and have their mass-charge ratio measured. In addition to the modeled fragments in $\mathcal{F}(p)$, unexpected fragments or contaminant fragments might be observed. The experimental conditions in the collision chamber and the physical and chemical properties of a peptide govern the probability of observing a particular fragment. A fragmentation pattern F is the outcome of this random experiment on the sample space $\phi(p)$.

Measurement: Each fragment generated by a peptide must have its mass-charge ratio measured by the particle detector. Due to measurement error, each fragment with a particular mass-charge ratio generates a mass-charge ratio observation close to, but not precisely at its true mass-charge ratio. The observation of many fragments with the same mass-charge ratio leads to the formation of a

distinctive peak close to the true mass-charge ratio of these fragments. The observed peak can then be represented by a single real number, an estimate of the true mass-charge ratio of the fragments that generated it. The deviation of this mass-charge ratio of a peak from its true value is modeled according to a probability distribution, typically the normal distribution.

SCORING SPECTRA

We score a peptide by the probability that the observed spectrum S was generated by this peptide. Let $\psi(S|p)$ denote the probability density function for the random vector S representing the MS/MS spectrum, given peptide p . Typically, we are searching a database for the peptide p^* that satisfies

$$p^* = \arg \max_p \psi(S|p)$$

A formal description of our two-step model of fragmentation followed by measurement is given by

$$\psi(S|p) = \sum_{F \subseteq \mathcal{F}(p)} \psi(S|F, p) \Pr(F|p)$$

The quantity $\Pr(F|p)$ represents the probability of a particular fragmentation pattern of a peptide. It is in the computation of $\Pr(F|p)$ that the complex process of fragmentation can be modeled. Section **Fragmentation probability estimation**, we describe a process for estimating fragmentation probability. The quantity $\psi(S|F, p)$ represents the probability that the spectrum S was generated by the fragmentation pattern F of a peptide p . We describe a model for $\psi(S|F, p)$ in section **Spectrum generation by fragments**.

As shown, the computation of $\psi(S|p)$ involves the summation over the fragmentation space, an exponentially sized set. Fortunately, the probability that a particular fragmentation results in a particular spectrum is negligible for all but a small number of fragmentation patterns. Even so, the cost of evaluating $\psi(S|p)$ as part of a database search is prohibitive. In section **Spectrum generation by fragments**, we demonstrate how an understanding of the fragmentation process will allow us to place constraints on our stochastic model of spectrum generation and compute a good approximation for $\psi(S|p)$ quickly. We present an algorithm that computes an approximation of $\psi(S|p)$ in time $O(|\mathcal{F}(p)|k + k^2)$.

Fragmentation probability estimation

Scoring a peptide based on the likelihood of it producing the observed fragments permits the incorporation of the physical and chemical intuition of an expert operator and statistically significant observed phenomena. We

briefly sketch a framework within which we can estimate $\Pr(F|p)$.

Given a database of tandem MS spectra each with an expertly curated peptide identification, we can generate, for each peptide, a list of the observed fragments. From this, we can estimate the likelihood of each peptide backbone cleavage event merely by counting the number of times it is observed. However, we expect that some peptides will undergo some cleavage events significantly more or less often than occurs in general. By using clustering and data mining techniques on this peptide-observed fragment dataset, we will find those properties of a peptide which lead to significant changes in the likelihood of particular cleavage events.

For example, experienced operators know that the presence of acidic amino-acids in a peptide makes the neutral water loss ion type cleavages much more likely. We expect the above data mining approach to demonstrate this, and many other properties of a peptide which govern the likelihood of observing a particular cleavage event.

At this point in time, we have no such database of expertly curated spectra labeled with a peptide, so we must estimate these probabilities by some other means. For the preliminary results contained in this paper, we have chosen probabilities in consultation with experienced mass spectrometer operators.

Spectrum generation by fragments

In the previous section, we focused on computing $\Pr(F|p)$ based on estimated probabilities of fragment generation. In this section, we develop a model for computing $\psi(S|F, p)$, and use this model to compute $\psi(S|p)$ efficiently.

The probability density $\psi(S|F, p)$ describes the probability of observing a collection of spectral peaks given a particular fragmentation pattern. Unfortunately, it is not at all obvious which fragment(s) are responsible for which peak(s), and which peaks should be considered noise. In order to compute $\psi(S|F, p)$, we need to either sum over all the possible explanations of each peak, or use our understanding of the mass spectrometer to limit the number of terms. Without loss of generality, we assume that each fragment in the fragment space has a unique mass-charge ratio. This assumption can be lifted in a straightforward way, the section **Implementation details** explains how this can be done. We can now make the following observations and assumptions about the relationship between a fragmentation pattern and the observed peaks.

1. Each unique mass-charge ratio in the fragment space generates at most one spectral peak.
2. Each spectral peak is the observed mass-charge ratio of at most one of the (unique) mass-charge ratios in the fragment space.

3. The assignment of spectral peaks to fragments must be *non-crossing*. For all fragments f_1, f_2 and spectral peaks S_1, S_2 , if $m/z(f_1) < m/z(f_2)$ and $S_1 < S_2$, then peak S_1 must have been generated by fragment f_1 and peak S_2 must have been generated by fragment f_2 .

In addition, we augment the fragment space $\mathcal{F}(p)$ with *noise fragments*, one for each spectral peak. Each noise fragment has the same mass-charge ratio as its spectral peak. We denote this augmented fragment space $\mathcal{F}'(p)$ and the corresponding fragmentation space $\phi'(p)$.

We assign zero probability mass to fragmentation patterns which violate these assumptions, and reconsider the computation of $\psi(S|F, p)$. We notice first that due to the addition of noise fragments to the fragment space, all spectral peaks must either be assigned to a unique fragment from our original fragment space $\mathcal{F}(p)$ or to a noise fragment. Therefore, only fragmentation patterns $F \subseteq \mathcal{F}'(p)$ with $|F| = k$ have non-zero probability mass.

However, we can say something even stronger. Let $S_i \stackrel{M}{=} f$ denote the event that peak S_i is generated by fragment f , and $S = (S_1, S_2, \dots, S_k)$ be a tandem MS spectrum ordered by mass-charge ratio. Further, let $F \subseteq \mathcal{F}'(p), |F| = k$ be an arbitrary fragmentation pattern, whose observed fragments $f_1, f_2, \dots, f_k \in F$ are ordered by mass-charge ratio. The non-crossing and uniqueness assumptions guarantee that only one assignment of spectral peaks to fragments has non-zero probability mass. All of the probability mass for $\psi(S|F, p)$ is captured by this unique non-crossing assignment. Under this Bayesian notion of the valid assignments, then

$$\psi(S|F, p) = \psi(S \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p)$$

We now choose an appropriate function to model $\psi(S \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p)$. As a probability density function, ψ must satisfy

- $0 \leq \psi(S \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p) \leq 1$; and
- $\int_{S_1 < S_2 < \dots < S_k} \psi(S \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p) = 1$.

In addition, we require that

- $\psi(S \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p)$ is maximized when $S_1 = m/z(f_1), \dots, S_k = m/z(f_k)$ and decreases rapidly to 0 as each S_i moves away from $f_i, i = 1, \dots, k$; and
- $\psi(S \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p)$ can be computed efficiently.

In isolation, the distribution of one measured mass-charge ratio about its true value is independent of any other measured mass-charge ratio about its true value. We model

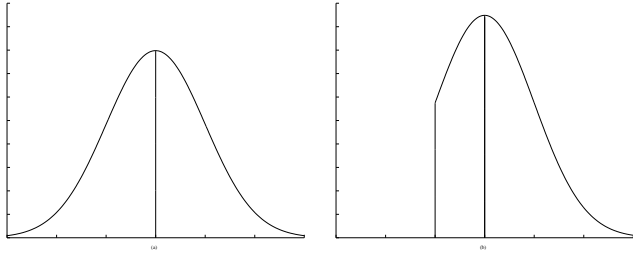


Fig. 3. (a) Normal measurement error distribution. (b) Truncated and scaled normal measurement error distribution.

the distribution of the measured mass-charge ratios as either normal or uniform distributions centered at the fragment mass-charge ratio. See Figure 3(a). On the other hand, we model the distribution of the measured mass-charge ratio of noise fragments by an impulse function at the mass-charge ratio of its spectral peak. In other words, only the spectral peak a noise fragment represents has a non-zero probability of generating it. We must expand $\psi(S | \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p)$ into its components in order to compute it.

$$\begin{aligned} & \psi(S | \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p) \\ &= \psi(S_1 | \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p) \times \\ & \quad \prod_{j=2}^k \psi(S_j | S_1, \dots, S_{j-1}, \cap_{i=1}^k [S_i \stackrel{M}{=} f_i], F, p) \\ &= \psi(S_1 | [S_1 \stackrel{M}{=} f_1], F, p) \times \\ & \quad \prod_{j=2}^k \psi(S_j | S_{j-1}, [S_j \stackrel{M}{=} f_j], F, p). \end{aligned}$$

All that remains is to specify the distribution of the measured mass-charge ratio S_j with respect to f_j , given that $S_j > S_{j-1}$. We model this by truncating the left-hand tail of the measurement distribution and rescaling its total probability density to one. See figure 3(b). Let ρ_f be the distribution of the observed peak about the true mass-charge ratio of fragment f . Then

$$\psi(S_1 | [S_1 \stackrel{M}{=} f_1], F, p) = \rho_{f_1}(S_1)$$

$$\psi(S_j | S_{j-1}, [S_j \stackrel{M}{=} f_j], F, p) = \begin{cases} \frac{\rho_{f_j}(S_j)}{\int_{S > S_{j-1}} \rho_{f_j}(S)}, & S_j > S_{j-1}; \\ 0, & \text{otherwise.} \end{cases}$$

Therefore, given our observations and assumptions, ψ satisfies all of our requirements for a probability density function, it concentrates its probability mass about the true mass-charge ratios, and it is efficiently computable. We must next show how this choice of ψ allows an efficient algorithm for computing $\psi(S|p)$.

An algorithm for computing $\psi(S|p)$

Having established an efficient method for computing $\psi(S|F, p)$, we must now ensure that we can compute $\psi(S|p) = \sum_{F \subseteq \mathcal{F}(p)} \psi(S|F, p) \Pr(F|p)$ without summing an exponential number of terms. In order to do this, we must impose a stringent assumption — for each fragment f , the probability of observing f may not be dependent on the event that some other fragment is observed. The probability of observing f may depend on peptide properties and experimental conditions, but must be independent of the probability of observing other fragments. Our own experience suggests that this is not, in general, true. Without this assumption, we cannot compute ψ efficiently. Experiments suggests that even with this assumption, enough of the stochastic behavior of fragmentation is captured. With the independence of fragment generation in hand, we can now describe how to compute $\psi(S|F, p)$ using dynamic programming.

As before, consider the tandem MS spectrum $S = (S_1, \dots, S_k)$ and the fragments $\mathcal{F}'(p) = \{f_1, \dots, f_m\}$ to be ordered by mass-charge ratio. Define $\mathcal{F}'_j(p) = \{f_1, \dots, f_j\}$ to be the first j fragments of $\mathcal{F}'(p)$. Then the dynamic programming recurrence function $\Phi(i, j)$ represents the probability mass associated with the event that the first i peaks were generated by i fragments from the first j fragments of $\mathcal{F}'(p)$. Formally,

$$\Phi(i, j) = \sum_{F \subseteq \mathcal{F}'_j(p)} \psi(S_1, S_2, \dots, S_i | F, p) \Pr(F|p)$$

Clearly, $\Phi(k, m) = \psi(S|p)$ is the value we are interested in. The following recurrence holds:

$$\Phi(i, j) = \begin{cases} 1, & \text{if } i = 0, \\ 0, & \text{if } i > j, \\ \Phi(i-1, j-1) \\ \quad \times \psi(S_i | S_{i-1}, S_i \stackrel{M}{=} f_j) \\ \quad \times \Pr(f_j|p) \\ \quad + \Phi(i, j-1) \Pr(\bar{f}_j|p), & \text{otherwise.} \end{cases}$$

Therefore, $\psi(S|p)$ can be computed in time $O(k|\mathcal{F}'(p)|)$. This computation sums over all possible fragment assignments simultaneously. The most likely assignment F^* is given by

$$F^* = \arg \max_{F \subseteq \mathcal{F}'(p)} \psi(S|F, p) \Pr(F|p)$$

It is easy to see how this assignment can be computed in the same time, using a max in place of a sum in the dynamic programming recurrence. A score function related to the maximum likelihood computation is

$$\psi'(S|p) = \max_{F \subseteq \mathcal{F}'(p)} \psi(S|F, p) \Pr(F|p)$$

In practice, $\psi(S|p)$ and $\psi'(S|p)$ produce values that are very similar. Almost all of the probability mass of $\psi(S|p)$ is contained in the most likely fragmentation pattern. The big advantage of computing $\psi'(S|p)$ instead of $\psi(S|p)$ is that we can take logarithms of intermediate values to avoid numerical instability in the computation of $\Phi(k, m)$.

Implementation details

The preceding theoretical description of scoring glossed over several important details. Most importantly, there is the issue of many fragments with the same theoretical mass. Second, this algorithm is only as good as the probabilities assumed for each fragment of a peptide and the presumed shape of the distribution of measured mass-charge ratio about the true value. Lastly, we must convert the above probability density into a suitable function for scoring spectra against a peptide database.

The first issue is easily solved. Note that any of the fragments with a particular mass is sufficient for the observation of a peak at a particular mass-charge ratio. The event that this mass is represented is the complement of the event that none of the possible fragments with this theoretical mass are represented.

Secondly, as discussed in Section **Fragmentation probability estimation** we will, in time, use a human curated database of identified spectral to compute empirical estimates of the probabilities this model requires. Until this database is available we have chosen these parameters in consultation with experienced mass spectrometer operators.

Lastly, as set out in Section **Scoring Spectra**, the function ψ is a continuous probability density function, and consequently isn't appropriate as a scoring function. In particular, since the observation of any particular spectrum is a zero probability event, we must decide how ψ will be used to score a peptide. Instead of considering a tandem MS spectrum as a vector of reals, we now consider each spectral peak observation to have a small width, related to the accuracy with which the mass spectrometry instrument can measure mass-charge ratios. This allows us to compute the probability of observing this spectrum as a result of the fragmentation of a particular peptide. In fact, for primarily aesthetic reasons, if this probability computation produces a probability p , we report a score of $-\log(p)$.

The algorithm of Section **Scoring Spectra** was implemented in C++ using the LEDA class library and was compiled on Digital True64 and RedHat Linux.

RESULTS AND DISCUSSION

In order to evaluate this scoring schema we have obtained a set of tandem MS spectra from the prototype Applied Biosystems MALDI-TOF-TOF instrument, in partnership

with the Proteomics Research Center in Framingham, MA. Each set of MS/MS spectra is generated from a tryptic digest of a single protein and can therefore be labeled with the protein undergoing fragmentation. The test spectra were generated from bovine serum albumin, enolase, carbonic anhydrase, and human serum albumin. A subset (13/41) of the provided spectra have been discarded due to exceptionally low signal-to-noise ratio. As mentioned earlier, fragmentation probabilities differ widely depending on the nature of the instrument. The MALDI ionization process usually results in singly charged peptides. Also, relatively high energy of collision makes internal fragments more likely.

We must evaluate the scoring schema not only on whether or not it correctly ranks the known peptide as the most likely to have generated the given spectrum, but also on how discriminating it is. If the top two scoring peptides have almost identical scores, then we have no way to choose between them. A database of peptides was generated by a tryptic digest of each of the 607674 sequences in the NCBI non-redundant protein database. The total number of peptides obtained by tryptic digestion was 21489237. Each experimental spectrum was scored against its true peptide sequence, as well as a filtered set of peptides whose parent mass matched the given parent mass to within 2 Daltons.

Table 1 summarizes the results on the 28 test spectra used to evaluate the scoring schema. The table presents the score of the correct peptide, the correct peptide's rank with respect to the NR protein database peptides, and two empirical measures of how different the score of the correct peptide is from the other top scoring peptides. Each p -value represents the chance that a peptide selected at random from the top 20 ranking peptides could have scored as well as the correct peptide. The first p -value assumes that the top 20 scores are normally distributed, while the second, more conservatively, uses the non-parametric Chebyshev inequality. An asterisk ('*') in the rank column indicates that the correct protein was not in the top 10 scoring peptides. Where the correct peptide was not ranked first, we list the p -values associated with the top ranking peptide from the database.

The results of the scoring schema on the set of test spectra clearly demonstrate the effectiveness of the probability model. Not only does the correct peptide achieve the best score for all but two peptides, but it is also well separated from the other incorrect peptides. The correct peptides are significantly different from the other top scoring peptides. In addition, the score of the incorrectly identified peptides is *not* significantly different from the other top scoring peptides. In other words, both correct *and* incorrect identifications are clearly indicated by the significance test. Further research into a definitive significance test remains to be done.

Table 1. Test spectra scores with respect to the NCBI protein database.

Peptide	Score	Rank	Normal p -value	Cheb. p -value
FKDLGEEHFK	85.87	1	$0.0E + 00$	$1.3E - 02$
YLYEIAR	21.91	1	$1.3E - 06$	$4.5E - 02$
AVVQDPALKPALVYGEATSR	50.05	*	$5.0E - 03^*$	$1.5E - 01^*$
HNGPEHWHK	35.05	1	$0.0E + 00$	$8.4E - 03$
HWHKDFPIANGE	45.93	1	$0.0E + 00$	$5.2E - 03$
LLMLANWRPAQPL	104.4	1	$0.0E + 00$	$4.9E - 03$
NWRPAQPL	78.62	1	$0.0E + 00$	$7.6E - 03$
NWRPAQPLKNR	47.74	1	$0.0E + 00$	$1.1E - 02$
RLVQFHFHWGSSDDQGSE	96.87	1	$4.1E - 06$	$5.0E - 02$
SHHWGYGKHNGPE	57.71	1	$1.1E - 06$	$4.5E - 02$
TKAVVQDPALKPALVYGE	42.78	1	$0.0E + 00$	$1.4E - 02$
YAAELHLVHWNTK	75.04	1	$0.0E + 00$	$1.0E - 02$
AVSKVYARSVYDSRGNPTVE	96.62	*	$6.4E - 03^*$	$1.6E - 01^*$
FFKDGKYD	108.5	1	$0.0E + 00$	$4.9E - 03$
FMIAPTGAKTFAE	47.85	1	$0.0E + 00$	$1.6E - 03$
IEEELGDNAVFAGENFHHGDK	46.96	1	$1.5E - 08$	$3.3E - 02$
KGVFRSIVPSG	89.26	1	$0.0E + 00$	$6.0E - 03$
KNVPLYKHLAD	102.5	1	$0.0E + 00$	$2.9E - 03$
LTKKRYG	56.11	1	$1.1E - 11$	$2.2E - 02$
LTVTNPKRIATAIE	116.1	1	$0.0E + 00$	$8.1E - 03$
LVVGLRTGQIKTG	95.06	1	$4.1E - 07$	$4.1E - 02$
NFHHGDKL	20.05	1	$2.0E - 14$	$1.7E - 02$
RLAKLNQLLRIEEE	73.96	1	$3.1E - 11$	$2.3E - 02$
SFAAGWGVMSVSHRSGETE	189.7	1	$0.0E + 00$	$2.2E - 03$
DVFLGMFLYFYAR	49.80	1	$0.0E + 00$	$4.6E - 03$
HPDYSVVLRLR	74.51	1	$0.0E + 00$	$1.2E - 02$
HPYFYAPELLFFAK	33.20	1	$0.0E + 00$	$1.2E - 02$
VFDEFKPLVEEPQNLIK	44.21	1	$0.0E + 00$	$4.7E - 03$

In concluding, we note again that this work represents research in progress. The scoring module will be part of a larger package with novel algorithms for interpretation of MS/MS data, and smart filtering of databases. A future course of action will be to incorporate the stochastic model presented here into a *de-novo* interpretation module. Another interesting extension will be to extend the scoring to mutation, and modification tolerant database searching. However, the true test of these ideas will be in their applicability to a proteomics pipeline and their role in discovering novel diagnostics and therapeutics of the 21st century.

ACKNOWLEDGMENTS

The authors would like to thank Peter Juhasz, Dale Patterson of the Applied Biosystems Proteomics Research Center as well as Rob Christian, Scott Patterson and other

members of the proteomics team at Celera for numerous discussions and helpful suggestions. The authors are also grateful to the PRC for generously supplying test spectra.

REFERENCES

- Bartels, C. (1990). Fast algorithm for peptide sequencing by mass spectrometry. *Biomedical and Environmental Mass Spectrometry* 19, 363–368.
- Dancik, V., T. Addona, K. Clauser, J. Vath, and P. Pevzner (1999). De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology* 6, 327–342.
- de Cossio, J. F., J. Gonzales, and V. Besada (1995). Protein identification using mass spectrometric information. *Comput. Appl. Biosci.* 11, 427–434.
- Eng, J., A. McCormack, and J. Yates (1994). 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.

-
- Fenyo, D., J. Qin, and B. Chait (1998). Protein identification using mass spectrometric information. *Electrophoresis* 19(6), 998–1005.
- Johnson, R. and K. Biemann (1989). 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.
- Lipman, D. and W. Pearson (1985). Rapid and sensitive protein similarity searches. *Science* 227, 1435–1441.
- Mann, M. and M. Wilm (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry* 66, 4390–4399.
- Perkins, D., D. Pappin, D. Creasy, and J. Cottrell (1997). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18), 3551–3567.
- Pevzner, P., V. Dancik, and C. Tang (2000). Mutation-tolerant protein identification by mass-spectrometry. In R. Shamir, S. Miyano, S. Istrail, P. Pevzner, and M. Waterman (Eds.), *International Conference on Computational Molecular Biology (RECOMB)*, pp. 231–236. ACM Press.
- Qin, J., D. Fenyo, Y. Zhao, W. Hall, D. Chao, C. Wilson, and R. Young (1997). A strategy for high-confidence protein identification. *Analytical Chemistry* 69, 3995–4001.
- Taylor, J. and R. Johnson (1997). Sequence database searches via *de novo* peptide sequencing by mass spectrometry. *Rapid Communications in Mass Spectrometry* 11, 1067–1075.