research articles Journal of proteome research

Comparison of Probability and Likelihood Models for Peptide Identification from Tandem Mass Spectrometry Data

William R. Cannon,*,[†] Kristin H. Jarman,*,[‡] Bobbie-Jo M. Webb-Robertson,[†] Douglas J. Baxter,[§]
Christopher S. Oehmen,[†] Kenneth D. Jarman,[||] Alejandro Heredia-Langner,[#]
Kenneth J. Auberry,[⊥] and Gordon A. Anderson[⊥]

Pacific Northwest National Laboratory, Richland, Washington 99352

Received May 19, 2005

We evaluate statistical models used in two-hypothesis tests for identifying peptides from tandem mass spectrometry data. The null hypothesis H_0, that a peptide matches a spectrum by chance, requires information on the probability of by-chance matches between peptide fragments and peaks in the spectrum. Likewise, the alternate hypothesis H_A, that the spectrum is due to a particular peptide, requires probabilities that the peptide fragments would indeed be observed if it was the causative agent. We compare models for these probabilities by determining the identification rates produced by the models using an independent data set. The initial models use different probabilities depending on fragment ion type, but uniform probabilities for each ion type across all of the labile bonds along the backbone. More sophisticated models for probabilities under both H_A and H_0 are introduced that do not assume uniform probabilities for each ion type. In addition, the performance of these models using a standard likelihood model is compared to an information theory approach derived from the likelihood model. Also, a simple but effective model for incorporating peak intensities is described. Finally, a support-vector machine is used to discriminate between correct and incorrect identifications based on multiple characteristics of the scoring functions. The results are shown to reduce the misidentification rate significantly when compared to a benchmark cross-correlation based approach.

Keywords: tandem mass spectrometry • peptide identification • fragmentation model • likelihood • hypothesis test • support vector machine

Introduction

High-throughput proteomic technologies seek to characterize the state of the proteome in a cell population in much the same manner that DNA microarrays seek to characterize the state of gene expression in a cell population. Characterization of the proteins can be done using several different methods. A typical procedure may involve extracting cellular proteins followed by tryptic digestion and then separating the peptides with liquid chromatography.[1−3] The separated peptides are then identified by MS/MS. Ideally, peptides could then be quantitated, post-translational modifications determined and the information assembled into a picture of the proteomic state of a cell population.

Just as with DNA microarrays, quality assurance of the high-throughput process is of paramount importance in order for proteomics to be of value to biologists. If peptides are initially identified improperly, then this information and the information on post-translational state and quantitation of protein expression are not of much value. There is much work needed to be done in this field, as evidenced by identification rates in which typically only 7−25% of all MS/MS spectra are annotated with a peptide when using any one tool. Although a large fraction of the spectra are not of high enough quality to analyze successfully and multiple tools can be used to improve the identification rates,[4] it is also clear that the identification rate could be increased with more sophisticated analyses.

Early probability-based models essentially employed a single hypothesis test to determine the significance of a by-chance match between a candidate peptide and an experimental spectrum. In particular, Mann and Wilm[5] present a scheme that divides an experimental spectrum into three regions containing added masses and a partial sequence. The resulting *sequence tag* is then input into a protein database and scores for database hits are constructed from the estimated probability

* To whom correspondence should be addressed. (W.R.C.) P.O. Box 999/MS K5-12, Richland, Washington, 99352, Tel: (509) 375-6732, Fax: (509) 375-6631, E-mail: william.cannon@pnl.gov. (K.H.J.) P.O. Box 999/MS K1−83, Richland, Washington, 99352, Tel: (509) 375−4539, Fax: (509) 375−2604, E-mail: kristin.jarman@pnl.gov.

[†] Computational Biology and Bioinformatics Group, Computational and Information Sciences Directorate.

[‡] Decision and Sensor Analytics Group, Computational and Information Sciences Directorate.

[§] Molecular Sciences Computing Facility, Environmental Molecular Sciences Laboratory.

[||] Computational Mathematics Group, Computational and Information Sciences Directorate.

[#] Statistical Sciences Group, Computational and Information Sciences Directorate.

[⊥] Instrument Development Lab, Environmental Molecular Sciences Laboratory.

10.1021/pr050147v CCC: $30.25

Published on Web 09/10/2005

that the hits are the result of chance. Mascot[6] estimates the probability that the match between a peptide and the spectrum would occur by chance and reports a level of significance of a match, however, as far as we are aware of, no details of the statistical model and scoring scheme are published.

In more recent work, the single hypothesis test models have been refined to include more realistic probabilities for by-chance matches between peptides and spectra. These new tools use more refined statistical methods in order to build a foundation for increasing the identification rate of peptides from MS/MS spectra. *SCOPE*[7] scores a given peptide through a detailed probabilistic model of peptide fragmentation based on the single hypothesis that the peptide was causally related to the spectrum, and then uses a test of significance to see if the score is high enough to be significant. Sadygov and Yates[8] present a scoring method based on a single hypothesis test where the null (random match) hypothesis probabilities are taken as the hypergeometric probability distribution. Likewise, Fridman, et al.[9] use a a significance test based on the hypergeometric distribution and use a goodness of fit measure to score each peptide. *ProbID*[10] take a decision analytic approach to peptide identification, where Bayes' Theorem is used to estimate the posterior probability that a given peptide is the correct one. In this case, the peptide with the highest posterior probability is assumed to be a correct hit.

Recent analyses based on hypothesis comparisons are especially promising,[11−14,36] as these analyses quantify the differences that would be expected if the peptide under consideration did in fact result in the spectrum, as compared to a by-chance match between the peptide and the spectrum. Underlying both the null hypothesis that the peptide matched the spectrum by chance and the alternate hypothesis that the peptide is causally related to the spectrum are probability models for matching each peptide fragment to a peak in the spectrum. To date, the probability models under the causal match hypothesis have considered different probabilities for different ion types such as *b* ions, *y* ions and neutral loss ions, but there has been no attempt to assign specific fragmentation probabilities to each labile bond. Likewise, most probability models for observing a by-chance match between a fragment and a peak in the spectrum assume that this probability of a match is uniform throughout the range of the spectrum.

To compare the two hypotheses and decide whether a match between a peptide and a spectrum is real or not, likelihood ratios have generally been used.[11−14,36] The likelihood ratio is a convenient measure not only because it directly compares the fragmentation matches under each hypothesis, but also because it allows for peptides of different lengths and charges to be directly comparable. Otherwise, each score characterizing the match between a peptide and a spectrum would have to be adjusted for length and charge state. One potential shortcoming of all likelihood models used to date, however, is that they give equal weight to each factor in the likelihood ratio regardless of the ion fragment under consideration. Some ion fragments, such as *b* and *y* fragments, are more useful in identifying peptides than others, such as fragments resulting from a neutral loss. Accordingly, weighting the factors in the likelihood ratio due to these more informative fragments relative to factors due to less informative fragments may result in improved discrimination between correct identifications and misidentifications.

In this paper, we provide the results on a study of these issues. For matches between spectral peaks and the expected ion fragments from a candidate peptide, we introduce a probability model that depends not only on each ion type, but also on each labile bond along the peptide backbone. For by-chance matches between spectral peaks and expected ion fragments under the null hypothesis, we introduce a truer accounting of the frequency by which peaks in the observed spectrum will randomly match a candidate peptide's fragment. With regard to the likelihood models used to quantify the differences between a by-chance match to a peptide and a causal match, we present an information theory measure that is related to the log-likelihood model which weights each term in the log-likelihood, with *y* and *b* ions having greater weights than less informative fragments due to neutral loss. We also provide a simple but effective measure for taking peak intensities into account also based on information theory. We illustrate these methods on a dataset of 18 999 spectra for peptides of varying length, charge and composition.[15] Finally, we use several measures of the match between spectrum and peptide in a support vector machine learning method that results in a significant reduction in the misidentification rate when benchmarked against a cross-correlation based approach.

Methods

Description of Spectra. For the training set, peptides were derived from *Deinococcus radiodurans* by tryptic digestion and mass analyzed in the laboratory of Richard Smith at the Pacific Northwest National Laboratory. Details can be found in refs 2 and 16. Briefly, the 16 134 CID spectra discussed herein were obtained using electrospray ionization sources feeding Finnigan LCQ Classic ion traps. The spectra were all output in centroid mode. Initial independent identifications were done with *SEQUEST*[17] using an organism-specific sequence database and using a multirun MS/MS strategy. Each peptide was independently identified with the LCQ multiple times on multiple days, and at least one spectrum for each peptide resulted in *SEQUEST*[17] scores exceeding a *DelCN* score of 0.1 and *Xcorr* scores of 2.0 for charge 1 peptides, 2.5 for charge 2 peptides, and 3.5 for charge 3 peptides. Next, the mass of each peptide parent ion was examined as to whether it confirmed to within one part-per-million of the theoretical mass for that peptide by the use of an 11.5 T ion-cyclotron resonance mass spectrometer and a 15% elution time tolerance.[18] If so, the initial *SEQUEST* identification was kept, otherwise it was rejected. The error rate for this process is unknown, but expected to be small.

Details of methods used for preparing the set of peptides used in the evaluation of the computational methods can be obtained from Keller, et al.[15] Briefly, a standardized mixture of peptides was developed and analyzed repeatedly. *SEQUEST* was used to identify peptides from the standard mixture and a decoy database of 88 000 proteins derived from the human genome. Correct identifications were determined when peptides from the control mixture were the top scoring peptide and passed minimum cutoff values for *SEQUEST* scores.

Statistical Model

The peptide identification procedure relies on comparing two hypotheses:

H_A: The spectrum is due to fragmentation of the candidate peptide.

H_0: The match between the candidate peptide and the spectrum is due to chance.

We assume a peptide fragments according to some statistical distribution where different fragment ions have different probabilities of appearing in an experimental MS/MS spectrum. The probability of a peak being generated at a specific location is much higher if the candidate peptide is present (the alternative hypothesis H_A) than purely by chance (the null hypothesis H_0). To compare the two hypotheses, we estimate the probability of appearance for each fragment ion under H_0 and H_A. The likelihood of the collection of observed spectral peaks matching the expected ion fragments for a given peptide is then computed under H_0 and H_A and compared using a likelihood ratio. This likelihood ratio serves as the basis for our scoring algorithm.

For a given sequence, the scoring procedure identifies peaks at locations corresponding to the specific ion fragments using a prediction interval based on the tolerance parameter s for each peak. Under the alternate hypothesis, H_A, the probability of observing a peak at location l_i is estimated by the value p_i computed a priori from training sets. Under the null hypothesis, H_0, the probability of appearance of a peak at location l_i is given by q_i, the estimated probability of a peak appearing at that location purely by chance.

H_A: Development of a Fragmentation Model and Probabilities from a Training Dataset. The probabilities of fragmentation occurring at a specific location along the peptide backbone are computed using a modification to the method for "learning ion types" presented in Dančík et al.[11] The approach presented by these authors allows for automated determination of the frequencies of occurrence of ion types as a function of their difference in mass from the sum of the masses of amino acid residues in a given peptide fragment. Using terminology of previous authors, including Fernandez-de-Cossio et al.[19] and Dančík et al.,[11] we refer to these mass differences as *mass offsets*. For example, y- and b-ions have mass offsets of +19 and +1 from the sum of neutral residue masses, respectively. For a length M ion fragment i, a spectral peak is produced within tolerance s of location l_i with some probability of appearance p_i, where p_i is estimated to be the fraction of spectra in a training dataset in which an M length ion fragment peak is observed. That is, we approximate the probability of appearance of an ion fragment by the frequency with which that fragment appears in the training dataset. While the concept of assigning different probabilities or weights to different ion types is commonly applied in both database and de novo methods, we have extended this approach to computing a different p_i for each ion type at *each position* along the peptide backbone to obtain a more realistic approximation to probabilities of appearance across the spectrum. The result of applying this fragmentation model to a candidate peptide is a model spectrum that is used in scoring. The statistical method and data structures used in the code allow for further extension to incorporate sequence dependent effects such as those described in the recent literature.[20−22] The difficulty in doing so lies not in deploying the probabilities, but rather in determining them from training data.

The fragmentation frequencies at each position along the peptide backbone were computed for charge +1 peptides using approximately 300 spectra for each set of peptides from length 5 to 20-mers, for a total of 4905 spectra. Likewise, for charge +2 peptides, approximately 300 peptides for each length from 6 to 30-mers were used in determining fragmentation frequencies for a total of 7098 spectra. For charge +3 peptides, the lengths trained on were from 6-mers to 55-mers with an average of 84 spectra for each length, with the total number of peptides being 4131. Multiply charged fragments are considered if the probability of observing that fragment is significant.

The probabilities of appearance for each ion fragment as a function of fragment length and type are estimated from peaks appearing by chance in addition to peaks associated with a given ion. Therefore, these probabilities tend to be overly optimistic. If the occurrence of peaks in a particular offset bin purely by chance is low, this false increase of frequencies will not be a serious problem. In the training, we examined the effects of peaks falling in offset bins by chance by filtering small, insignificant peaks from the spectra prior to computing frequencies of appearance. We found that for determining probabilities and scoring peptides, there was no statistical difference between using filters and not using them.

H_0: The Development of Peak Probabilities from Matches to Random Peptides. Here, we present two models for the generation of by-chance match probabilities between a spectral peak and an ion fragment. In the first model, these probabilities are generated by comparing a specific set of random peptides with the spectrum of interest. In the second model of by-chance matches, the probabilities are generated by considering common phenomena that occur in model 1 and observed probabilities for correct matches, as determined under H_A.

Model 1. Under the null hypothesis that the match between a spectrum and a peptide are by chance, we determine the fragment probabilities q_i as follows. We estimate the by-chance probability of a match as the frequency by which random peptides produce fragments at the same position as fragment i in our candidate peptide. More specifically, we generate a set of random peptides and we scan each model spectrum of the random peptides to determine whether a peak is predicted to occur at each discrete location of the actual spectrum. The tally across all random peptides results in an empirical distribution of the by-chance match frequency for each fragment, q_i.

Furthermore, the random peptides are taken as those obtained from the organism's protein sequence database in which each randomly selected peptide has a mass-to-charge consistent with the mass-to-charge value observed for the precursor ion. An example distribution for tryptic peptides having a precursor mass-to-charge ratio of 773.8 is shown in Figure 1. We take this distribution to be the distribution of probabilities that a peak would occur at each location by chance.

Model 2. We assume that a peak may occur at location l_i for two reasons: (1) the fragment ion i observed at l_i is a y (b, a, b-H_2O, y-H_2O) fragment composed of the same amino acids as the candidate peptide but is a juxtaposition of the amino acids in the sequence; (2) The peak at l_i is due to some other fragmentation product that does not simply consist of a juxtaposition of amino acids of the candidate peptide. The latter situation includes both fragmentation ions that are used in the fragmentation model under the alternate hypothesis (y, b, a, b-H_2O, y-H_2O, etc.), as well as other fragmentation ions such as loss of side chains and internal fragmentation ions.

The probability of observing a peak under assumption 1 is determined as follows. For a sequence of M distinct amino acids there are $M!$ ways of ordering the sequence. If each amino acid j occurs M_j times, the multinomial coefficient determines the relative number of ways of ordering the sequence. The probability of observing each of these sequences is the prob-

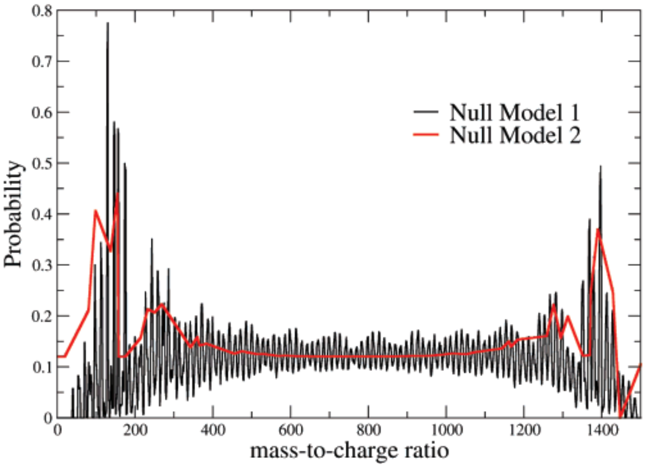

Figure 1. Plot of probability distributions for null hypothesis models 1 and 2. Model 1 is the empirical distribution of the frequency that a fragment occurs at the specified mass-to-charge ratio. This distribution is derived from a set of tryptic peptides that have a mass-to-charge ratio consistent with the observed mass-to-charge ratio of the precursor ion. Large peaks just below 1400 *m*/*z* are due to *b* ion fragments that have lost the C-terminal Arg or Lys, while large peaks below 200 are in part due to *y* ion fragments containing only Arg or Lys residues. The Model 2 distribution is shown for the peptide PGIDFTNDPLLQGR and is described by eqs 1−2 in which $q_0 = 0.12$.

ability of observing any of the sequences multiplied by the multinomial coefficient

$$\pi_i = \frac{M!}{M_1!M_2! \cdots M_r!}p_i \qquad (1a)$$

The probability of observing any of these sequence fragments, $p_{\rm i}$, is taken from the training data, as described above under the alternate hypothesis. Thus, π_i gives a probability of observing a peak due to assumption 1.

This probability of a random match between a peak in the spectrum and a peptide clearly declines as a fragment increases in length. However, the probability of such a match again increases as the fragment approaches its full length. This occurs because we are starting with a set of candidate peptides that have very similar masses. As a result, when these similar-mass parents lose similar 1−3mer fragments the probability of a chance match due to the juxtaposition of amino acids of *the shorter of the two fragments* results in an increased probability of a match of the longer of the two fragments at the high end of the spectrum. We model this as follows

$$\rho_i = p_{\rm parent}({\rm mass}) \cdot \frac{M!}{M_1!M_2! \cdots M_r!}p_i \qquad (1b)$$

Here M is the number of amino acids in the lost (shorter) fragment, M_j is the number of each unique amino acids j in the lost fragment, and p_i is the probability of observing fragment i derived from the training data, and $p_{\rm parent}$(mass) is the fraction of candidate peptides having a parent mass similar to the parent mass of the candidate peptide in question. We can estimate $p_{\rm parent}$(mass) as in the following example. If we use a mass window of 3 daltons to pull candidate peptides from the protein database and we define "similar" to mean those parent peptides that have a mass within ± 0.5 daltons of the candidate peptide in question, then $p_{\rm parent}$(mass) will be approximately 1/3 if we assume that parent masses that are uniformly distributed throughout the 3 dalton range.

Under assumption 2 of this model, the peak at l_i may also be a by-chance match to a fragment ion of the candidate peptide for some unknown reason. The probability of observing a peak at l_i in this manner is q_0, where q_0 can be determined empirically from the training data set or taken as

$$q_0 = N_{\rm pks}\frac{\rm tol}{\max\,(mz) - \min\,(mz)} \qquad (1c)$$

for a test spectrum containing $N_{\rm pks}$ peaks, with m/z tolerance tol and an instrumental mass range max (mz) − min (mz). We note that q_0 approximates the probability of a random peak appearing at any specific location assuming peaks are uniformly distributed about the mass range of interest. The overall probability q_i of a peak appearing at that location purely by chance when some random peptide is present is then given by

$$\begin{aligned} q_i &= 1 - (1 - \pi_i)(1 - \rho_i)(1 - q_0) \\ &= \pi_i + \rho_i + q_0 - \pi_i\rho_i - \pi_i q_0 - \rho_i q_0 + \pi_i\rho_i q_0 \qquad (2) \end{aligned}$$

The last equation can be interpreted in terms of a Venn diagram consisting of three partially overlapping regions, π_i, ρ_i, and q_0. The pair cross-terms correct for over-counting the regions of overlap while the tertiary cross-term corrects for the region in which all three probability regions overlap. In the lower mass-to-charge range, peaks due to π_i will dominate. These peaks are primarily small fragments consisting of 1−3 amino acids in length and can be either from the N-terminus or the C-terminus. Short fragments derived from the C-terminus of random tryptic peptides are especially likely to overlap, since most candidate peptides will have a terminal lysine or arginine, and the chance of a random match to these peaks is relatively high. In the high mass range, peaks due to ρ_i will dominate for similar reasons.

Null Models 1 and 2 are compared in Figure 1. The model 2 probabilities were generated using the charge +2 tryptic peptide PGIDFTNDPLLQGR. Here q_0 = 0.12 for comparison purposes, but different values simply shift the curve by a constant amount for most of its range. Model 2 is a modification of the more commonly used constant-value probability of appearance that takes into account important features that occur in the low and high mass-to-charge range.

Scoring Method. Let the vector x represent appearance of ion fragment peaks in the test spectrum where $x_i = 0$ if peak i is not observed in the test spectrum, and $x_i = 1$ if peak i is observed in the test spectrum. The likelihood ratio for H_0 versus $H_{\rm A}$ is given by the probability of observing x under $H_{\rm A}$ divided by the probability of observing x under H_0. Assuming that the appearance of peaks at different locations is independent, then the likelihood ratio score for a given candidate is L, where

$$\begin{aligned} L &= \frac{P\{\text{observing } x \text{ under } H_{\rm A}\}}{P\{\text{observing } x \text{ under } H_0\}} \\ &= \frac{\prod_i p_i^{x_i} \prod_i (1 - p_i)^{1 - x_{\rm i}}}{\prod_i q_i^{x_i} \prod_i (1 - q_i)^{1 - x_{\rm i}}} \qquad (3) \end{aligned}$$

In practice, we use only fragment peaks whose probability of appearance exceeds q_i (the probability of observing a peak at random) when forming the likelihood ratio. This ensures that the scoring procedure is using peaks that have a different probability of appearance under H_0 and H_A so that the occurrence of each peak for a given candidate is expected to be more frequent than by chance alone.

Next, we take the log-likelihood ratio

$$\Lambda = \sum_i \log\left(\frac{1-p_i}{1-q_i}\right) + \sum_i x_i \log\left[\frac{p_i(1-q_i)}{q_i(1-p_i)}\right] \qquad (4)$$

and apply the following decision rule:

if $\Lambda \leq K_c$, then accept H_0,
if $\Lambda > K_c$, then reject H_0.

Here K_c is the critical decision threshold. Traditional hypothesis testing then dictates that if H_0 is rejected, then H_A is accepted and the candidate sequence is determined to be present in the unknown sample. The cutoff criterion K_c can be determined empirically to be the value that minimizes the combined false and missed positive rates for a test dataset. Or, if the cost of misclassifying a match between a spectrum and an incorrect peptide as correct is judged to be higher than the cost of misclassifying a true match as an incorrect match, K_c can be adjusted accordingly.

We can directly compare two peptides *i* and *j* by taking the ratio of the likelihood ratio from eq 3, or similarly the difference of the log-likelihoods from eq 4. This leads us the log-likelihood, Λ_{ij}, that a peptide *i* is a better match to the spectrum than peptide *j*, relative to by-chance matches. We will later show that this is a very useful criterion.

The log-likelihood ratio in eq 4 gives equal weight to each peak that is expected from the model spectrum when comparing the alternate hypothesis to the null hypothesis. However, a more useful approach may be to weight each term in the log-ratio, giving peaks that are more useful in identifying peptides, such as *b* and *y* ions, more weight. If we take these weights to be the probability with which each peak is expected to be observed, then we get the information theory scoring function

$$\Omega_p = \sum_i x_i\, p_i \log\left(\frac{p_i}{q_i}\right) + \sum_i (1-x_i) p_i \log\left[\frac{(1-p_i)}{(1-q_i)}\right] \qquad (5)$$

The first term is simply the relative entropy between the alternate-hypothesis probability space and the null-hypothesis probability space for fragments that are observed, as averaged over the alternate-hypothesis probability space. Likewise, the second term is the relative entropy between the alternate-hypothesis probability space and the null-hypothesis probability space for fragments that are *not* observed; however, now the average is over the alternate-hypothesis probability space when fragments are expected to be observed. This latter term tells how informative the missing information is, given that we that we expect it with a probability of p_i. In analogy to the likelihood ratio criterion discussed above, a decision criterion for the information theory scoring function can be likewise established. However, now the acceptance of the alternate hypothesis is not based on likelihood, but rather on whether the alternate hypothesis provides more explanatory information than the null-hypothesis.

Just as we can form a difference using eq 4 to compare which of two peptides *i* or *j* is a better match to a spectrum, relative to chance, we can likewise form an analogous difference of the log of each ratio, from eq 5. We will refer to the difference of the logs of these ratios as Ω_{Pij}.

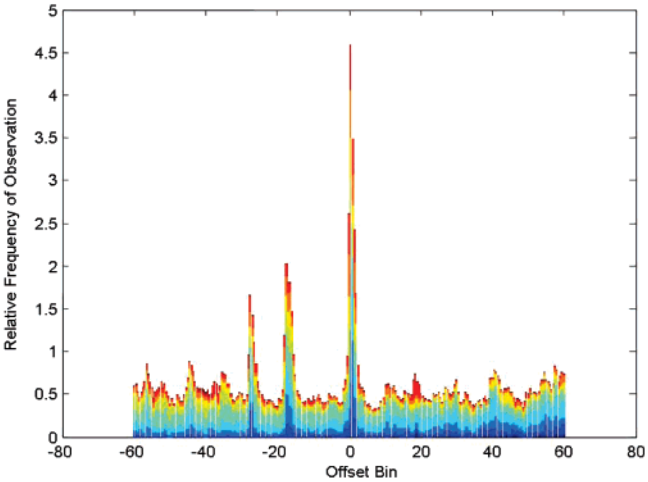

Figure 2. Histogram of ion frequencies versus offset bin for N-terminus partial peptide sequences generated from 10mers. Individual histograms for ion offsets for each partial peptide from length 1 to 9 are colored and stacked to present a summary view of the ion offset patterns that are found.

Intensity-Based Score. At this time, the method does not take into account peak intensities in the log-likelihood scoring. Several schemes have been used for accounting for peak intensities. One method is to sum the intensities that can be accounted for by a given peptide, and then rank peptides by their total intensity. However, this approach is problematic. A peptide that can only account for a single large peak in the spectrum may outscore a peptide that accounts for three significant but smaller intensity peaks. In this approach, there is a tradeoff between total intensity and the number of peaks that can be accounted for.

Instead, as a measure of how well a peptide can account for the ion current in the spectrum, we measure the extent of the spectrum's total ion current that is covered by the peptide by the entropy statistic

$$\Omega_I = -\sum_i I_i * \log(I_i) \qquad (6)$$

Here, I_i is the relative intensity of peak *i*, and the summation is over all peaks that a given peptide can account for. The entropy score will be largest for peptides that cover the greatest extent of the spectrum's ion current. Ideally, this could include all internal fragments and other minor peaks, but at this point we limit ourselves only to the major peaks found in the training data (Figures 2 and 3).

Support Vector Machine. As discussed above, we calculate several decision criteria on which to decide which peptide, if any, is a correct match to a spectrum. In particular, we are interested in these six critieria: the likelihood ratio, Λ or Ω_p; likelihood ratio difference between the top two scoring peptides, Λ_{12} or Ω_{P12}; the intensity statistic, Ω_I; the rank of Ω_I; the total number of histidines, lysines, and arginines in the peptide; and finally, the total number of lysines and arginines. The number of basic residues are used because they impose thermodynamic restraints on the charge state of the parent ions.

We combine these metrics into an overall score using a support-vector machine (SVM). SVMs have been noted as an excellent generalized supervised learning approach to classification based on statistical optimization theory developed

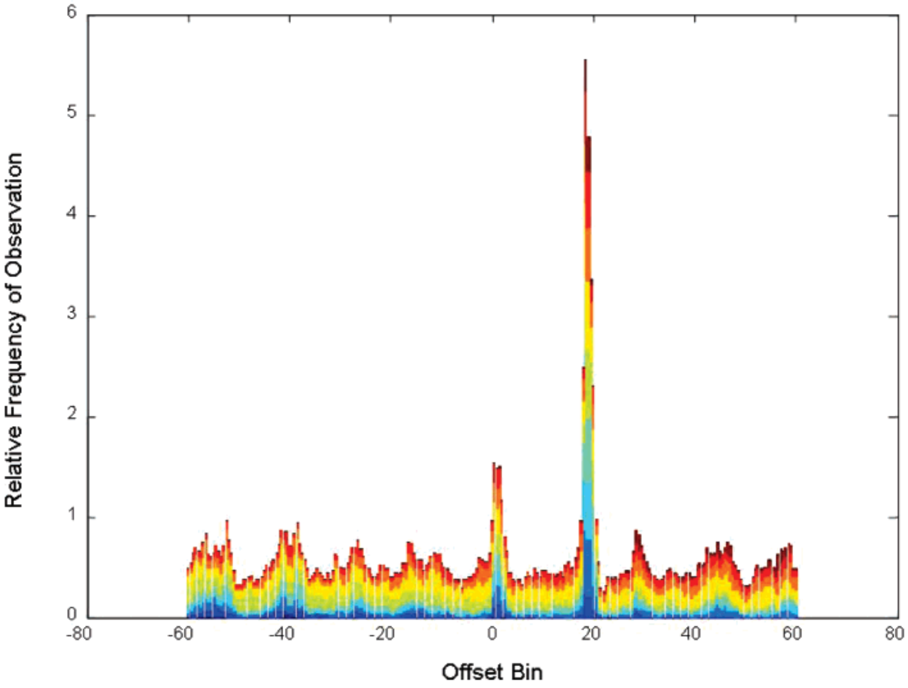

Figure 3. Histogram of ion frequencies versus offset bin for C-terminus partial peptide sequences generated from 10mers. Individual histograms for ion offsets for each partial peptide from length 1 to 9 are colored and stacked to present a summary view of the ion offset patterns that are found.

by Vapnik[23] and others.[24–27] An SVM is exceptionally attractive in this application as it is able to handle both large datasets and noise, and has been used by Anderson, et al.[28] to evaluate peptide identifications using *SEQUEST*-generated scores. The SVM will hence yield a single score associated with a peptide based on the six parameters defined. The decision boundary is defined as the function

$$f(x) = \sum_{i=1}^{N} \lambda_i K(x,x_i) + b$$

where the SVM draws an optimal hyperplane in a high-dimensional feature space determined by λ and b. The classification of a peptide as present or not is based on the sign of this function, i.e., the peptide is in the sample if $f(x) > 0$ and not in the sample if $f(x) < 0$. The function $K(\mathbf{x},\mathbf{x}_i)$ is called the kernel, which allows the problem to be embedded in a higher dimensional space. We utilize a quadratic kernel with a constant of one and a coefficient of 10 and the sequential minimization optimization algorithm of Platt to perform the SVM.[24,26]

Code Implementation. The analysis was prototyped in Matlab and implemented in C as the program *NWPolygraph*. Two versions of the production code exist, a serial version and a parallel version, which differ only in the top-level routine. The serial code is written in ANSI standard C and should run on any computer with an ANSI standard C compiler including any Unix/Linux platform.

On a single processor HP L1820 with an IA64 900 MHz chip and 961 MB of memory, the analysis takes two seconds per spectrum to evaluate all tryptic peptides from a database of 88 000 proteins that match within 3 daltons of the predicted parent peptide mass. The number of candidate peptides in case of the dataset analyzed here is approximately 200 000. When considering all possible peptides that match within 3 mass units of the parent peptide mass, the number of candidates for this data set increases to approximately 3 million and the analysis takes approximately 30 s per spectrum.

The parallel version of the code is designed to score many (thousands to hundreds of thousands) spectra against a reference database of similar size in parallel. This version of the code can compare a spectrum with tryptic candidates in the NR database in approximately 40 s. The MPI programming paradigm is used with dynamic scheduling and a shared global disk resource. The shared global disk system used is the *LUSTRE* file system (http://luster.org) on the HP Linux cluster, Mpp2, in the Molecular Science Computing Facility in the Environmental Molecular Science Laboratory at Pacific Northwest National Laboratory. Mpp2 is a 980 node/1960 Itanium-2 processor machine recently put into place at PNNL. We intend to make the program freely available for academic and government research.

Results and Discussion

Fragmentation Model Development from Training Data. Ultimately, we are striving to develop fragmentation models that can be reconciled with statistical mechanics. At this point, we address only the charge and position dependence of fragmentation processes. Consequently, the training data set was partitioned according to charge and length of the parent ion. Each partitioned set consisted of approximately 300 spectra. For each set, the MS/MS fingerprints are constructed from the partial peptide masses and most frequent ion offsets as described in the previous section, where the bin width is set to 0.5 μm. Figures 2 and 3 illustrate the cumulative offset frequencies for a test set of 10-mer spectra as a function of offset from the N- and C-termini, respectively. The figures represent a histogram of offset frequencies constructed from many spectra and are a summary view in that the peaks in Figures 2 and 3 represent the cumulative effect across all

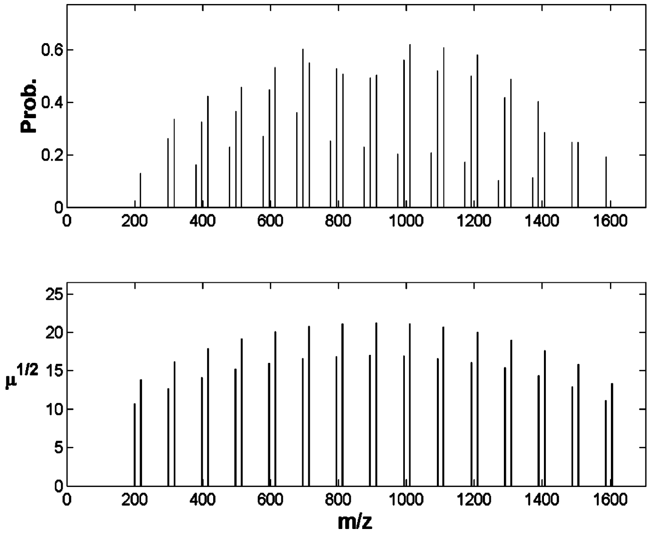

Figure 4. Comparison of fragmentation patterns independently predicted for polyvaline from the training set (top plot; details in *Preliminary Results*) and a simple theoretical model (bottom) in which the number of product molecules produced due to fragmentation at a peptide bond is proportional to the reduced mass at that peptide bond. In the theoretical model only *b* and *y* ion fragments are shown, and the *b* ion fragments have been scaled to be 80% of the *y* ion fragments.

positions in 10-mer peptides. Each bin shown in Figures 2 and 3 thus represent approximately 300 × 10 events. The effect at each position across the peptides is represented by the color-coding in the figure. Consistent with work by Dančík, et al.[11] and Havilio et al.,[14] and with common assumptions about the frequency of appearance of the principal ion types, the most prominent offsets observed in this test set correspond to the *y*, *y*-H_2O ($-NH_3$), *b*, and *b*-H_2O ($-NH_3$) ions. Due to the instrument resolution and the inherent averaging of offsets without regard to sequence composition, peaks for the ion types *y*-H_2O and *y*-NH_3, and ion types *b*-H_2O and *b*-NH_3 were not well resolved from each other. The ion offsets are very consistent across the daughter fragments of all peptides, and are also consistent with the most frequent ion types reported by Dančík, et al.[11] and Havilio, et al.[14] In particular, the most frequent offsets for the C-terminus ions are consistently 19 and 1, and for the N-terminus ions are consistently 1 and −17. In addition to these latter two ion offsets, significant peaks were also observed at approximately −27 from the N-terminal fragment, which are the *a* ions which result from a loss of CO.

The probability of a fragmentation varies considerably as a function of position along the peptide backbone. Fragmentation events are relatively less likely near either terminus and much more likely toward the middle of the peptide. Figure 4 (top) displays the fragmentation probabilities along the backbone of polyvaline. The probability of a fragmentation event in a real peptide is a function of the energy landscape of the peptide and its dynamics. Formation of secondary structure of the peptide in the gas phase can have a strong affect on the probability of a fragmentation, as can mass effects at the fragmenting bond.

Previous work has sought to characterize the fragmentation probabilities as a function of the mass.[14,22] To make this connection between mass and properties of a labile bond, knowledge of the ensemble of three-dimensional structures of the peptide must be used in calculations at some level of molecular theory. However, the *reduced mass* can be directly related to the frequency of the vibrating bond in a model system with out resorting to structural models. As an example, consider the simple model of a peptide in which a peptide bond *i* is modeled as a spring with force constant *k*. The classical energy at each model bond is given by

$$\begin{aligned} E_i &= h\nu_i \\ &= \frac{h}{2\pi}\sqrt{\frac{k}{\mu_i}} \end{aligned}$$

Here, *h* is Plank's constant and the reduced mass at peptide *i*, μ_i, is determined by the mass of the groups that are N-terminal to the bond (m_N) and the masses that are C-terminal to the peptide bond (m_C) by

$$\mu_i = \frac{m_N m_C}{m_N + m_C}$$

A striking correlation of fragment ion abundance derived from the training data with the reduced mass at the peptide bond is demonstrated in Figure 4. The top plot in Figure 4 shows the fragmentation pattern predicted from the training data for the model peptide polyvaline. The bottom plot in Figure 4 shows the fragmentation pattern predicted independently from a simple theoretical model in which the number of product molecules produced at each peptide bond *i* is proportional to $\sqrt{\mu_i}$. Additionally, analysis of the characteristic motions of simple peptide models in extended conformations (coupled harmonic oscillators) would demonstrate that the bonds with the largest reduced mass also have the largest amplitude motions along the reaction coordinate. The correlation shown in Figure 4 is present because the training data treat the effects of three-dimensional structure in an average manner, and the reduced mass likewise consolidates the position and mass information into a single statistic. In this sense, the relationship is general and not limited to any specific peptide.

A simple way to interpret this is that the frequency of the bond vibration increases inversely with reduced mass. For two bonds that differ only in their reduced mass, the bond with the greater reduced mass will have vibrational energy states that lie at lower energy levels. Hence, vibrationally excited state levels are more easily populated, which leads to a lower reaction barrier and increased rate of reaction.

However, even the association of reduced mass with fragmentation probabilities is not straightforward. Length of the peptide also plays a critical role in that relatively short peptides provide fewer opportunities for internal solvation of so-called "mobile protons".[29,30] Functional groups of amino acid side-chains can also play this role. Thermodynamics tells us that it is not generally possible to separate the effects of mass, length, and amino acid composition on fragmentation probabilities. More specifically, it is generally not possible to break down the free energy contributions to fragmentation into separate mass, length and composition terms. However, it may be possible to develop approximations that are a compromise between nonspecific average probabilities used here and elsewhere and full structural models.

Probabilities for Chance Matches. The probability of a chance match between a predicted fragment and a spectral peak as a function of mass-to-charge ratio is calculated under the null hypothesis using model 1, as described under the *Methods* section. An example is shown in Figure 1 for a parent peptide having a mass-to-charge ratio of 773.8 ± 1.5 m/z. Besides the obvious capability to calculate the by-chance probabilities as a function of mass-to-charge, there are other several important features of the model. As shown in Figure 1, the probability of a by-chance match of a predicted ion fragment to a peak in the spectrum increases toward both the low and high end of the spectrum. At the low end, these peaks are due to small fragments consisting of 1−3 amino acids, and there is an increased chance of observing a match to an incorrect peptide because there is a relatively large population of fragments in which the amino acid order can be juxtaposed to come up with a fragment occurring at the same mass-to-charge ratio as that observed for the correct peptide. This is especially true for C-terminal fragments from tryptic peptides. The majority of these peptides will contain either a lysine or an arginine at the C-terminus, and this reduces the possible number of juxtapositions that would lead to a fragment having that mass-to-charge ratio. For instance, there are only two ways to form a 3-mer peptide that consist of a C-terminal lysine and two other unique peptides, while there are six ways of forming the peptide without the constraint of ordering the lysine at the C-terminal position.

At other positions in the low end, this same phenomenon also leads to a *decrease* in the chance of seeing a match between a peak in the spectrum and a fragment from an incorrect peptide. This is simply a consequence of peptides losing discrete masses when fragmentation occurs. If some locations have an increased probability of a match, then adjacent locations must have decreased probabilities of a match. In fact, this alternating pattern of regions of high match probability and low match probability due to the discrete nature of the fragmentation process can be seen throughout the range of the spectrum.

This phenomenon again becomes accentuated at the high end of the spectrum. In this case, the effect is due to the fact that we are starting with a set of peptides that have very similar mass-to-charge ratios for the full-length peptide, and then in the case of tryptic peptides most of these peptides will have either a lysine or arginine that is fragmented off from the C-terminus to leave a set of N-terminal fragments that are all very close in mass-to-charge ratio.

The probabilities obtained in this manner show very little variability as a function of the database size. Figure 5 shows the probabilities for chance matches as a function of database size for a parent peptide of 773.8 ± 1.5 m/z. The databases were generated by randomly removing sequences from the database of 88 000 proteins used in the study by Keller, et al.[15] As can be seen, the profile of the chance probabilities does not vary when the number of sequences in the protein database is increased from 5000 to 88 000. This is reassuring for building a scoring model in which the rate of misidentifications can be accurately characterized as a function of protein database size.

Model Spectra and Scoring Method. Frequently, empirically predicted spectra are incorrectly referred to as theoretical spectra. Model spectra generated by the use of training sets or expert opinion are based on empirical data, however, and are not based on molecular theory. As the field progresses, it will be increasingly important to keep this distinction in mind, since theory-based calculations[29,30] may increasingly contribute to our understanding of sequence specific fragmentations and formulation of model spectra. Ideally, there should be a correlation between the empirically derived model spectra and expectations from molecular theory, such as the relationship of mass and fragmentation probabilities discussed above.

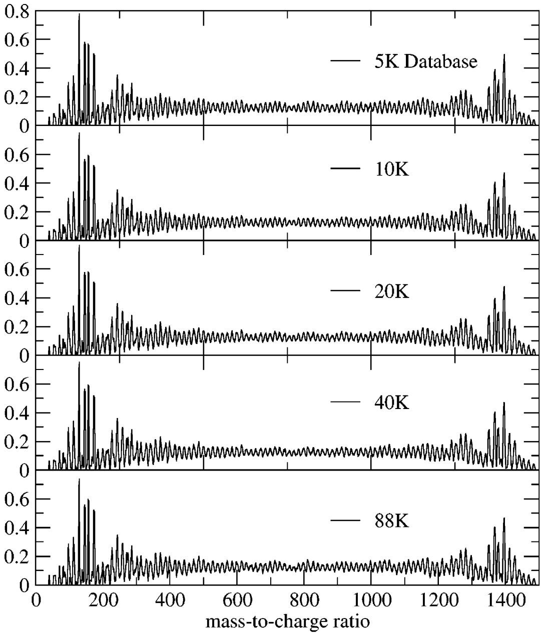

Figure 5. Plot of probability distributions for the null hypothesis model 1 as a function of database size. The database with 5K proteins had 1692 candidate proteins that matched within ± 1.5 m/z of the precursor ion, the 10K protein database had 3288 matching peptide candidates, the 20K protein database had 5420 matching peptide candidates, the 40K protein database had 9775 matching peptide candidates, and the 88K protein database had 17 841 matching peptide candidates.

The peptide scoring method is illustrated in Figure 6. The top plot shows the model spectrum generated for the 14-mer PGIDFTNDPLLQGR. The y-axis of this plot represents the frequency of appearance for each spectral peak, rather than relative intensity typically plotted for MS data. (Given sufficient sampling in a sequence-specific manner in the training data, an experimentally sufficient number of molecules being fragmented in the mass spectrometer, and accurate representation of the number of fragments in the peak intensities, these values should converge.) We substituted frequency of appearance for relative intensity in this plot since relative intensities are not used in the likelihood scoring currently. Rather, the frequency of appearance is the key parameter for scoring each peak. We also note that the frequency of appearance for each peak is different because the offset frequencies are computed separately for each position along the peptide backbone as well as for each fragment ion type.

The bottom plot in Figure 6 illustrates the scoring method. The spectral peaks are plotted in light gray, while the peaks for the candidate peptide PGIDFTNDPLLQGR that match the spectrum are plotted in black. The horizontal line in the top plot shows the probability of observing a peak at any location purely by chance (q_0, eq 1c). In this case, the log-likelihood ratio Ω_p is 21.3, resulting in a correct positive match between the test spectrum and the candidate.

Comparison of Scoring Models. First, we will compare the sensitivity and precision of relative and absolute scores within a scoring model, and then we will compare the scoring models to each other.

The comparison between absolute and relative performance of the likelihood and weighted likelihood scoring models are

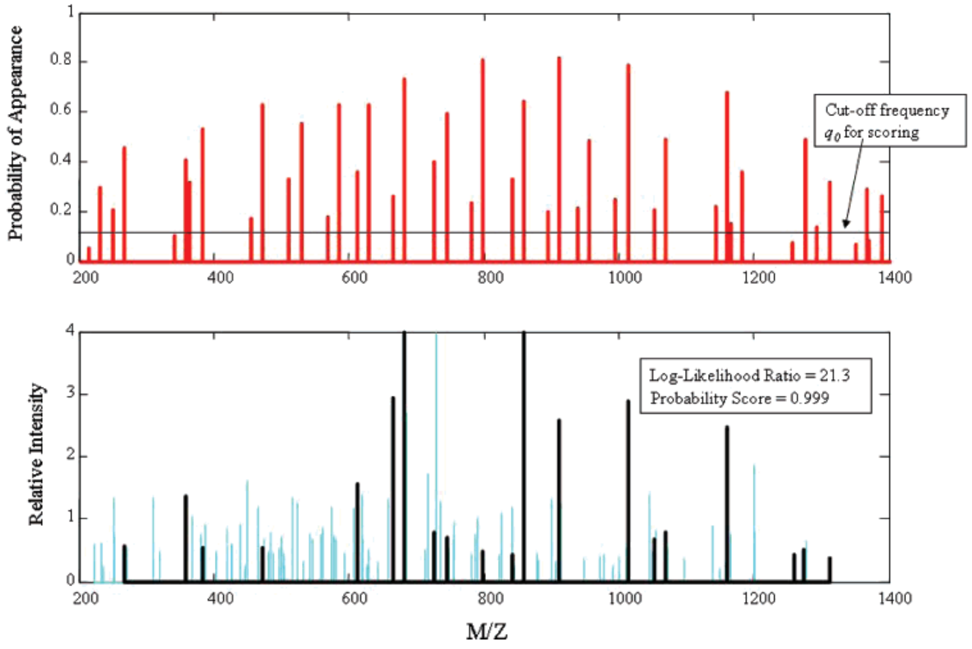

Figure 6. Illustration of peptide scoring method for PGIDFTNDPLLQGR. The top plot shows the candidate fingerprint where peak location is plotted on the *x*-axis and frequency of appearance is plotted on the *y*-axis. The bottom plot illustrates the scoring method on a spectrum for PGIDFTNDPLLQGR, where the gray lines denote nonfingerprint peaks, and the black lines denote observed fingerprint peaks.

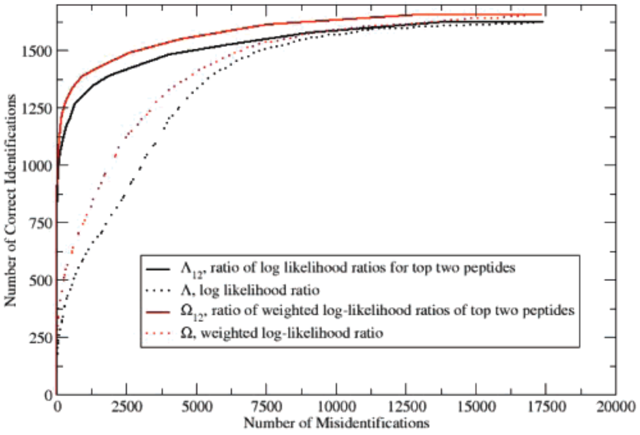

Figure 7. Receiver Operating Characteristic curve comparing identification rates of scoring functions based on absolute log-likelihood ratio scores and relative log-likelihood ratio scores. In both the case of unweighted and weighted log-likelihood ratio scores, the relative score between the top two peptides is a better indicator of a correct identification than are the absolute scores.

determined by comparing the identification rates of Λ with Λ_{ij}, and Ω_P with Ω_{Pij}, respectively. For the relative performance, we will always be comparing the scores of the top two peptides, so we will call these Λ_{12} and Ω_{P12}. We used a previously published dataset of 18 999 spectra for the comparisons.[15] The performance of Λ, Λ_{12}, Ω_P, and Ω_{P12}, are compared in Figure 7 using a Receiver Operating Characteristic (ROC) plot of the number of correct identifications and the number of misidentifications as a function of the cutoff criteria. Since the ratio of likelihood ratio, Λ_{12}, directly compares two peptides, it is reasonable to assume that this would be a more sensitive metric for choosing the best peptide match to a spectrum than the absolute likelihood ratio, Λ. Indeed, as shown in Figure 7, Λ_{12} clearly outperforms Λ. Likewise, the relative score Ω_{P12} also outperforms the weighted log-likelihood score, $\Omega_{P.}$ One reason the likelihood ratios (Λ and Ω_p) underperform is that the scores are very sensitive to the quality of the spectrum. A spectrum for a given peptide containing numerous fragment ion signals will always score much better than a spectrum having relatively fewer signals even if there is sufficient information in the latter to identify the correct peptide. As such, as high likelihood scores primarily reflect the abundant information in the spectrum.

Also, as anticipated, the weighted log-likelihood scoring outperforms the analogous unweighted scoring scheme. This is because major peaks due to fragments such as *b* and *y* ions are more important for identifying peptides than peaks due to neutral loss such as *a*, *b*-H_2O and *y*-H_2O ions.

Next, we compare the information theory metric of the top two scoring peptides Ω_{P12} to the likelihood ratio scoring, Λ_{12}, using different probability models for fragmentations at each of the labile bonds. The probability models that we use are as follows. First, we consider a uniform probability for fragmentation that is independent of the amino acid position in the sequence under H_A. These values are average values obtained from our training set for each ion type and are very similar to those reported in Havillo, et al.[14] and Dancik et al.[11] For the by-chance match (H_0) of a peak in the spectrum to a fragment of the peptide under consideration, we assume that the by-chance match is independent of the location of the peak in the spectrum. We estimated the most appropriate value, 0.12, for this probability from actual matches of incorrect peptides to spectra, using H_0 model 1, as discussed in the methods section and illustrated graphically in Figure 1.

The second model that we use involves independent fragmentation probabilities for each ion type at each position along the peptide backbone under both H_A and H_0. Under H_A, the probabilities used are described in the Methods section and shown in Figures 2 and 3. Under H_0, we use the null hypothesis model 1 probabilities described in the Methods section, in which the frequency of a by-chance match of a peak to

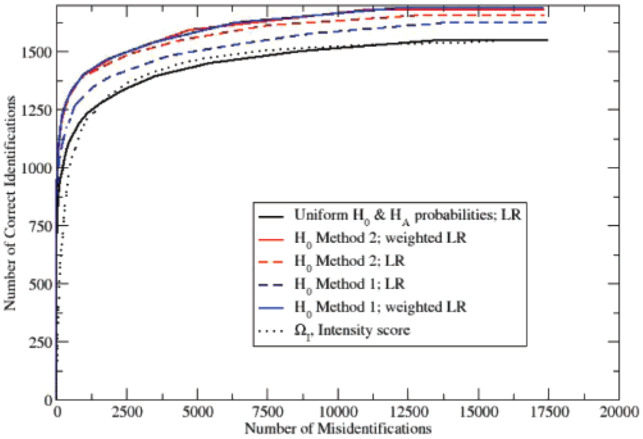

Figure 8. Receiver Operating Characteristic curve comparing weighted and unweighted log-likelihood scores using different probability models for both the null and alternate hypotheses. Intensity score (equation 7) shown for comparison.

fragments from incorrect peptides is estimated directly from the spectrum and the set of candidate peptides.

The third probability model that we use is identical to the second, with the exception that we use the null hypothesis model 2 probabilities under H_0, as described in the Methods section. These values are relatively high at the low and high end of spectra due to the loss of small fragments that can have similar mass-to-charge ratios, and become essentially constant throughout the mid-section of a spectrum, as shown in Figure 1.

Figure 8 shows the performance of these probability models for matching peaks in conjunction with the two scoring models for matching peptides to spectra, Λ_{12} and Ω_{P12}. The use of position-specific probabilities under H_A combined with more realistic probabilities for by-chance matches under H_0 results in a performance that is significantly better than the performance obtained using position-independent probabilities under H_A and a constant for H_0 probabilities. In addition, the weighted log likelihood score Ω_{P12} generally outperforms the unweighted log-likelihood score Λ_{12}.

We also compared the performance of the intensity score, Ω_I, to the likelihood scores. Surprisingly, the intensity metric performs as well as the likelihood ratio score using uniform probabilities, even though the values involved in calculating the score are not derived from training data. The two scores differ somewhat in philosophy. In the likelihood scoring, we were measuring how well a peptide matches an experimental spectrum based on the properties of the model spectrum of the peptide. In contrast, the intensity score measures how well a spectrum matches a particular peptide based on the properties of the observed spectrum. In a rather abstract manner, however, there is a relationship between the probability terms that make up the scores. Assuming that peaks are correctly matched to ion fragments, the key distinction between these terms is that the relative intensity in eq 7 represents the relative probability of formation of each ion fragment if one assumes that transition state theory[31] applies in this case. The transition state theory assumption may not be justified, however, due to the small number of molecules being sampled in the collision chamber. However, possibly more troublesome may be the assumption that the relative intensity is an accurate estimate of the true count of molecules of each fragment type, due to the processing of the raw data.

Next, we benchmark these methods to the cross-correlation approach contained in *SEQUEST* on the independent dataset of 18 999 spectra described previously [15] of which 1662 are known to be due to specific charge 1 and 2 peptides, and 17 337 are due either to known charge 3 peptides or are associated with false hit sequences from the human genome used as a decoy genome. The identifications in this dataset were made using *SEQUEST*, in which the criteria for a correct identification were that the best hit to the spectrum of interest was due to a peptide from one of the control set of protein sequences and a set of *SEQUEST* scores above a predefined cutoff. Tryptic enzyme rules allowing for an unlimited number of missed tryptic termini were used in the database search. However, a number of the reported peptides did not fall into the category of being classic tryptic peptides. A set of 524 peptides either (a) did not begin after a lysine or arginine (b) did not end in lysine or arginine, or (c) neither began after a lysine or arginine or ended with a lysine or arginine. Since the specific enzyme cleavage rules used in the original analysis were not clear, we chose to rerun the same data set specifying tryptic enzyme rules with up to 12 missed cleavages and using *y* ions, *b* ions, *y*-H_2O-(NH_3) ions, *b*- H_2O(NH_3) ions and *a* ions. The *SEQUEST* results are attained through filters generated by experts on the *SEQUEST* score parameters.[32] We used only charge +1 and +2 peptide hits for the comparison because with *SEQUEST* scores there is no clear way to choose between a top scoring charge +2 peptide and a top scoring charge +3 peptide. The use of empirically derived pre-filters for ion current and continuity of ion fragment series,[17] and expert-based filters[32] for *SEQUEST Xcorr* (1.8 for charge 1 parent ions, 2.5 for charge 2 parent ions, and 3.5 for charge 3 parent ions) and *DelCN* values (0.8) reduce the dataset to 7850 *Xcorr* values, 1562 of which correspond to true identifications and 6288 which correspond to false identifications. Thus, filtering returned a true positive rate (TPR) = 0.940 and false positive rate (FPR) = 0.363. Our approach does not use a pre-filtering based on intensity or continuation of fragment ion series to reduce the size of the dataset. Starting with 18 999 vectors of six parameters, the sign of the SVM can be used to perform a hard classification similar to the rules used for *SEQUEST*. We perform the SVM using 10-fold cross-validation to assign each peptide a SVM score. Selecting all peptides that achieve a positive score, this filtering reduces the likelihood ratio/SVM dataset to 2983 identifications of which 1425 are true and 1,558 are false. This results in true positive and false positive rates of TPR = 0.857 and FRP = 0.090. It is especially interesting to note the large decrease in false positives from 6288 for *SEQUEST* to 1558 for the likelihood ratio/SVM approach while the number of true positives retained is only 137 less than for *SEQUEST*. This is a small fraction of true positives lost by the SVM compared to the extra 4730 false identifications that were eliminated.

To observe a more comprehensive comparison of the sensitivity versus specificity of each algorithm we generate the Receiver Operating Characteristic (ROC) curves, Figure 9. To attain a ratio of true to false peptide identifications similar to that of *SEQUEST* this ROC curve is generated by using a randomly selected set of 6690 false identifications and all true identifications, 1662. The measure of quality of a classifier can be quantified by the area under the curve. This results in areas relating to very accurate classifiers of 0.973 and 0.959 for the likelihood/SVM method reported here and *SEQUEST*, respectively. To assess the statistical significance of the difference among the two methods, a two-tailed signed rank test[33,34] was

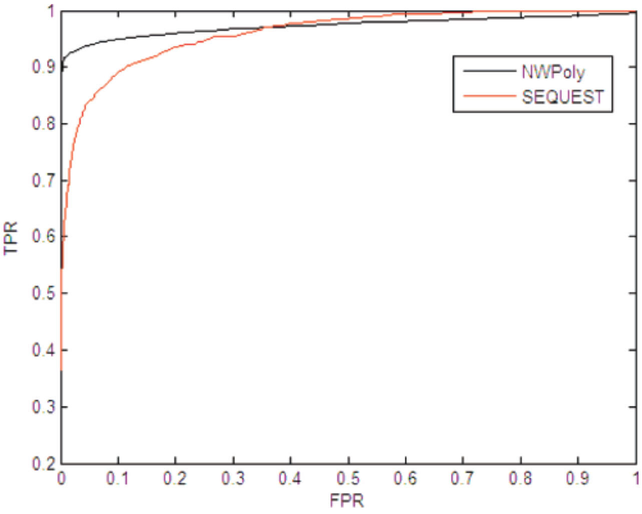

Figure 9. Distributions of true positives (TPR) and false positives (FPR) for the collection of 18 999 spectra using *SEQUEST* and the likelihood/SVM method implemented as the program *NWPolygraph* (NWPoly). In both cases, classic tryptic peptides from the sequence database are used as candidate peptides, 12 missed cleavages are allowed, and *b*, *y*, *b*-NH_3/H_2O, *y*-NH_3/H_2O, and *a* fragments are used in matching peptides to spectra.

utilized. The signed rank test evaluates the hypothesis that the difference between the two curves comes from a distribution with a median of zero. This comparison determined that the ROC curves for are significantly differerent with a *p*-value of approximately 4×10^{18}. From Figure 9, it is clear that this approach achieves a higher sensitivity when requiring higher specificity.

Future Work. There are several areas that would increase the discriminatory power of the analysis. First, the fragmentation model and model spectrum can be made more specific by making them sensitive to the composition or sequence of the peptide. For instance, in the current fragmentation model, no distinction is made between peptides that either contain or do not contain serine, threonine, glutamic acid or aspartic acid. These peptides are the most likely to undergo a neutral loss of water and have peaks for the ion series *y*-H_2O and *b*-H_2O. Likewise, peptides containing side chain amine groups are more likely to undergo a neutral loss of NH_3. No attempt to use this information was made in our initial study.

Along these same lines, the incorporation of conditional probabilities for the appearance of *y*-H_2O and *y*-NH_3 ions based on the presence/absence of *y* series ions may provide greater discriminatory power. This is part of a more fundamental problem in that the probabilities of appearance of peaks in a spectrum may not be independent of each other, as is generally assumed in current database search methods that use likelihood scoring. A more appropriate model may be a likelihood model that takes into account the ion fragment dependencies in a conditional probability network such as the Bayesian network used by Frank and Pevzner in de novo peptide identification.[13] The characterization of these dependent probabilities under the alternate hypothesis is fairly straightforward when the effects of sequence composition are treated in an average manner. However, more challenging is the characterization of the null hypothesis probabilities and the formulation of prior probabilities that incorporate additional chemical information.

Extending the fragmentation model to be sequence specific is also approachable. The most difficult aspect of this will not be the incorporation of such a model into the analysis, but rather the careful analysis of a training set. It is not possible to consider every possible combination of amino acids into peptide sequences because the data is simply not available. For a peptide of length 12, for instance, this would require 20^{12} combinations of amino acids. A feasible approach may be to assume that the effects of protein length and peptide composition are independent, and to then study the problem in a manner similar to that done by Tabb, et al. [22] Although, this would assume an additive free energies for amino acid composition, length, and mass, it may still provide a more accurate estimation of fragmentation probabilities, A significant problem, however, is that gold standard training sets do not exist. For the purpose of training a statistical procedure such as that presented here, it is desirable to have a training set in which the peptides are known a priori, rather than using an independent tool such as *SEQUEST* for the initial identifications in the training set. This can bias the training set so that the new tool is only trained on peptides that are identifiable with existing tools.

We will later report a complete analysis on using a support vector machine to discriminate between correctly identified peptides and incorrectly identified peptides. There are several issues to investigate with regard to which parameters make the most important contributions, the dependencies between parameters, the choice of kernels and optimization steps. In our application of the SVM here, we used the SVM to identify spectra that were correctly matched with peptides from spectra that were incorrectly matched with peptides. In this application, it was assumed that the peptide with the highest likelihood ratio score was the appropriate peptide to consider. However, one application of the SVM that we will investigate and that should increase the number of correctly identified peptides is to also use an SVM to choose the best candidate that matches a spectrum.

Finally, we will report at a later date on the implementation of this peptide scoring procedure in a sequence optimization approach that is similar to de novo peptide identification, but is not limited by incomplete sets of peaks in the MS/MS spectrum.[35]

Summary

The statistical framework presented here is a step toward using highly specific model fragmentation patterns and model spectra to statistically analyze MS/MS spectra and identify peptides with high discrimination. Currently, using nonsequence specific, average fragmentation patterns that depend only on the peptide length and amino acid position, we are able to identify peptides in an evaluation set of spectra significantly better than one current industry standard, *SEQUEST*. The method analyzes peptides of all charge states on equal footing, is fast enough for high-throughput studies, and is flexible enough to be easily extended to identify peptides based on their composition. The current program is fast and can run on any computer with an ANSI C compiler.

Acknowledgment. We would like to thank researchers in the laboratory of Richard Smith at PNNL who generated the datasets herein, as well as Dr. Smith, for providing the data. This research was funded by the U.S. Department of Energy Office of Advanced Scientific Computing Research under

Contract No. 47901, and by the Office of Biological and Environmental Research under Contract Nos. 41966 and 43930. The calculations were performed in the William R. Wiley Environmental Molecular Sciences Laboratory at the Pacific Northwest National Laboratory. The EMSL is funded by the Office of Biological and Environmental Research in the U.S. Department of Energy. PNNL is operated by Battelle for the U.S. Department of Energy under Contract DE-AC06-76RLO 1830.

References

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.

(2) Smith, R. D.; Anderson, G. A.; Lipton, M. S.; Pasa-Tolic, L.; Shen, Y.; Conrads, T. P.; Veenstra, T. D.; Udseth, H. R. An accurate mass tag strategy for quantitative and high-throughput proteome measurements. *Proteomics* **2002**, *2* (5), 513−523.

(3) Wolters, D. A.; Washburn, M. P.; Yates III, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **2001**, *73* (23), 5683−5690.

(4) Resing, K.; Meyer-Arendt, K.; Mendoza, A.; Aveline-Wolf, L.; Jonscher, K.; Pierce, K.; Old, W.; Cheung, H.; Russell, S.; Wattawa, J.; Goehle, G.; Knight, R.; Ahn, N. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **2004**, *76* (13), 3556−3568.

(5) Mann, M.; Wilm, M. Error Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Anal. Chem.* **1994**, *66*, 6 (24), 4390−4399.

(6) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.

(7) Bafna, V.; Edwards, N. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinform.* **2001**, *17*, 13S−21S.

(8) Sadygov, R. G.; Yates, J. R. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Anal. Chem.* **2003**, *75* (15), 3792−3798.

(9) Fridman, T.; Razumovskaya, J.; Verberkmoes, N.; Hurst, G.; Protopopescu, V.; Xu, Y. The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. *J. Bioinform. Comput. Biol.* **2005**, *3* (2), 455−476.

(10) Zhang, N.; Aebersold, R.; Schwikowski, B. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics* **2002**, *2* (10), 1406−1412.

(11) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **1999**, *6* (3/4), 327−342.

(12) Elias, J. E.; Gibbons, F. D.; King, O. D.; Roth, F. P.; Gygi, S. P. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotech.* **2004**, *22* (2), 214−219.

(13) Frank, A.; Pevzner, P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964−973.

(14) Havilio, M.; Haddad, Y.; Smilansky, Z. Intensity-based statistical scorer for tandem mass spectrometry. *Anal. Chem.* **2003**, *75* (3), 435−444.

(15) Keller, A.; Purvine, S.; Nesvizhskii, A. I.; Stolyar, S.; Goodlett, D. R.; Kolker, E. Experimental protein mixture for validating tandem mass spectral analysis. *Omics* **2002**, *6* (2), 207−212.

(16) Lipton, M. S.; Pasa-Tolic, L.; Anderson, G. A.; Anderson, D. J.; Auberry, D. L.; Battista, J. R.; Daly, M. J.; Fredrickson, J.; Hixson, K. K.; Kostandarithes, H.; Masselon, C.; Markillie, L. M.; Moore, R. J.; Romine, M. F.; Shen, Y.; Stritmatter, E.; Tolic, N.; Udseth, H. R.; Venkateswaran, A.; Wong, K. K.; Zhao, R.; Smith, R. D. Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99* (17), 11049−11054.

(17) Eng, K.; McCormack, A. L.; Yates III, J. R. 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.

(18) Harkewicz, R.; Belov, M. E.; Anderson, G. A.; Pasa-Tolic, L.; Masselon, C. D.; Prior, D. C.; Udseth, H. R.; Smith, R. D. ESI−FTICR mass spectrometry employing data-dependent external ion selection and accumulation. *J. Am. Soc. Mass Spectrom.* **2002**, *13* (2), 144−154.

(19) Fernandez-de-Cossio, J.; Gonzalez, J.; Satomi, Y.; Shima, T.; Okumura, N.; Beseda, V.; Betancourt, L.; Padron, G.; Shimonishi, Y.; Takao, T. Automated interpretation of high-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. *Rapid* Commun. Mass Spectrom. **1998**, *12*, 1867−1878.

(20) Huang, Y.; Wysocki, V. H.; Tabb, D. L.; Yates III, J. R. The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Intl. J. Mass Spectrom.* **2002**, *219* (1), 233−244.

(21) Breci, L. A.; Tabb, D. L.; Yates III, J. R.; Wysocki, V. H. Cleavage N-terminal to proline: Analysis of a database of peptide tandem mass spectra. *Anal. Chem.* **2003**, *75* (9), 1963−1971.

(22) Tabb, D. L.; Smith, L. L.; Breci, L. A.; Wysocki, V. H.; Lin, D.; Yates III, J. R. Statistical characterization of ion trap mass spectra from doubly charged tryptic peptides. *Anal. Chem.* **2003**, *75* (5), 1155−1163.

(23) Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

(24) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and other Kernel-Based Learning Methods*; Cambridge University Press: New York, 2000.

(25) Furey, T. S.; Cristianini, N.; Duffy, N.; Bednarski, D. W.; Schummer, M.; Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinform.* **2000**, *16* (10), 906−914.

(26) Platt, J. C. Sequential minimal optimization: A fast algorithm for training support vector machines. *Technical Report MSR-TR-98-14, Microsoft Research* **1998**.

(27) Yang, Z. R.; Chou, K.-C. Bio-support vector machines for computational proteomics. *Bioinform.* **2004**, *20*, (5), 735−741.

(28) Anderson, D. C.; Li, W.; Payan, D. G.; Noble, W. S. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2003**, *2* (2), 137−146.

(29) Dongre, A. R.; Jones, J. L.; Somogyi, A.; Wysocki, V. H. 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*, *8* (35), 8365−8374.

(30) Somogyi, A.; Wysocki, V. H.; Mayer, I. The Effect of Protonation Site on Bond Strengths in Simple Peptides—Application of Ab Initio and Modified Neglect of Differential-Overlap Bond Orders and Modified Neglect of Differential-Overlap Energy Partitioning. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (8), 704−717.

(31) Steinfeld, J. I.; Francisco, J. S.; Hase, W. L. *Chemical Kinetics and Dynamics*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, 1999; p x, 518 p.

(32) Le Roch, K. G.; Johnson, J. R.; Florens, L.; Zhou, Y. Y.; Santrosyan, A.; Grainger, M.; Yan, S. F.; Williamson, K. C.; Holder, A. A.; Carucci, D. J.; Yates, J. R.; Winzeler, E. A. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Res.* **2004**, *14* (11), 2308−2318.

(33) Hollander, M.; Wolfe, D. A. *Nonparametric Statistical Methods*, 2nd ed.; John Wiley & Sons: New York, 1999; p xiv, 787 p.

(34) Salzberg, S. On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Min. Know. Discov.* **1997**, *1* (3), 317−328.

(35) Heredia-Langner, A.; Cannon, W. R.; Jarman, K. D.; Jarman, K. H., Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinform.* **2004**, *20* (14), 2296−2304.

(36) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: Towards high-throughput tandem mass spectrometry data identification *Proteomics* **2003**, *3*, *3*(8), 1454−1463.

PR050147V