Published in final edited form as:

J Proteome Res. 2009 May ; 8(5): 2226–2240. doi:10.1021/pr800677f.

Predicting Intensity Ranks of Peptide Fragment Ions

Ari M. Frank

Department of Computer Science and Engineering, University of California, San Diego. 9500 Gilman Drive, Mail Code 0404 La Jolla, CA 92093-0404, USA

Abstract

Accurate modeling of peptide fragmentation is necessary for the development of robust scoring functions for peptide-spectrum matches, which are the cornerstone of MS/MS-based identification algorithms. Unfortunately, peptide fragmentation is a complex process that can involve several competing chemical pathways, which makes it difficult to develop generative probabilistic models that describe it accurately. However, the vast amounts of MS/MS data being generated now make it possible to use data-driven machine learning methods to develop discriminative ranking-based models that predict the intensity ranks of a peptide's fragment ions. We use simple sequence-based features that get combined by a boosting algorithm in to models that make peak rank predictions with high accuracy. In an accompanying manuscript, we demonstrate how these prediction models are used to significantly improve the performance of peptide identification algorithms. The models can also be useful in the design of optimal MRM transitions, in cases where there is insufficient experimental data to guide the peak selection process. The prediction algorithm can also be run independently through PepNovo+, which is available for download from <http://bix.ucsd.edu/Software/PepNovo.html>.

Keywords

MS/MS; peptide; fragmentation; prediction; machine learning; ranking; boosting; MRM

Introduction

Tandem mass spectrometry (MS/MS) has emerged as the leading method for high-throughput protein identification and quantitation in proteomics.^{1,2} Recent technological breakthroughs have led to a dramatic increase in the volume of MS/MS data being generated. While analyzing this growing amount of data encounters computational bottlenecks – which need to be met with ever-improving algorithms – this surplus of the data also opens a window of opportunity: it allows us to train more sophisticated and powerful models to be used in the analysis.

Many traditional statistical analysis methods rely on creating probabilistic models that describe the process being studied. However, except for very elementary cases, such models tend to over-simplify the phenomenon they describe and are consequently inaccurate. The tremendous growth witnessed in recent years, both in the volume of experimental data that is collected and in the computational resources made available for its processing, have given rise to new machine learning-based analysis approaches. These new computational methods do not require us to explicitly understand the underlying mechanism behind the processes being modeled. Rather, they rely on finding patterns and correlations concealed in the data and then utilize

them to draw insightful conclusions (this computational idea is at the heart of novel methods like metagenomics³).

In this paper we explore how a machine learning boosting algorithm⁴ (in the context of ranking⁵) can be used to leverage large amounts of experimental MS/MS data to create models for predicting the intensity ranks of a peptide's fragment ion peaks based only on the peptide's amino acid sequence. Using this data-driven machine learning method, we are able to create powerful predictive models without having to fully understand the complex dynamics of peptide fragmentation. Our models are constructed from simple ("weak") sequence-based features that get combined by the boosting algorithm into a strong accurate prediction models. In addition, examining the values of the parameters in the trained models can gain insights on the dynamics of fragmentation pathways.

What makes our method effective, and is also an underlying assumption in all other machine learning-based MS/MS sequencing algorithms, is the fact that MS/MS spectra are reproducible. This makes them suitable targets for prediction-based algorithms. The reproducibility of mass spectra stems from the fact that they are generated by recording the stochastic fragmentation process of many molecules. Thus, repeated experiments with the same peptides produce similar looking spectra.

The reproducible nature of mass spectra makes it possible to use simple but powerful comparison-based methods like spectral libraries for peptide identification.^{6,7} While this method is accurate, it is not suitable for identifying many less common peptides (post-translationally modified peptides, products of miscleavages, peptides from rarely expressed proteins or alternative splice variants, etc.) To be able to identify all spectra using a similarity based method, one would have to be able to predict a theoretical mass spectrum for any peptide sequence. The predicted theoretical spectrum must be sufficiently similar to the observed experimental spectrum in order for the similarity based identification to succeed. However, due to the complex nature of peptide fragmentation is often difficult to make such predictions accurately. The popular Sequest algorithm,⁸ which relies on the cross-correlation between the theoretical spectrum of a database peptide and the experimental mass spectrum, skirts the issues involved in predicting theoretical spectra by relying on a naïve fragmentation model that assigns a fixed equal intensity to the major ions (e.g., *b*- and *y*-ions) and a lower fixed intensity to their neutral losses. In practice, most experimental spectra are not very similar to their Sequest-predicted theoretical spectra, which leads to missed identifications.

It is widely accepted that better understanding of peptide fragmentation is required to produce more accurate and sensitive peptide sequencing algorithms.^{9–13} To this end, several data-mining studies have been performed on large sets of mass spectra with the goal of discovering and quantifying the factors that play a role in peptide fragmentation. Some of the notable influences on fragmentation that were examined include: the influence the amino acids adjacent to the cleaved bond have on its observed intensity;^{10,11,13–15} the effects ceratin amino acids have on the abundance of neutral losses;^{10,12,16} and how the composition of basic amino acids in the peptide influences the fragmentation.^{11,15,17} In some cases this information was incorporated into the scoring models used by sequencing algorithms.^{13,18–25}

There have also been additional efforts to incorporate predictions of theoretical spectra into the peptide identification process. A recent method developed by Zhang^{26,27} for creating realistic theoretical mass spectra relies on a kinetic model that simulates the peptide fragmentation process in order to predict theoretical MS/MS spectra from peptide sequences, which can then be compared with experimental mass spectra for identification purposes. This method was used as the basis for a new scoring function for a database search program that outperformed Sequest and Mascot.²⁸ However, the spectrum prediction process can be difficult

at times, and becomes less accurate and more computationally intensive with large peptides and higher charge states. The ByOnic database search algorithm²³ takes a different approach, using heuristic-based rules for predicting fragment ion ranks, which it incorporated into its scoring function.

Due to the complexity of peak intensity prediction problem we took on a slightly relaxed problem, and concentrated only on predicting a peptide's relative fragment-ion peak ranks. We call this problem the *Peak Rank Prediction Problem* and it is outlined in Figure 1. In this problem, we are given a peptide sequence $P = p_1 \dots p_n$ and a set of fragment ions $\mathcal{F}$, and we need to predict π , which is a permutation of P 's fragment ions. The goal is for the permutation π to resemble, as much as possible, the ordering of peaks, from strongest to weakest, as observed in an experimental mass spectrum of the peptide P . Since the weakest peaks' intensities are usually below the instrument's detection level, not all n peaks for a fragment type are likely to be observed. We are therefore mainly concerned with accurately predicting the order of the strongest peaks.

We note that while peak intensity predictions can possess more discriminatory power than peak rank predictions, it is often beneficial for sequencing algorithms to rely on peak ranks instead of intensities. In particular, using peak ranks makes an sequencing algorithm less vulnerable to the idiosyncrasies often observed in individual MS/MS spectra.²⁹ In an accompanying manuscript,³⁰ we describe how our peak rank prediction models are incorporated into a scoring function for peptide-spectrum matches that greatly improves the performance of both de novo sequencing and database search algorithms.

Peak rank prediction models can also be useful for tasks out side the realm of peptide sequencing. For example, being able to predict which peptide fragments are likely to be the strongest can be invaluable for selecting optimal transitions in multiple reaction monitoring (MRM) experiments in which one is trying to quantify proteins for which there is insufficient previous MS/MS experimental data to guide to peak selection.³¹

Methods

Peptide Fragmentation

Peptide fragmentation is a complex process, often involving several competing chemical pathways, including some that have yet to be discovered or thoroughly understood. Much effort has been devoted towards studying the processes that govern peptide fragmentation.^{32–36} Below we describe some of the prominent mechanisms involved in peptide fragmentation that are described in the literature. Since the data used to train our models was acquired in experiments involving collision induced dissociation (CID), we focus on factors that are known to influence this type of fragmentation.

The most widely accepted model for peptide fragmentation is the *mobile proton model*.^{9,37–42} According to this model, during the ionization stage, peptides are protonated at various sites (mainly terminal amino groups and basic side chain groups), with some protonation sites being more favored than others (e.g., the amino group of arginine). The peptide cleavage is initiated by migration of the sequestered charge (proton) from the initial site of protonation to an amide carbonyl oxygen along the peptide backbone. A nearby carbonyl, *N*-terminal to the protonated carbonyl oxygen, serves as a nucleophile to attack the electropositive carbon of the protonated carbonyl.^{9,43,44} This cleavage usually produces either a prefix (*N*-terminal) *b*-ion or a suffix (*C*-terminal) *y*-ion. This type of fragmentation is considered a “charge-directed” process, since it requires the involvement of a proton near the cleavage site.⁹

The energy required to mobilize a proton from a basic side chain or the terminal amino group depends on the identity of the protonated amino acid, with dissociation energy requirements mimicking the order of amino acid gas phase basicity,³⁶ i.e., the energy required to mobilize a proton that is on arginine > lysine > histidine > non-basic amino acids. In some cases, such as peptides containing arginines, the energy required to mobilize the proton might be too high, making it possible for alternative fragmentation pathways to dominate; pathways that do not involve the ionizing proton (a “charge-remote” fragmentation pathway). These fragmentation events often get initiated by ions containing long side-chains or poly-ring structures.⁴⁵ The dissociation of protonated peptides can be described as a competition between charge-remote and charge-directed fragmentation pathways, in a complicated reaction pattern where fragment ions are formed with substantially different probabilities.³⁶

The activity of charge-direct and charge-remote pathways can be significantly enhanced or reduced depending on the identity of the amino acids adjacent to the peptide's cleavage site. For instance, the fragmentation of the bond *N*-terminal to proline is known to be very active when there is a mobile proton.^{14,46} Often this is the most active site in the peptide, and it is responsible for a very intense peak in the observed spectrum. Similar enhanced fragmentation, though slightly less intense, occurs if we replace the proline with a glycine residue. In contrast, when no mobile protons are available, the charge-remote pathways can dominate and cleavage is enhanced near acidic residues.^{9,47} In addition to fragmentation of the backbone, there are also pathways that lead to the peptides losing neutrally charged molecules such as CO , NH_3 and H_2O ,³² which diversifies the set of observed fragment ions.

MS/MS Datasets

The training data for our experiments consisted of approximately 320000 unique peptide-spectrums pairs collected from various MS/MS experiments involving low-resolution CID ion-trap mass spectrometers. Most of the data we used was generated in the Briggs lab at UCSD (samples of from human HEK293 cell culture⁴⁸ and samples from *Dictyostelium discoideum*⁴⁹), and the Smith lab at PNNL (samples taken from *Shewanella oneidensis* MR-1^{50,51}).

We used the InsPecT database search tool²¹ to perform peptide identifications (release 20070613), using the default search parameters (precursor mass tolerance 2.5 Da, fragment ion tolerance 0.5 Da). All searches were performed using a shuffled decoy database.^{52,53} The InsPecT *F*-score threshold values for accepting identifications were selected to ensure a true positive peptide identification rate of 98% (i.e., only 2% of the peptide hits came from the decoy database).

Since a peptide's charge and precursor mass can greatly influence the nature of its observed spectrum, we partitioned the training data into distinct sets as described in Table 1, and trained separate models for each partition. The number of partitions generated for each charge depended on the number of training peptides available (we divided the doubly-charged peptides into 6 parts, while the singly-charge were only divided into 3 parts). Partitioning the data this way also gave computational benefits, since many of the partitioned models required several days to train and also used up most of the available RAM on the machines.

A peptide's mass and charge are not the only factors that influence its mass spectrum. The peptide's amino acid composition also has a great influence on the fragmentation. In particular, the number and location of the basic amino acids (namely arginine and lysine) play an important role in influencing which peptide fragmentation mechanisms are most active. To reflect the important role amino acid composition plays in peptide fragmentation, we applied an additional partitioning of the training data according to proton mobility. Following the division suggested

by Kapp et al.,¹⁵ we placed peptides in three categories, according to their amino acid composition:

- **Mobile peptides** - the number of basic residues (i.e., combined arginine, lysine, and histidine residues) is less than the peptide's charge.
- **Nonmobile peptides** - number of arginine residues is greater than or equal to the peptide's charge.
- **Partially mobile peptides** - all peptides not classified as mobile or nonmobile.

Ranking Algorithms

To create models for predicting peptide fragment we used a discriminative ranking algorithm called RankBoost due to Freund et al.⁵ The RankBoost algorithm uses a machine learning method called boosting,^{54,55} which produces highly accurate prediction rules by combining many “weak” rules that, each on their own, might be only moderately accurate.

To train models with RankBoost we need to supply the algorithm with several inputs:

- $\mathcal{X}$ - a instance space. In our case, $\mathcal{X}$ is a training set of peptide fragment peaks, in which each instance $x \in \mathcal{X}$ is a pair $x = (P, I)$, where P is a peptide sequence and I is a single fragment ion (e.g., $b_7, y_5, b_5 - H_2O$, etc.)
- Φ - an ordering of the instances in the training set. This information is described using a feedback function $\Phi: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ that contains ordered pairs of instances of the form (x_0, x_1) . In our case, we used a feedback function that assigns a pair of instances (x_0, x_1) a positive value if both x_0 and x_1 represent fragments belonging to the same peptide P and x_1 has a higher intensity than x_0 (the value of $\Phi(x_0, x_1)$ is proportionate to the intensity difference between x_0 and x_1 , as explained below). All other pairs of instances were given a weight of 0.
- F - a set of feature functions. Each feature $f \in F$ can be expressed as a function $f: \mathcal{X} \rightarrow \mathbb{R}$. Our features measure different aspects that can influence the detected intensity of a fragment ion such as the location of the cleavage, identity of adjacent amino acids, and more. We go into further details about the feature functions below.

The goal of the RankBoost learning algorithm is to create a scoring function $H: \mathcal{X} \rightarrow \mathbb{R}$ that weights and combines feature functions from F , in order to induce a ranking that misorders as few pairs as possible in Φ . For the problem at hand, this means that the ranking function H is optimized to give high scores to strong peaks and low scores to weak peaks (when the peaks being compared belong to the same peptide).

RankBoost Models For Predicting Peak Ranks

The structure of the peak rank prediction problem lends itself naturally to a ranking algorithm-based solution, and in this section we demonstrate how this can be done using the RankBoost algorithm. First, we note that the intensities observed in a peptide's experimental spectrum naturally induce a ranking of the fragment ions (this is done simply by ordering the the peptide's fragment ions according to the intensities observed for them in the spectrum). We can also easily extract from a peptide's mass spectrum a set of ordered pairs of peaks which are required for the algorithm's feedback function Φ . However, the key components in our approach, as in any machine learning-based method, are the features we use to describe the data.

Our feature functions derive all their needed information from the peptide's sequence $P = p_1 \dots p_n$. Though the peptide's precursor charge also influences the observed peak pattern, as we noted above, we train separate models based on charge, size, and proton mobility state, so this information does not need to be conveyed to the feature functions. The feature functions

are used to create a feature vector to describe each peptide fragment. For each peptide cleavage position $1 \leq i < n$ and fragment type $t \in \mathcal{F} = \{y, b, \dots\}$, we fill a separate feature vector. Ultimately, these vectors are scored by the model to assign each fragment with a ranking score. The rank scores are then used to induce a permutation of the peptide's fragment ion peaks (e.g., if b_4 's rank score is higher than y_7 's, this is interpreted as the fact that we expect b_4 's intensity to be higher than y_7 's in the experimental spectrum).

There are several different types of sequence-based features that are predictive of a fragment's intensity, which we describe below. The total number of features that can be used to predict a particular fragment's rank is over 200. In each model we only consider the four most abundant fragment types (according to the statistics observed in the training examples for the model's specific charge/size/mobility state). This set of fragments, denoted by F , usually contains the abundant prefix b - and suffix y -ions, and their neutral losses (e.g., $b - H_2O$). Thus, our models can potentially use over 800 features. However, in practice, not all features get included in the models since during the training, many of them do not yield a significant improvement to the ranking performance compared to the features already selected to be in the model (features are only added to a model if they reduce the training ranking error⁵).

Even though a model might contain hundreds of features, the actual feature vectors that are created for peaks are extremely sparse. On average, most of the features are not relevant to the given peak instance (in this case we say the feature abstains). Typically less than 10–20 feature functions do not abstain (for example see Figure 5). Most of the features are only applicable to a specific fragments type, so for example, only the features belonging to y -ions participate in the scoring of a y -peak, the rest of the features abstain. In addition, many of the features, especially the amino acids adjacency features described below, come in sets of 20 (one feature for each standard amino acid), and for every fragment scored, only one of the features in the set is assigned a nonzero value.

The first feature we mention is an indicator function which holds the type of fragment being examined (e.g., $b, y, b - H_2O$, etc.) This feature is used to give a prior score for the different fragment types (usually the b - and y -ions receive a positive score while the neutral losses get zero). This feature assigns a value with all fragment ions. The rest of the features are designed to assign values to specific fragment types (this is denoted by the superscript t in the feature names). Following is a description of the three classes of features we use: peak location features, adjacent amino acid features, and peptide composition features.

Peak location features—The relative mass of a peak has a significant influence on the peak's observed intensity. Peaks located near the center of the peptide generally have higher intensity than the peaks near the N - and C -terminals.^{10,56} This could be because the cleavage of the amide bonds might be more energetically favorable near the center of the peptide, than it is near the terminals. Another possible reason for this could be that peaks near the terminals are often close to the edges of the “visible” range of masses which the instruments are calibrated to detect, and can thus suffer from instrumental bias that records a weaker signal for them. Note that the mass spectrometers that generated our training data were calibrated to detect peaks between a minimal mass that is approximately $\sim 0.25 \cdot$ precursor mass, and a fixed maximal mass of 2000 Da. For example, if the peptide has a precursor mass of 2500 Da, only b and y peaks that have a mass of 625–2000 Da can be detected. Since other instruments can have significantly different visible mass ranges, this would require training specific models for such new instruments even if the same method for peptide fragmentation method is used.

Let Min be the minimal peak mass that can be detected with a peptide with P 's precursor mass; Max be the maximal mass that can be detected (the minimum between P 's precursor mass and

2000); and m be theoretical mass of the peak being ranked. There are 3 features (per fragment ion type $t \in \mathcal{F}$) used to describe the location of a peak:

- $f_{rel}^t(m) = \frac{m - Min}{Max - Min}$ - The relative location of the peak with mass m in the range $[Min, Max]$.
- $f_{min}^t(m) = m - Min$ - The distance from the minimal detected mass.
- $f_{max}^t(m) = Max - m$ - The distance from the maximal detected mass.

Note that for each instance, only one of the features $f_{min}^t(m)$ and $f_{max}^t(m)$ gets assigned a value depending on whether the peak is closer to the N -terminal or the C -terminal. The superscript t in the feature names reflects the fact that there is a separate feature created for each fragment ion type $t \in \mathcal{F}$.

Adjacent amino acid features—The identity of the amino acids adjacent to a cleavage site play an important role in determining the propensity of the peptide to undergo fragmentation at that site.^{10,11,14,15} These influences can be very significant. For instance, the peaks that are products of a cleavage N -terminal to proline are extremely strong^{14,46} (usually they are much stronger than all other peaks in the spectrum).

We derived a simple set of pattern-based features to describe the amino acids adjacent to a cleavage site. We only look at the amino acids that are up to 3 positions away from the cleavage site, since amino acids farther away do not have as strong an influence on the fragmentation. In total, we created 6 sets of 20 features each, which we describe below. Let X denote one of the 20 standard amino acids; $P = p_1 \dots p_n$ be the peptide sequence for which we want to predict peak ranks; i be the index of the cleavage site we are examining (i.e., i is the number of amino acids that are on the N -terminal side of the cleavage position); and $t \in \mathcal{F}$ be the fragment type being scored. The indicator function $I[[p_i = X]]$ returns 1 if the amino acid p_i is X and 0 otherwise. Our 6 sets of features can be expressed using the following indicator functions:

- $f_{Cut-3=X}^t(P, i) = I[[p_{i-3} = X]]$ - The amino acid 3 positions before the cleavage site is X .
- $f_{Cut-2=X}^t(P, i) = I[[p_{i-2} = X]]$ - The amino acid 2 positions before the cleavage site is X .
- $f_{Cut-1=X}^t(P, i) = I[[p_{i-1} = X]]$ - The amino acid directly before the cleavage site is X .
- $f_{Cut+1=X}^t(P, i) = I[[p_{i+1} = X]]$ - The amino acid directly after the cleavage site is X .
- $f_{Cut+2=X}^t(P, i) = I[[p_{i+2} = X]]$ - The amino acid 2 positions after the cleavage site is X .
- $f_{Cut+3=X}^t(P, i) = I[[p_{i+3} = X]]$ - The amino acid 3 positions after the cleavage site is X .

These adjacent amino acid features are sparse. For any given peptide and cleavage index i , at most 6 of the 120 features takes a value 1 (less than 6 features might take a nonzero value if the index i being examined is near the terminals).

Peptide composition features—Besides examining the amino acids adjacent to cleavage site, we can also extract important information from a peptide's general amino acid composition. For instance, if a peptide has basic amino acids near the N -terminal, rather than the C -terminal, the observed b -ion fragments can be stronger than the y -ion fragments, contrary to what is typically observed with tryptic peptides. Furthermore, there are amino acids that are known to be more prone to neutral losses, such as asparagine which often loses NH_3 . A simple way to capture this sequence information is to count the number of occurrences of each amino

acid on both sides of the cleavage site. We created the following two sets of 20 features for each fragment type $t \in \mathcal{F}$:

- $f_{N\#X}^t(P, i) = \sum_{j=1}^{i-3} I[p_j=X]$ - The number of times X appears in P on the N -terminal side of the cleavage.
- $f_{C\#X}^t(P, i) = \sum_{j=i+4}^n I[p_j=X]$ - The number of times X appears in P on the C -terminal side of the cleavage.

Note that we exclude the amino acids that are 1–3 amino acids away from the cleavage site, since these are covered by the adjacent amino acid features.

We also take note of the amino acids on the N - and C -terminals, since these specific positions can have a special influence on the fragmentation outcome. We use two sets of features to express this information:

- $f_{N=X}^t(P) = I[p_1=X]$ - The amino acid on the N -terminal is X .
- $f_{C=X}^t(P) = I[p_n=X]$ - The amino acid on the C -terminal is X .

Creating Training and Validation Sets—In order to train a model with RankBoost, we need to supply the algorithm with samples from the instance space $\mathcal{X}$ and a feedback function $\Phi: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$. For each partition of the training data (see Table 1), we separated the peptides into three sets based on their proton mobility (mobile, partially mobile, and nonmobile), and repeated the following steps:

- **Selection of fragment types to be modeled** - We examine the set of peptides and their matching spectra, and assign the four most abundant fragment ions to the model's fragment set $\mathcal{F}$ (usually $b, y \in \mathcal{F}$). The model being created only predicts ranks for these fragment types.
- **Addition of instances to $\mathcal{X}$** - For each peptide $P = p_1 \dots p_n$ in the training set, we compute the masses of the $4 \cdot (n - 1)$ non-trivial peaks corresponding ion types $\mathcal{F}$, and ignore the ones that fall outside of the range detected by the mass spectrometer. We record the intensities detected in the experimental mass spectrum for each of these $k \leq 4 \cdot (n - 1)$ fragments (assigning an intensity of zero if a fragments has no detected peak in the spectrum). For each peak i , $1 \leq i \leq k$, we create a feature vector x_i^P using the ranking features described above, and add x_i^P to the instance space $\mathcal{X}$.
- **Addition pairs to feedback function Φ** - Let $x_1^P, \dots, x_k^P$ be the k instances created for P 's fragment peaks, and let I_i be the intensity observed in the mass spectrum for fragment peak i ($1 \leq i \leq k$). Since the feedback function Φ is symmetric,⁵ it only requires pairs (x_i^P, x_j^P) in which $I_j > I_i$, so we exclude all pairs where peak $I_j \leq I_i$. Let $I_{\max} = \max\{I_1, \dots, I_k\}$. Each remaining pair (x_i^P, x_j^P) is assigned a value

$$\Phi(x_i^P, x_j^P) = \frac{I_j - I_i}{I_{\max}} \quad (1)$$

When we examine repeated MS/MS experiments involving the same peptide, we often observe an element of randomness in the peak intensities. Pairs of low-intensity peaks often have different ranks in different experimental spectra. Therefore, with low intensity peaks, the order

observed is often determined not so much due to mechanisms that govern peptide fragmentation, but more due to the randomness of the peak intensity measurements. Applying the weighting scheme in Eq. 1 reduces the influence of these somewhat randomly ordered instance pairs on the models being trained.

During the training we assess the model's performance, by measuring its rank loss, which equals the weighted proportion of pairs of instances in Φ whose order is predicted incorrectly by the model. Measuring performance on the same set of data used to train the model is not recommended since this usually leads to model overfitting. To avoid this pitfall, we set aside a portion of the data to serve as a *validation set* which is used to test the performance of the model (and determine what configuration of the parameters is optimal in general). We typically set aside 1/4 of the training data for validation purposes.

Results

Training the RankBoost Models

A typical training set we used to train the models consisted of ~ 10000 peptides, which created ~ 400000 instances (peaks) in $\mathcal{X}$, and $\sim 4 \cdot 10^6$ pairs of the form $(x_0, x_1) \in \mathcal{X} \times \mathcal{X}$ in Φ . Our algorithm is capable of performing approximately 2 rounds of boosting per second on this data using a single CPU. The training typically required between 100000 to 400000 rounds to complete. We now examine in detail the training of a specific model for predicting fragments in doubly-charged "mobile" peptides of mass 1100–1300 Da, in order to observe the dynamics of the convergence of the RankBoost model.

Figure 2 depicts the statistics obtained from the training of the ranking model. The greatest reduction in error rates is achieved in the initial boosting rounds. The first feature to be selected in the models is f_{rel}^y (Relative peak location of y). Using this feature reduced the initial error from about 50% to 33%. By round 10 there are only 4 features in the model, and they alone reduce the ranking error to 26.6%. After 100 rounds, the model includes 17 features, the training error was reduced to 15.1%. From this stage on, adding more features and performing additional rounds of boosting yields a small but steady improvement in the model's performance. After 100000 rounds, using 423 features, the training error was 7.97%. At the stage of the model's convergence at round 200000, the validation error is minimal (7.89%), and the model includes 512 features. Continuing with additional rounds does not improve the validation error, but does start to widen the gap between the validation and training error (to a maximum of 0.08%). This gap, though small, is evidence that the model is starting to experience overfitting (it is being optimized specifically for the training data).

Figure 3 illustrates the dynamics of tuning the scores assigned to a ranking feature. The figure shows a graphical representation of the cumulative contribution of the feature f_{rel}^y (relative peak location of y), after various numbers of update rounds (in every boosting round only the most beneficial feature has its weights changed). The feature $f_{rel}^y(P, i)$, can have a value between 0, designating that the peak is located at the minimal detectable mass, and 1, for a peak at the maximal detectable mass. After the first round the rule that was most beneficial was to assign a score +0.5 if the y peak falls between 5%-80% of the mass range. Subsequent update rounds refine the function further, decreasing the score near the extremities and increasing it near the center. This feature's behavior corresponds to a known phenomenon observed with MS/MS spectra of peptides, in which the peaks near the center are, on average, much stronger than the peaks detected in the low and high mass ranges.^{10,56} These figures also illustrate the typical behavior of the learning algorithm, which makes large changes in the initial rounds, but only smaller refining changes thereafter.

In Figure 4 we compare the feature f_{rel}^y with peak location features for the other fragment types used in the model: b , $b - NH_3$ and $b - H_2O$. All four features show a similar trend, the scores near the terminals are much lower than the scores near the center of the detected mass range. However, the magnitude of the b and y feature scores is much greater than the magnitude observed for $b - NH_3$ and $b - H_2O$. This reflects the fact that b - and y -ions are typically much stronger (and are ranked higher) than ions with neutral losses.

Table 2 lists the most positive and most negative adjacent amino acid features in the model for the y , b , $b - NH_3$, and $b - H_2O$ ions, trained on doubly-charged mobile peptides of mass 1100–1300 Da. In the interest of clearer presentation the names of the features were shortened, omitting $f_{(,...)}^t(P, i) = 1$ (since all features in the same column belong to the same fragment type). Note that with the features listed in the table, as with all our indicator function-based features, we assume that if the indicator returns a value 0, the feature's score is also 0. In addition, since with low-energy CID fragmentation it is impossible to distinguish between the isobaric amino acids leucine and isoleucine, these two amino acids are represented by a single feature (I/L) in all the tables below.

The features listed in Table 2 capture much of the previously observed phenomena regarding the influence of amino acids on peptide fragmentation.^{10,11,15} With mobile peptides, the most active fragmentation pathways tend to be charge-directed, and most prominent among these is the cleavage N -terminal to proline. This is reflected in the tables with the highest scoring feature for all fragments being “Cut +1 = P”. The influence of proline goes beyond the position directly N -terminal to the cleavage site, since “Cut +2 = P” and “Cut +3 = P” also have high scores. When the proline is on the other side of the cleavage the resulting peaks tend to be extremely weak, and often they do not get detected. Consequently, the features “Cut –1 = P” tend to have negative scores. Glycine (G) is known to have a similar effect to proline (though often weaker).

The tables also show other known effects, such as the positive influence of the aliphatic amino acids (e.g., A,V,I,L) when they appear N -terminal to the cleavage. There are also known negative effects such as when the amino acids serine (S) or threonine (T) appear N -terminal to the cleavage, or when histidine (H) is C -terminal to the cleavage.^{15,57} The table also shows amino acid influences that are particular to specific fragment ions. For instance serine and threonine that are known to promote water loss,^{16,32} have positive scores in $b - H_2O$, but not with the other fragment types. Similarly, glutamine (Q) and asparagine (N) are known to increase loss of ammonia which explains their features' positive scores with $b - NH_3$. Another interesting phenomenon is the negative influence that the acidic amino acids (D,E) have on the intensity of neutral losses, especially $b - NH_3$, which seems to be more prominent than the effect they have on the b - or y -ions. The tables also show the negative effect of having arginine on the C -terminal side of the cleavage (“Cut +1 = R”) with the b , $b - NH_3$ and $b - H_2O$ fragments. This can be explained by the fact with doubly-charged mobile peptides, if there is an arginine on the C -terminal side, that means that there is no additional one on the N -terminal side. This sequesters most of the charge on the suffix fragments, which makes the prefix b fragments much weaker. In addition, the fact that this feature is most negative at the +1 position means that the arginine itself also interferes with the fragmentation (this fact was also previously observed^{15,57}).

Table 3 compares the amino acid adjacency features for b -ions across different mobility states (mobile, partially mobile, and nonmobile). The main difference between these models is the role that the charge-remote pathways play (in particular the fragmentation that is initiated by having aspartic acid (D) on the N -terminal side of the cleavage site). With the mobile peptides, the most positive feature is “Cut +1 = P”, while “Cut –1 = D” is not even among the top ten. This starts to change with the partially mobile peptides, where both features have similar strong scores. However, with the nonmobile peptides, the charge-remote pathways become more

dominant, which is manifested in the table by a much higher score for “Cut $-1 = D$ ”. Another difference, between mobility states is the positive effect of having histidine *N*-terminal the cleavage site (e.g., “Cut $-1 = H$ ”). Note that these feature do not appear in the mobile model because doubly-charged mobile peptides with a tryptic end cannot contain additional basic amino acids like histidine (see definitions of proton mobility states). When examining the most negative features, we see that the lists for the different mobility states are quite similar.

Table 4 examines the most positive and most negative peptide composition features in the model for doubly-charged mobile tryptic peptides with mass 1100–1300 Da. The strongest positive features involve the counts of basic amino acids on both sides of the cleavage site. For the *y*-ion, the highest scores are given to the presence of basic amino acids on the *C*-terminal side of the cleavage, while for the *b*-ion the presence of basic amino acids on the *N*-terminal side is rewarded. Interestingly, having histidine on the *C*-terminal is also highly rewarded with the *y*-ion (the score for having at least one histidine is 0.78). This might be because, in general, having histidine in the vicinity of a cleavage cite increases the fragmentation.⁵⁷ The presence of aliphatic amino acids (such as leucine and alanine) is also rewarded with high scores. For $b - NH_3$ we note that as expected, having asparagine (N) on the *N*-terminal side of the cleavage is highly rewarded (asparagine has a tendency to lose NH_3). In addition, the positive score for having multiple phenylalanines (F) can be explained by the known fragmentation mechanisms that involve loss of NH_3 from aromatic amino acids.^{58,59} With $b - H_2O$ we see positive effect of having histidine (H) in the peptide. We also note the positive scores associated with having threonine (T) and serine (S) on the *N*-terminal side of the cleavage (the presence of these amino acids is known to increase water loss^{16,32}). As far a the negative features are concerned, it is worthy to note that all feature types are penalized for having proline on the *N*-terminal side of the cleavage. This reflects the far reaching effects of proline (as seen in Table 3).

Table 5 lists the most positive and most negative *N*- and *C*-terminal amino acid features. The strongest features for the *y*-ion add premiums when the *C*-terminal is basic (“*C*-term aa is R” and “*C*-term aa is K”). There is also a high premium when the *N*-terminal is proline. This is most likely given to counter the negative scores assigned for the amino acid adjacency features like “Cut $-1 = P$ ”, since when a proline is near the terminals its positive and negative effects on fragmentation are diminished. Interestingly, the highest positive terminal feature scores belong to “*N*-term aa is Q” and “*N*-term aa is W” for $b - NH_3$ and “*N*-term is E” for $b - H_2O$. These feature scores are a manifestation of the glutamine/glutamic acid effect,⁶⁰ in which extensive loss of NH_3 occurs when the *N*-terminal amino acid is glutamine, and a loss of H_2O when the amino acid is glutamic acid. Our results also indicate that there is substantial loss of NH_3 when tryptophan is on the *N*-terminal (though involving a different fragmentation mechanism⁶¹). The most outstanding negative scoring features are “*N*-term is Q” and “*N*-term is E” with the *b*- and *y*-ions. This is also due to the glutamic acid/glutamine effects. When the *N*-terminal amino acid is either glutamine or glutamic acid, the scores of *b*- and *y*-ions get reduced in order to increase the ranks of the $b - NH_3$ and $b - H_2O$, respectively.

Rank Prediction Results

We trained a total of 39 models for peak rank predictions (3 mobility states $\times$ 13 partitions described in Table 1). The experiments below test the accuracy of these models for the peak rank prediction task.

First we examine how peak ranks get predicted in practice. Figure 5 gives a detailed example of the calculation of the peak rank scores of the top three peaks in the doubly-charged peptide GEEVTPISAIR. Part *a* displays the experimental mass spectrum, in which the top three peaks are y_6 (intensity 89.9), y_7 (intensity 67.5), and b_5 (intensity 45.9). Part *b* of the figure holds a table that shows how the peak rank scores are calculated (only features that have a nonzero score are listed). The calculated rank scores induce the same rank ordering that is observed in

the experimental spectrum (y_6 has a score of 4.67, y_5 has a score of 4.22, and b_5 has a score of 3.22). These scores were computed using the model for mobile doubly-charged peptides of mass 1100–1300 (this model's most prominent features are listed in Table 2 through Table 5). In this model, both the y - and b -ion fragment types receive a score of +0.43 from the “Fragment Indicator” feature ($b - NH_3$ and $b - H_2O$, which are the other two fragments in the model, receive a score of 0). The table shows that the peak location features give a large score premium to all three peaks since they are b - and y -ions situated near the center of the peptides. Generally, the adjacent amino acid features have a lesser weight than the position features, with exception of the features “Cut $-1 = P$ ” which gives a high score premium of +1.52 for y_6 and +1.37 to b_5 (note that y_6 and b_5 get different scores since they are different fragment ions, and are assigned different features: $f_{Cut-1=P}^y$ and $f_{Cut-1=P}^b$ respectively). The peptide composition features generally contribute low scores, with exception of the “# C-side R=1” which gives +0.36 to the y -ions. The major factor that contributes to raising the y -ions scores compared to b_5 's is the feature “C-term aa is R” which gives +0.5 to the y -ions and -0.5 to the b -ion. This feature can be explained by the fact that with a mobile peptide, if the C-terminal is R, this means that most of the charge is sequestered at the C-terminal which leads to strong y -ions and weaker b -ions.

In order to examine how well the peak rank predictions fit experimental spectra, we selected a test case of peptides of length 10 and ran benchmark experiments with them. Figure 6 shows histograms of the ranks predicted for the peaks observed in the experimental spectra with ranks 1,3,5, and 10. The figures show that the rank predictions for the stronger peaks are much more accurate than the rank predictions for weaker peaks. For example, in over 60% of the cases the peak predicted to be strongest was in fact the strongest peak in the spectrum. However, when we examine the tenth strongest peak in the spectrum, we see that only in 9% of the cases, the predicted rank is correct. In addition, with the stronger peaks (ranks 1 and 3), the predicted ranks tend to be more concentrated around the observed rank, while with the weaker peaks (ranks 5 and 10), the predictions are more spread out.

We ran additional benchmark experiments to test the peak rank prediction models on peptides of various lengths and charges. Table 6 holds results of these experiments. In this analysis we examined only the top 7 peaks, since these are the ones that are most likely to be observed in experimental spectra (in addition, the predictions with lower ranked peaks are less accurate, and thus less helpful when matching to experimental spectra.) To evaluate how well our predictions match the observed data, we examine two parameters. First, we computed Spearman's rank correlation coefficient (ρ)⁶² between a peptide's predicted peak ranks and the ones observed in the experimental spectra:

$$\rho = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

where d_i is the difference between i (the predicted rank) and the actual rank observed for the i^{th} fragment in the experimental spectrum, and n is the number of peaks we examined ($n = 7$). ρ can take values between +1 and -1 , where $\rho = +1$ indicates perfect rank correlation, $\rho = -1$ indicates perfect negative rank correlation, and ρ with values near 0 indicates that there is very weak, or non existent, correlation between the sets of variables examined. We also examined how well we predicted each of the ranks 1,...,7. Since it is often difficult to make exact rank predictions, we examined the proportion of cases in which our predictions were close to the observed, with the predicted rank p being at most 3 positions away from the observed rank r (i.e., $r - 3 \leq p \leq r + 3$).

There are a few general trends we can observe in the table. First, as evident by the positive values of ρ , with all charge states and peptide lengths examined, our predictions have strong positive correlations with the fragment ranks observed in the experimental spectra. These correlations are strongest with the shorter peptides with charges 1 and 2, where our rank predictions are most accurate. This poorer results with triply-charged peptides are due to the more complex fragmentation processes involved with multiply-charged peptides which also often results in poorer fragmentation patterns that are difficult to predict. Examining the rest of the table's columns reveals that the predictions for stronger peaks are more accurate than weaker peaks (this is evident in all cases where the accuracy deteriorates as we move from rank 1 to rank 7). In addition, the peak rank predictions with shorter peptides are generally more accurate than the ones done with longer peptides. However, the decline in prediction accuracy is not that severe. For instance, there is only a 5.6% reduction in the accuracy of the prediction of the strongest peak when go from doubly-charged peptides of length 10 (92.9% correct) to peptides of length 20 (87.3% correct); even though the number of peak ranks that need to be predicted is doubled.

Discussion

In this paper we explored how ranking algorithms can be used to predict the ranks of fragment ion peaks in a peptide's experimental mass spectrum based solely on the peptide's sequence (the peak rank prediction problem). This is not a trivial task. Peptide fragmentation is a complex process that involves many competing chemical pathways. It is quite difficult to create generative probabilistic models for mass spectra that consider this wide range of chemical processes. Yet generative models are not required to solve the peak rank prediction problem. The structure of this problem lends itself nicely to a ranking-based solution since the required output is an ordering of instances (we want to predict a permutation for fragment ions, from strongest to weakest). We based our solution to this problem on ranking-based discriminative models; created using a pool of simple sequence-based features. These features get combined by the RankBoost algorithm into strong discriminative models, using a training procedure that is optimized to minimize the number of pairs of peaks on which the model makes ordering mistakes. What makes this approach feasible are the large amounts of MS/MS training data that have become available (we used approximately 320000 unique peptide-spectrum pairs), which enabled us to create detailed models with minimal overfitting.

As opposed to many “black box” machine learning algorithms, like neural networks, the RankBoost algorithm is relatively transparent, and allows us to observe the dynamics involved in creating the models, and later on, the scoring of new instances. We are able to examine the contribution of single features, and in many cases understand the logic that guided the scoring of different instances. The high-scoring features in our models reflect known chemical pathways in peptide fragmentation. Since the CID fragmentation method is well studied, we did not find evidence of novel fragmentation pathways that were not previously described in the literature. However, it is likely that given training data that was generated using a fragmentation method that has not been studied as extensively as CID, our models could supply some interesting insights into the dynamics of the peptide fragmentation.

Our experimental results show that our peak rank predictions are fairly accurate, and can therefore be helpful in detecting correct peptide-spectrum matches. In an accompanying manuscript,³⁰ we incorporate our peak rank prediction models into a scoring function for peptide-spectrum matches that greatly improves the performance of both de novo and database search algorithms. The models we described can also be useful for multiple reaction monitoring (MRM) experiments,³¹ where they can assist in the selection of transitions in cases where there is no experimental MS/MS data to identify a peptide's strongest fragments.

Our models' predictions for peak ranks in triply-charged peptides were slightly inferior to the predictions for singly and doubly-charged peptides. This is due to the fact that the dynamics of the fragmentation pathways in triply-charged peptides are more difficult to predict (these peptides are typically longer and contain more basic amino acids). This underscores the need for more complex features that can improve the modeling of such cases (e.g., features that simultaneously describe the location of several amino acids). Another possible way to increase the models' accuracy is to incorporate into RankBoost a more powerful and expressive learning algorithm (instead of the regular boosting). Alternating decision trees⁶³ might be good choice since they can easily account for relationships and dependencies between simple features and might provide a more accurate model of the complex interactions between fragmentation pathways.

Accurate peak models become essential when trying to identify large peptides with high charge states. CID can reach charges +4 and +5, though spectra with these charges are not identified often. ETD⁶⁴ spectra routinely possess even higher charges. Top-down mass spectra,⁶⁵ record the fragmentation of whole proteins and can possess precursor charges that reach dozens and even hundreds of protons. When such large peptides (or proteins) are fragmented, they often undergo several fragmentation events, giving rise to many internal fragments. Often, the internal fragments end up accounting for most of the spectrum's intensity. Since there is a quadratic number of possible internal ions, it becomes especially important for scoring functions to be able to predict which are the few internal fragments that are likely to be observed.

Acknowledgement

I am grateful to Pavel Pevzner for many inspiring discussions. This research was supported by the "National Center for Research Resources" of the NIH via grant P-41-RR24851. I would also like to acknowledge the UCSD FWGrid Project for the availability of their computational infrastructure. The UCSD FWGrid Project is funded in part by NSF Research Infrastructure Grant number NSF EIA-0303622.

References

1. Washburn M, Wolters D, Yates J III. Large Scale analysis of the yeast proteome via multidimensional protein identification technology. *Nature Biotech* 2001;19:242–247.
2. Aebersold R, Mann M. Mass spectrometry-based proteomics. *Nature* 2003;422:198–207. [PubMed: 12634793]
3. Venter J, et al. Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 2004;304:66–74. [PubMed: 15001713]
4. Freund, Y.; Schapire, R. A decision-theoretic generalization of on-line learning and an application to boosting.. *European Conference on Computational Learning Theory*; 1995. p. 23-37.
5. Freund Y, Iyer R, Schapire R, Singer Y. An Efficient Boosting Algorithm for Combining Preferences. *Journal of Machine Learning Research* 2003;4:933–969.
6. Stein S, Scott D. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J. Am. Soc. Mass. Spectrom* 1994;5:859–866.
7. Yates J III, Morgan S, Gatlin P, and Griffin CL, Eng J. Method To Compare Collision-Induced Dissociation Spectra of Peptides: Potential for Library Searching and Subtractive Analysis. *Anal. Chem* 1998;70:3557–3565. [PubMed: 9737207]
8. Eng J, McCormack A, Yates J III. An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino Acid Sequences in a Protein Database. *J. Am. Soc. Mass Spectrom* 1994;5:976–989.
9. Wysocki V, Tsaprailis G, Smith L, Brei L. Mobile and Localized Protons: A Framework for Understanding Peptide Dissociation. *J. Mass Spectrom* 2000;35:1399–1406. [PubMed: 11180630]
10. Tabb D, Smith L, Brei L, Wysocki V, Lin D, Yates J III. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem* 2003;75:1155–1163. [PubMed: 12641236]

11. Huang Y, Triscari J, Tseng G, Pasa-Tolic L, Lipton M, Smith R, Wysocki V. Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns. *Anal. Chem* 2005;77:5800–5813. [PubMed: 16159109]
12. Martin D, Eng J, Nesvizhskii A, Gemmill A, Aebersold R. Investigation of Neutral Loss during Collision-Induced Dissociation of Peptide Ions. *Anal. Chem* 2005;77:4870–4882. [PubMed: 16053300]
13. Barton S, Richardson S, Perkins D, Bellahn I, Bryant T, Whittaker J. Using Statistical Models To Identify Factors That Have a Role in Defining the Abundance of Ions Produced by Tandem MS. *Anal. Chem* 2007;79:5601–5607. [PubMed: 17579495]
14. Brei L, Tabb D, Yates J III, Wysocki V. Cleavage N-terminal to Proline: Analysis of a Database of Peptide Tandem Mass Spectra. *Anal. Chem* 2003;75:1963–1971. [PubMed: 12720328]
15. Kapp E, Schutz F, Reid G, Edes J, Moritz R, O'Hair R, Speed T, Simpson R. Mining a Tandem Mass Spectrometry Database To Determine the Trends and Global Factors Influencing Peptide Fragmentation. *Anal. Chem* 2003;75:6251–6264. [PubMed: 14616009]
16. Savitski M, Kjeldsen F, Nielsen M, Zubarev R. Relative Specificities of Water and Ammonia Losses from Backbone Fragments in Collision-Activated Dissociation. *J. Proteome Res* 2007;6:2669–2673. [PubMed: 17503798]
17. Tabb D, Huang Y, Wysocki V, Yates J III. Influence of Basic Residue Content on Fragment Ion Peak Intensities in Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem* 2004;76:1243–1248. [PubMed: 14987077]
18. Schutz F, Kapp E, Simpson R, Speed T. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochem. Soc. Trans* 2003;31:1479–1483. [PubMed: 14641094]
19. Elias J, Gibbons F, King O, Roth F, Gygi S. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. biotech* 2004;22:214–219.
20. Frank A, Pevzner P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem* 2005;77:964–973. [PubMed: 15858974]
21. Tanner S, Shu H, Frank A, Mumby M, Pevzner P, Bafna V. InsPecT: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra. *Anal. Chem* 2005;77:4626–4639. [PubMed: 16013882]
22. Wan Y, Chen T. PepHMM: A hidden Markov model based scoring function for tandem mass spectrometry. *Anal. Chem* 2006;78:432–7. [PubMed: 16408924]
23. Bern M, Cai Y, Goldberg D. Lookup Peaks: A Hybrid of de Novo Sequencing and Database Search for Protein Identification by Tandem Mass Spectrometry. *Anal. Chem* 2007;79:1393–1400. [PubMed: 17243770]
24. Colinge J. Peptide Fragment Intensity Statistical Modeling. *Anal. Chem* 2007;79:7286–7290. [PubMed: 17713966]
25. Klammer A, Reynolds S, Bilmes J, MacCoss M, Noble W. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. *Bioinformatics* 2008;24:i348–356. [PubMed: 18586734]
26. Zhang Z. Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Anal. Chem* 2004;76:3908–3922. [PubMed: 15253624]
27. Zhang Z. Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides with Three or More Charges. *Anal. Chem* 2005;77:6364–6373. [PubMed: 16194101]
28. Sun S, Meyer-Arendt K, Eichelberger B, Brown R, Yen C-Y, Old W, Pierce K, Cios K, Ahn N, Resing K. Improved Validation of Peptide MS/MS Assignments Using Spectral Intensity Prediction. *Mol. Cell. Proteomics* 2007;6:1–17. [PubMed: 17018520]
29. Frank A, Savitski M, Nielsen M, Zubarev R, Pevzner P. De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry. *J. Proteome Res* 2007;6:114–123. [PubMed: 17203955]
30. Frank A. A Ranking-Based Scoring Function For Peptide-Spectrum Matches. To appear in *J. Proteome Research*. 2009
31. Anderson L, Hunter C. Quantitative Mass Spectrometric Multiple Reaction Monitoring Assays for Major Plasma Proteins. *Mol. Cell. Proteomics* 2006;5:573–588. [PubMed: 16332733]

32. Papayannopoulos I. The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrometry Reviews* 1995;14:49–73.
33. O'Hair R. The role of nucleophile-electrophile interactions in the unimolecular and bimolecular gas-phase ion chemistry of peptides and related systems. *Journal of Mass Spectrometry* 2000;35:1377–1381. [PubMed: 11180627]
34. Cole R. Some tenets pertaining to electrospray ionization mass spectrometry. *Journal of Mass Spectrometry* 2000;35:763–772. [PubMed: 10934430]
35. Gabelica V, De Pauw E. Internal energy and fragmentation of ions produced in electrospray sources. *Mass Spectrometry Reviews* 2005;24:566–587. [PubMed: 15317019]
36. Paizs B, Suhai S. Fragmentation pathways of protonated peptides. *Mass Spectrometry Reviews* 2005;24:508–548. [PubMed: 15389847]
37. Tang X, Thibault P, Boyd R. Fragmentation reactions of multiply-protonated peptides and implications for sequencing by tandem mass spectrometry with low-energy collision-induced dissociation. *Anal. Chem* 1993;65:2824–2834. [PubMed: 7504416]
38. Jones J, Dongre A, Somogyi A, Wysocki V. Sequence Dependence of Peptide Fragmentation Efficiency Curves Determined by Electrospray Ionization/Surface-Induced Dissociation Mass Spectrometry. *J. Am. Chem. Soc* 1994;116:8368–8369.
39. Dongré A, Jones J, Somogyi A, Wysocki V. Influence of Peptide Composition, Gas-Phase Basicity and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model. *J. Am. Chem. Soc* 1996;118:8365–8374.
40. Harrison A, Yalcin T. Proton mobility in protonated amino acids and peptides. *International Journal of Mass Spectrometry and Ion Processes* 1997;165:339–347.
41. Summerfield S, Gaskell S. Fragmentation efficiencies of peptide ions following low energy collisional activation. *International Journal of Mass Spectrometry and Ion Processes* 1997;165:509–521.
42. Csonka I, Paizs B, Lendvay G, Suhai S. Proton mobility in protonated peptides: a joint molecular orbital and RRKM study. *Rapid Communications in Mass Spectrometry* 2000;14:417–431. [PubMed: 10717650]
43. Polce M, Ren D, Wesdemiotis C. Dissociation of the peptide bond in protonated peptides. *Journal of Mass Spectrometry* 2000;35:1391–1398. [PubMed: 11180629]
44. Schlosser A, Lehmann W. Five-membered ring formation in unimolecular reactions of peptides: a key structural element controlling low-energy collision-induced dissociation of peptides. *Journal of Mass Spectrometry* 2000;35:1382–1390. [PubMed: 11180628]
45. Cheng C, Gross M. Applications and mechanisms of charge-remote fragmentation. *Mass Spectrometry Reviews* 2000;19:398–420. [PubMed: 11199379]
46. Vaisar T, Urban J. Probing the Proline Effects in CID of Protonated Peptides. *J. Mass Spectrom* 1996;31:1185–1187. [PubMed: 8916427]
47. Huang Y, Wysocki V, Tabb D, Yates J III. The Influence of Histidine on C-leavage C-terminal to Acidic Residues in Doubly Protonated Tryptic Peptides. *Int. J. Mass Spectrom* 2002;219:233–244.
48. Tanner S, Shen Z, Ng J, Florea L, Guigó R, Briggs S, Bafna V. Improving gene annotation using peptide mass spectrometry. *Genome Res* 2007;17:231–239. [PubMed: 17189379]
49. Tanner S, Payne SH, Dasari S, Shen Z, Wilmarth PA, David LL, Loomis WF, Briggs SP, Bafna V. Accurate Annotation of Peptide Modifications through Unrestrictive Database Search. *J. Proteome Res* 2008;7:170–181. [PubMed: 18034453]
50. Masselon C, Pasa-Tolic L, Tolic N, Anderson G, Bogdanov B, Vilkov A, Shen Y, Zhao R, Qian W, Lipton M, Camp D, Smith R. Targeted Comparative Proteomics by Liquid Chromatography-Tandem Fourier Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem* 2005;77:400–406. [PubMed: 15649034]
51. Gupta N, Tanner S, Jaitly N, Adkins J, Lipton M, Edwards R, Romine M, Osterman A, Bafna V, Smith R, Pevzner P. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res* 2007;17:1362–1377. [PubMed: 17690205]
52. Higdon R, Hogan J, Belle GV, Kolker E. Randomized sequence databases for tandem mass spectrometry peptide and protein identification. *OMICS* 2005;9:364–379. [PubMed: 16402894]

53. Elias J, Gygi S. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007;4:207–214. [PubMed: 17327847]
54. Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 1997;55:119–139.
55. Schapire R, Singer Y. Improved Boosting Using Confidence-rated Predictions. *Machine Learning* 1999;37:297–336.
56. Havilio M, Haddad Y, Smilansky Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem* 2003;75:435–444. [PubMed: 12585468]
57. Huang Y, Triscari J, Wysocki V, Pasa-Tolic L, Anderson G, Lipton M, Smith R. Dissociation Behaviors of Doubly-Charged Tryptic Peptides: Correlation of Gas-Phase Cleavage Abundance with Ramachandran Plots. *J. Am. Chem. Soc* 2004;126:3034–3035. [PubMed: 15012117]
58. ElAribi H, Orlova G, Hopkinson A, Siu K. Gas-Phase Fragmentation Reactions of Protonated Aromatic Amino Acids: Concomitant and Consecutive Neutral Eliminations and Radical Cation Formations. *Journal of Physical Chemistry A* 2004;108:3844–3853.
59. Lioe H, O'Hair R. Neighbouring group processes in the deamination of protonated phenylalanine derivatives. *Org. Biomol. Chem* 2005;3:3618–3628. [PubMed: 16211098]
60. Harrison A. Fragmentation reactions of protonated peptides containing glutamine or glutamic acid. *J Mass Spectrom* 2003;38:174–87. [PubMed: 12577284]
61. Lioe H, O'Hair R, Reid G. Gas-phase reactions of protonated tryptophan. *J. Am. Soc. Mass Spectrom* 2004;15:65–76. [PubMed: 14698557]
62. Lehmann, E.; D'Abrera, H. *Nonparametrics: Statistical Methods Based on Ranks*. 1975. Holden-Day
63. Freund, Y.; Mason, L. The Alternating Decision Tree Algorithm.. *Proceedings of the 16th International Conference on Machine Learning*; 1999. p. 124-133.
64. Syka J, Coon J, Schroeder M, Shabanowitz J, Hunt D. Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *PNAS* 2004;101:9528–9533. [PubMed: 15210983]
65. Kelleher NL. Top down proteomics. *Anal. Chem* 2004;76:197A–203A. [PubMed: 14697051]

Peak Rank Prediction Problem

Input:

- Peptide sequence $P = p_1 p_2 \dots p_n$, where p_i , $1 \leq i \leq n$, are amino acids.
- Set of fragment ion types $\mathcal{F}$, e.g., $\mathcal{F} = \{b, y, a, y - H_2O, b^{+2}, \dots\}$.

Output:

- A permutation π of the set of all possible fragment peaks ($\mathcal{F} \times \{1, \dots, n\}$), where π is ordered according to decreasing intensity (e.g., $\pi = y_8, y_6, b_3, b_4, b_3 - H_2O, \dots$).

Figure 1.
The peak rank prediction problem.

Figure 2.

Statistics of RankBoost training for 300000 rounds. the left-side graph displays the training and validation error rates after running the RankBoost algorithm for a given number of rounds, and the right-side graph displays the number of features included in the model (i.e., the features that have a nonzero weight). The x-axis displays the number of boosting rounds using a logarithmic scale. The figures were generated using a training set of doubly-charged “mobile” peptides with precursor masses 1100–1300.

Figure 3.

Updating Relative y-ion peak location feature (f_{rel}^y). The figure shows the cumulative score plots for the y-ion peak's relative location after k rounds in which the feature was selected to be modified by the learning algorithm ($k = 1, 2, 4, 8, 16, 32, 128, 512$). The x-axis holds the values of $f_{rel}^y(P, i)$, which are in the range 0–1 and are binned into 20 bins of equal length. The y-axis holds the cumulative score assigned by the ranking model to different values of $f_{rel}^y(P, i)$.

Figure 4.

Comparison of peak location features for different ion types. The figure shows the score plots for the same feature (relative peak location) with different fragment ion types: y , b , $b - NH_3$, and $b - H_2O$. The x -axis represents the peak's relative location (values in the range 0–1 that are binned into 20 bins of equal size). The y -axis holds the score assigned by the ranking model to different values of the relative peak locations. The different magnitudes of the scores reflect the fact that on average, b and y -ions are ranked much higher than $b - NH_3$ and $b - H_2O$.

Figure 5.

Example of computation of peak rank scores for the three strongest peaks in the spectrum of the doubly-charged peptide GEEVTPLSALR. Part *a* displays the experimental spectrum of the peptide GEEVTPLSALR. Part *b* lists the feature vectors used to compute the rank scores for the three strongest peaks in the spectrum: y_6 , y_7 , and b_5 . The table only lists features that were assigned nonzero values. (*) Different adjacent amino acid feature, peptide composition features and terminal amino acid features are created for each fragment type (b , y). This explains why the scores for the same type of feature (e.g., Cut -1 is T) are different between y_6 and b_5 .

Figure 6.

Peak Rank Prediction Histograms. The figure shows histograms of the predicted peak ranks for peaks observed in the experimental spectra with ranks 1,3,5 and 10. The results were collected from 5000 doubly-charged peptides of length 10. Each observed peak was assigned a predicted rank between 1 and 36 (9 internal cleavage sites $\times$ 4 fragment types).

Table 1

MS/MS training dataset. The set of 319578 pairs of unique peptides and MS/MS spectra was partitioned according to charge and precursor mass. For each partition we list its precursor mass range, the number of peptides that fell in that range, and the typical length of those peptides (the lengths listed cover at least 95% of the peptides in each partition).

Charge 1			Charge 2			Charge 3		
Precursor mass (Da)	#Unique peptides	Typical lengths	Precursor mass (Da)	#Unique peptides	Typical lengths	Precursor mass (Da)	#Unique peptides	Typical lengths
0–1150	20971	7–12	0–1100	25709	7–12	0–1950	13198	10–19
1150–1400	18984	9–15	1100–1300	33167	9–14	1950–2450	13131	16–24
1400+	16231	11–20	1300–1600	45595	10–16	2450–3000	12684	20–29
			1600–1900	43054	13–20	3000+	13824	25–48
			1900–2400	43225	15–25			
			2400+	19805	20–32			
	56186			210555			52837	

Most positive and negative adjacent amino acid features. The features are taken from the model for the y , b , b - NH_3 , and b - H_2O ions, trained on doubly-charged mobile peptides of mass 1100–1300 Da.

Most positive features					
y		b		b-NH ₃	
				b-H ₂ O	
Feature	Score	Feature	Score	Feature	Score
Cut+1 is P	1.52	Cut+1 is P	1.37	Cut+1 is P	0.95
Cut-1 is V	0.79	Cut+2 is P	0.78	Cut+2 is P	0.56
Cut-1 is L	0.68	Cut-1 is V	0.69	Cut-3 is N	0.53
Cut+2 is P	0.57	Cut-1 is I/L	0.57	Cut-1 is V	0.50
Cut-2 is I/L	0.50	Cut-2 is I/L	0.43	Cut-2 is N	0.39
Cut+3 is H	0.45	Cut-2 is W	0.33	Cut-1 is I/L	0.35
Cut+1 is G	0.44	Cut-2 is V	0.26	Cut+3 is P	0.17
Cut-2 is W	0.37	Cut-1 is W	0.25	Cut-2 is W	0.13
Cut-2 is V	0.36	Cut-2 is A	0.24	Cut+3 is H	0.13
Cut-2 is A	0.32	Cut+3 is P	0.24	Cut-1 is Q	0.12
				Cut+3 is P	0.20

Most negative features					
y		b		b-NH ₃	
				b-H ₂ O	
Feature	Score	Feature	Score	Feature	Score
Cut-1 is G	-1.24	Cut-1 is P	-1.07	Cut+1 is R	-0.45
Cut-1 is P	-1.08	Cut+1 is R	-0.98	Cut-1 is G	-0.38
Cut-1 is S	-0.67	Cut-1 is G	-0.82	Cut-1 is S	-0.28
Cut-1 is N	-0.64	Cut+1 is H	-0.72	Cut-2 is E	-0.27
Cut-3 is P	-0.48	Cut-3 is G	-0.67	Cut-1 is P	-0.23
Cut-3 is Y	-0.46	Cut+2 is H	-0.64	Cut-2 is E	-0.21
Cut-3 is G	-0.46	Cut-3 is S	-0.63	Cut+1 is Q	-0.21
Cut-3 is T	-0.44	Cut-3 is T	-0.59	Cut-2 is D	-0.19
Cut-3 is S	-0.41	Cut-1 is S	-0.58	Cut+1 is E	-0.17
Cut-1 is D	-0.40	Cut+2 is K	-0.58	Cut+1 is D	-0.15

Table 3

Adjacent amino acid features for *b*-ion in peptides of different mobility states. The features are taken from the models trained on doubly-charged mobile peptides of mass 1100–1300 Da.

Most positive features					
Mobile			Nonmobile		
Feature	Score	Feature	Score	Feature	Score
Cut +1 is P	1.37	Cut –1 is D	1.01	Cut –1 is D	1.30
Cut +2 is P	0.78	Cut +1 is P	1.00	Cut +1 is P	0.77
Cut –1 is V	0.69	Cut –1 is H	0.75	Cut –2 is H	0.63
Cut –1 is I/L	0.57	Cut –2 is H	0.70	Cut –3 is R	0.63
Cut –2 is I/L	0.43	Cut –3 is R	0.62	Cut –1 is E	0.62
Cut –2 is W	0.33	Cut –1 is V	0.42	Cut –1 is H	0.62
Cut –2 is V	0.26	Cut +2 is P	0.37	Cut –1 is R	0.40
Cut –1 is W	0.25	Cut –1 is E	0.37	Cut –1 is V	0.28
Cut –2 is A	0.24	Cut –1 is I/L	0.31	Cut +2 is P	0.23
Cut +3 is P	0.24	Cut +1 is I/L	0.19	Cut –1 is I/L	0.23

Most negative features					
Mobile			Nonmobile		
Feature	Score	Feature	Score	Feature	Score
Cut –1 is P	–1.07	Cut –1 is P	–0.91	Cut –1 is P	–0.73
Cut +1 is R	–0.98	Cut –1 is G	–0.72	Cut –1 is G	–0.71
Cut –1 is G	–0.82	Cut –1 is S	–0.62	Cut –1 is S	–0.49
Cut +1 is H	–0.72	Cut –1 is T	–0.47	Cut –1 is T	–0.46
Cut –3 is G	–0.67	Cut +1 is R	–0.35	Cut +1 is R	–0.38
Cut +2 is H	–0.64	Cut –1 is N	–0.32	Cut +2 is R	–0.27
Cut –3 is S	–0.63	Cut –3 is G	–0.30	Cut +1 is D	–0.25
Cut –3 is T	–0.59	Cut +1 is D	–0.29	Cut –3 is P	–0.24
Cut –1 is S	–0.58	Cut –3 is P	–0.29	Cut –1 is N	–0.23
Cut +2 is K	–0.58	Cut –2 is P	–0.26	Cut –3 is G	–0.17

Table 4 Most positive and negative peptide composition features. The features are taken from the model for the $y, b, b-NH_3$, and $b-H_2O$ ions, trained on doubly-charged mobile peptides of mass 1100–1300 Da (only features for the b, y and $b-NH_3$ ions are shown). In all the features we assume that amino acid counts of 0 receive score 0.

Most positive features					
y		b		$b-NH_3$	
Feature	Score	Feature	Score	Feature	Score
#C-side H > 0	0.81	#N-side H > 0	1.23	#N-side N = $\begin{cases} 1 \\ 2 + \end{cases}$	0.46 0.74
#N-side H > 0	0.78	#N-side I / L = $\begin{cases} 1 \\ 2 \\ 3 + \end{cases}$	0.30 0.48 0.73	#C-side P = $\begin{cases} 1 \\ 2 + \end{cases}$	0.10 0.45
#C-side K > 0	0.67	#N-side K > 0	0.61	#N-side Q = $\begin{cases} 1 \\ 2 + \end{cases}$	0.09 0.26
#N-side K > 0	0.39	#N-side A = $\begin{cases} 1 \\ 2 \\ 3 \\ 4 + \end{cases}$	0.19 0.34 0.50 0.60	#C-side R > 0	0.25
#C-side R > 0	0.36	#N-side R > 0	0.55	#N-side F = $\begin{cases} 1 \\ 2 + \end{cases}$	0.02 0.21
				#N-side H > 0	0.19
				#N-side G = $\begin{cases} 1 \\ 2 \\ 3 + \end{cases}$	0.04 0.15 0.22

Most negative features					
y		b		$b-NH_3$	
Feature	Score	Feature	Score	Feature	Score
#N-side P = $\begin{cases} 1 \\ 2 \\ 3 + \end{cases}$	-0.33 -0.59 -0.54	#C-side Q = $\begin{cases} 1 \\ 2 + \end{cases}$	-0.11 -0.28	#N-side P = $\begin{cases} 1, 2 \\ 3 + \end{cases}$	-0.16 0.01
#C-side Q = $\begin{cases} 1 \\ 2 \\ 3 + \end{cases}$	-0.20 -0.48	#N-side P = $\begin{cases} 1 \\ 2 \\ 3 + \end{cases}$	-0.14 -0.10 0.01	#N-side S = $\begin{cases} 1 \\ 2 + \end{cases}$	-0.04 -0.05
				#C-side V = $\begin{cases} 1 \\ 2 + \end{cases}$	-0.08 0.00

Most negative features					
y		b		$b-NH_3$	
				$b-H_2O$	
Feature	Score	Feature	Score	Feature	Score
#C-side A = $\begin{cases} 1 \\ 2 \\ 3 \end{cases} +$	-0.06	#N-side T = $\begin{cases} 1 \\ 2 \\ 3 \end{cases} +$	-0.10	#C-side I / L = $\begin{cases} 1, 2 \\ 3 \end{cases} +$	-0.06
	-0.20		-0.09		-0.02
	-0.35		-0.03		-0.02
#C-side I / L = $\begin{cases} 1 \\ 2 \\ 3 \end{cases} +$	-0.10	#C-side I / L = $\begin{cases} 1 \\ 2 \\ 3 \end{cases} +$	-0.04	#N-side W = $\begin{cases} 1 \\ 2 \end{cases} +$	-0.04
	-0.17		-0.07		0.00
	-0.24		-0.10		
#C-side N = $\begin{cases} 1 \\ 2 \end{cases} +$	-0.12	#N-side G = $\begin{cases} 1 \\ 2 \\ 3 \end{cases} +$	-0.03	#C-side T = $\begin{cases} 1 \\ 2 \end{cases} +$	-0.02
	-0.22		-0.06		0.00
			-0.09		

Most positive and negative terminal amino acid features. The features are taken from the model for the y , b , b - NH_3 , and b - H_2O ions, trained on doubly-charged mobile peptides of mass 1100–1300 Da.

Table 5

Most positive features					
y		b		b-NH3	
Feature	Score	Feature	Score	Feature	Score
N-term is P	0.69	N-term is P	0.19	N-term is Q	1.01
C-term is R	0.50	N-term is D	0.13	N-term is W	0.70
C-term is K	0.12	C-term is F	0.12	N-term is Y	0.26
N-term is M	0.12	N-term is G	0.11	N-term is E	0.21
N-term is T	0.10	N-term is T	0.07	C-term is K	0.01
				N-term is V	0.01

Most negative features					
y		b		b-NH3	
Feature	Score	Feature	Score	Feature	Score
N-term is E	-0.20	N-term is E	-1.06	N-term is S	-0.09
N-term is Q	-0.10	N-term is Q	-0.81	N-term is T	-0.06
N-term is D	-0.08	C-term is R	-0.50	N-term is P	-0.04
N-term is G	-0.02	C-term is K	-0.49	N-term is N	-0.03
N-term is N	-0.02	C-term is H	-0.30	N-term is D	-0.01
				N-term is I/L	-0.02

Table 6
Peak rank prediction accuracy. The table holds statistics of the peak rank prediction accuracy for peptides of various lengths and charges. For each set of peptides, we computed the rank correlation and observed how often the peak predicted to have a rank p ($p = 1 \dots 7$) was observed in the spectrum with a rank r such that $r-3 \leq p \leq r+3$.

Charge	Peptide length	Avg. Rank Correlation (ρ)	Rank of predicted peak						
			1	2	3	4	5	6	7
1	7	0.736	0.918	0.845	0.774	0.745	0.688	0.652	0.609
	10	0.697	0.894	0.822	0.738	0.695	0.643	0.605	0.573
	15	0.671	0.883	0.774	0.696	0.607	0.547	0.503	0.466
	20	0.642	0.779	0.740	0.679	0.584	0.426	0.429	0.429
2	7	0.756	0.957	0.916	0.905	0.870	0.820	0.777	0.704
	10	0.714	0.929	0.896	0.848	0.818	0.755	0.698	0.639
	15	0.687	0.904	0.837	0.765	0.732	0.643	0.602	0.538
	20	0.663	0.873	0.801	0.739	0.675	0.636	0.583	0.497
	25	0.650	0.857	0.774	0.735	0.668	0.605	0.553	0.497
	30	0.671	0.884	0.848	0.749	0.692	0.624	0.567	0.510
3	15	0.563	0.783	0.699	0.649	0.641	0.568	0.490	0.445
	20	0.513	0.692	0.610	0.598	0.586	0.517	0.462	0.417
	25	0.441	0.598	0.571	0.546	0.522	0.450	0.392	0.354
	30	0.454	0.601	0.498	0.500	0.478	0.431	0.368	0.328
	35	0.435	0.591	0.533	0.490	0.452	0.416	0.380	0.317
	40	0.427	0.571	0.548	0.515	0.505	0.399	0.389	0.336