research articles Journal of proteome research

High-Accuracy Peptide Mass Fingerprinting Using Peak Intensity Data with Machine Learning

Dongmei Yang,[†] Kevin Ramkissoon,[†] Eric Hamlett,[†] and Morgan C. Giddings*[,†,‡,§]

Department of Microbiology and Immunology, Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, and Joint Department of Biomedical Engineering, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599 and North Carolina State University, Raleigh, North Carolina 27695

Received February 16, 2007

For MALDI-TOF mass spectrometry, we show that the intensity of a peptide–ion peak is directly correlated with its sequence, with the residues M, H, P, R, and L having the most substantial effect on ionization. We developed a machine learning approach that exploits this relationship to significantly improve peptide mass fingerprint (PMF) accuracy based on training data sets from both true-positive and false-positive PMF searches. The model's cross-validated accuracy in distinguishing real versus false-positive database search results is 91%, rivaling the accuracy of MS/MS-based protein identification.

Keywords: ion suppression • mass spectrometry • peptide mass fingerprinting • protein identification • peak intensity

Introduction

Mass-spectrometry-based (MS) proteomics relies upon the accurate identification of proteins, mainly by peptide mass fingerprinting (PMF) and tandem mass spectrometry-based (MS/MS) analysis of peptides. Though PMF was the first commonly used identification method, it suffers from a lack of accuracy compared to MS/MS. PMF is applied by proteolytically digesting a protein into short peptides (~5–30 amino acids), then measuring the mass of the peptides by one of the available methods, such as matrix-assisted laser desorption ionization (MALDI) time-of-flight (TOF) mass spectrometry. The resulting masses (fingerprint) are searched against a gene/protein database to determine which of the sequence entries is most likely to have generated the measured fingerprint (e.g., refs 1–4).

The limitation in PMF accuracy is a side effect of the nonspecific nature of peptide mass matching. Within the mass accuracy of commonly used mass spectrometers, e.g., 30–100 ppm or more, a single peptide mass will match many peptides in a sequence database. The frequent occurrence of false-positive peptide matches reduces the specificity of a PMF search.

Typical PMF search procedures consider only the mass of a match between a spectrum peak and a database peptide, but not the intensity of the peak in the spectrum. However, there is evidence that the peak intensity correlates to sequence properties of the peptide, due to ion suppression.[5,6] This correlation may be exploited by a search engine to distinguish peptides that have a good fit between sequence properties and the measured intensity, yet there do not appear to be simple, definitive rules that predict spectral peak intensity from a corresponding peptide sequence.[7]

To determine how this correlation could be exploited to improve PMF identification accuracy, we collected 234 PMF samples by MALDI-TOF MS of digested spots from gel-separated *E. coli* proteins. We derived statistics showing that both the N-terminal and C-terminal amino acid of a peptide affected peak intensity, along with the internal amino acid composition and the size of the peptide. We then applied a machine-learning approach that uses these statistics to improve PMF identification accuracy. We applied the system to analyze both real and false positive PMFs produced by the GFS (Genome-based peptide Fingerprint Scanning) protein identification system, a search engine developed in our laboratory to match PMF data to unannotated genome sequence.[1] To confirm the results obtained with GFS, we also applied Mascot,[3] which agreed with the GFS identifications for all the real protein samples.

Using 10-fold cross validation, we assessed the ability of our system to improve PMF identification accuracy. It was able to distinguish real PMF matches from false positive search results with 91% accuracy, greatly improving the discrimination power of PMF search over existing methods. To determine which parts of a peptide sequence contribute the most to our model's performance, we assessed model accuracy with each factor considered separately, finding that peptide length and its internal amino acid composition contributed the most to the model accuracy.

We performed a further test of the model using a validation data set that was never used for training, composed of chromatographically separated ribosomal protein fractions from *Escherichia coli*. These samples were analyzed by peptide

* To whom correpondence should be addressed. CB#7290, 804 Mary Ellen Jones, Chapel Hill, NC 27599, USA. Tel, (919) 843-3513; Fax, (919) 962-2963; E-mail, giddings@unc.edu.
† Department of Microbiology and Immunology.
‡ Department of Computer Science.
§ Joint Department of Biomedical Engineering.

10.1021/pr070088g CCC: $40.75
Published on Web 10/03/2007

mass fingerprinting, collecting both MS data and MS/MS data for selected peptides on a MALDI-TOF/TOF instrument. Because of the complex nature of these samples, each of which contained multiple proteins, standard PMF search in both Mascot and GFS performed poorly at identifying the ribosomal proteins in them. However, using just the MS data, our system was able to identify ribosomal proteins with a high accuracy, rivaling the identification results produced by the MS/MS search results. Notably, the model identified several high-scoring ribosomally associated proteins in these fractions that were not identified by MS/MS-based search using Mascot or GFS.

Materials and Methods

Laboratory Methods. Peptide mass fingerprint data for training and testing the system were obtained by 2-D PAGE and MALDI-MS/MS from wild type *E. coli* strain MG1655 and two laboratory generated streptomycin resistant derivative strains, following the procedures described in the Supporting Information S1. Tryptic peptide digestion data were obtained for 234 protein spots on an Applied Biosystems 4700 MALDI-TOF/TOF (Foster City, CA).

Protein Identification. We searched the *E. coli* MS and MS/MS data for matches against *E. coli* proteins, using GFS,[1,9] along with Mascot[3] for verification. GFS calculates theoretical peptides from an *in silico* translation and digestion of an entire genome sequence and then locates windows where a statistically significant number of matches have occurred. The identified sequence may or may not correspond to an annotated gene, depending on the accuracy of gene annotation. However, for the searches performed here, we filtered out all results except those matching an annotated gene, producing results that were qualitatively equivalent to those produced by Mascot.

The 234 mass lists were searched by GFS against the *E. coli* strain K-12 MG1655 genome[10] using a mass tolerance of 100 ppm and a window size of 500 nucleotides. Proteins were also identified by Mascot searching against the Mass Spectrometry protein sequence Database (MSDB).[11] The TOF/TOF system generated MS/MS data for selected peptides, which were used in the Mascot search to verify protein identities for the true-positive data set. However, these MS/MS data were not used in the training or testing of the model, since its goal is to increase PMF search accuracy using only MS data. We selected 100 top-scoring mass lists as true-positive samples, that had both a significant Mascot score ($p < 0.05$) and a significant GFS expectation value ($E < 0.0005$), the latter calculated by the method of Fenyö and colleagues.[12,13] We verified that the GFS and Mascot identifications were the same in each case, comparing the gene annotation for the match produced with GFS to the protein matched by Mascot.

False-Positive Data set. For each of the 100 true-positive mass lists, we generated randomized mass lists of the same size by randomly picking masses and intensities from the 234 real mass lists. Each randomly generated mass and intensity pair was generated as follows. First, a random peptide mass was selected from the 234 real mass lists. Then, a corresponding peak intensity was picked from a different peptide that had a mass of at least 10 but less than 30 Da away from the first selected mass. The rationale for this procedure was to break the correlation between the peptide sequence composition and the measured peak intensity, because there should be no such correlation in false matches, while retaining the overall trend of decreasing peak intensity with increasing mass. The minimum of 10 Da was used to ensure that a different peptide was picked for intensity than the one used for the mass.

Each list of peptides generated by the above procedure was searched by GFS using the same parameters as for the real data, and new lists of identical size were repeatedly generated until we obtained one having a significant sequence match ($E < 0.0005$) that fell within an annotated gene. In addition to excluding matches against unannotated sequences, we also eliminated matches to the same gene that one of the true-positive lists matched, which would occur occasionally if the random selection picked a set of masses that happened to have many masses in common with one of the real sample lists.

The result of this selection process was that for each of the 100 true-positive lists, a high-scoring false-positive mass and intensity list was produced with the same number of peaks as the true-positive.

Ribosomal Data set. We also obtained a data set from a ribosomally enriched sample, separated by reversed phase liquid chromatography and collected into a series of fractions, each of which was analyzed by MALDI-TOF/TOF MS/MS on an Applied Biosystems 4700 (Foster City, CA). Detailed procedures are described in Supporting Information S1.

GFS was used to produce a list of top-20 PMF matches for each of 72 ribosomal samples in the data set, searching against the *E. coli* MG1655 genome sequence. We assigned as positives all PMF matches against ribosomal proteins, i.e., those known to be ribosomally associated according to Ecocyc,[14] and those which had been observed in separate MS/MS analyses of ribosomal fractions by our group. This produced 124 positive cases in the data set, along with 740 negatives. These were used for further testing of the model, as described below.

Data Availability. The described data sets are downloadable from our GFS Web site http://bioinfo.unc.edu/Downloads/files/DataPackage.zip.

Intensity Range Binning. We classified all the matched spectral peaks for each identified protein into one of three categories: low, medium, or high intensity. The peaks with top one-third of the intensities were assigned as high-intensity peaks, the peaks with bottom one-third of the intensities as low-intensity peaks and the rest as medium intensity peaks. This is illustrated in Figure 1.

Statistical Assessment of Data. We examined amino acid frequency distribution at the N terminus, C terminus, and internal to the sequence, for each of the matched peptides in the true positive data set, to determine the nature of the correlation between sequence and peptide peak intensity in MS. For each of the three intensity bins (I1-low, I2-medium, I3-high), we calculated the frequency of each amino acid residue. To normalize the frequencies, we divided the frequency of occurrence for each amino acid in each intensity bin by the overall frequency of the residue in all matching sequences. This procedure was performed separately for the N termini, C termini, and internal amino acids. For each calculated value, we approximated 95% confidence intervals using the modified WALD method[15] to assess the sufficiency of the data to produce useful statistical results.

Model Description. The input to the model is a set of masses and their associated intensities, from a single PMF experiment, along with the corresponding sequence region identified by GFS as the best-scoring match, including a list of the peptide sequences that match peak masses from the spectrum, as illustrated in Figure 1. It considers one such match region at a time, producing a score for the list of matched masses and their

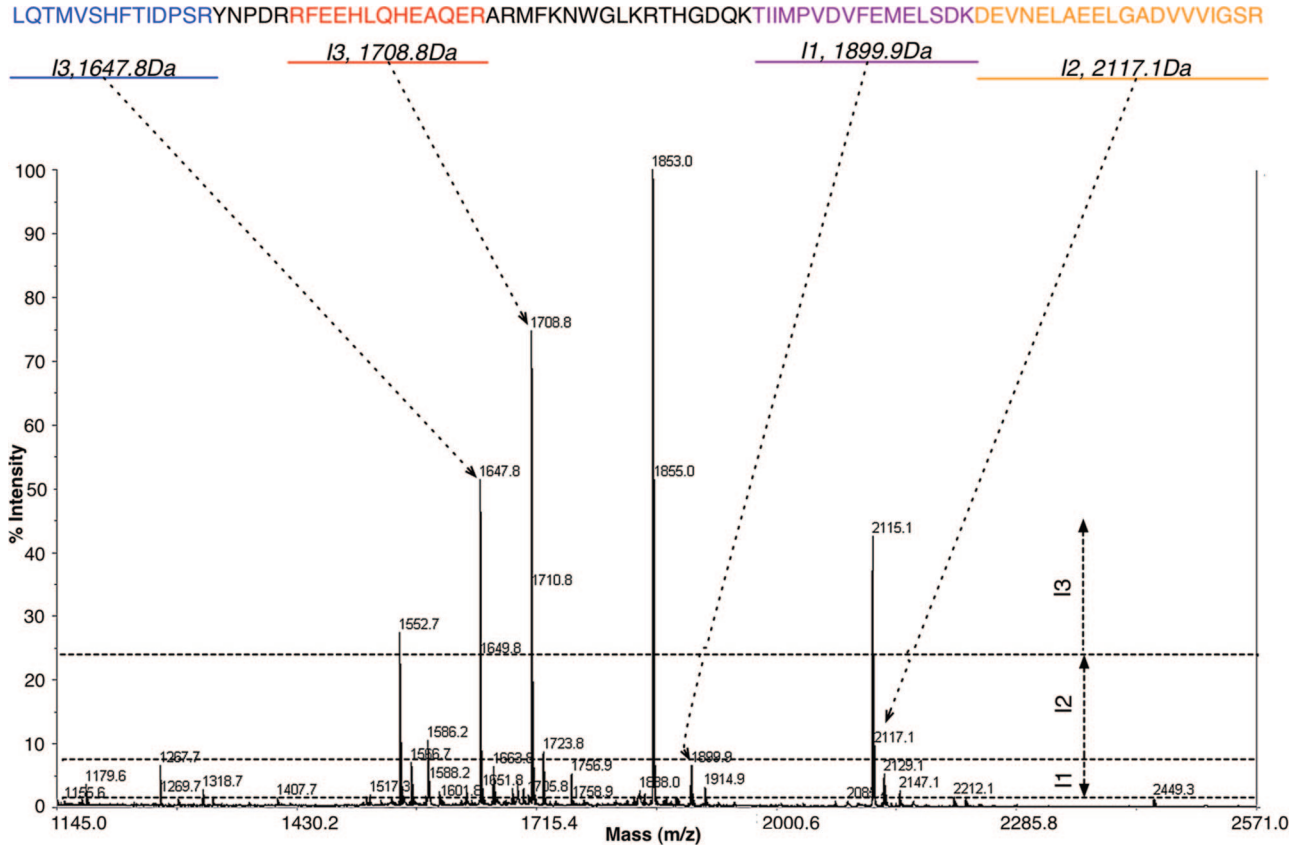

Figure 1. Model considers all peaks whose mass corresponds to a peptide in the matching sequence region (dotted arrows from peaks to sequence). Peaks are categorized into one of three intensity states: I1 (low intensity), I2 (medium intensity), or I3 (high intensity), as shown by the horizontal dotted lines. The bottom horizontal dotted line is the S/N (signal-to-noise ratio) cutoff line. The model considers the intensity category and sequence for each matched peak produced by searching the PMF.

sequences to determine whether the set of matches is due to a real protein or a false positive result. Specifically, for each spectral peak matched to a sequence, the program considers a triplet consisting of peak mass, peak intensity, and the matched peptide sequence, with all such triplets from the matched sequence comprising a set **P**. Upon the basis of an observable Markov model design, the system is used to score **P**.[8] The model is composed of three states, with an optional transition probability score between states and emission probabilities for each peptide sequence within the states. The model traverses the states according to the order of the peptides matched to the sequence, from left to right, with the transitions going between two adjacent matched peptides. Hereafter we refer to our peptide-based model as PMM, derived from Peptide Markov Model (though the Markov label does not necessarily apply, as addressed in the results and discussion section). The model is illustrated in Figure 2.

For each matched peak triplet, the model traverses into one of three states: I1, I2, and I3, corresponding to one of the three intensity categories into which peaks are divided. For each of the three states, there is a set of emission probabilities used to calculate the probability that a peptide sequence was produced by a peak in this intensity range (these are obtained as described in the Training Section, below). Specifically, the model independently considers the emission probability for each character at the N terminus, the C terminus, and internal to a matched peptide. The emission probabilities of residues differ for each of the three intensity states; for example, in I3 (the highest intensity state), it is more likely to have an Arginine than a Lysine present at the C terminus, as shown in the Results section.

We divided the sequence emission probabilities into three regions because prior research indicated that the residue at the C terminus (either Lysine or Arginine for trypsin digestion) had an effect on intensity.[16] Though we were unaware of prior evidence regarding the effect of the N-terminal residue on peak intensity, we decided to separate its emission probability in case it has an effect similar to that of the C terminus. Because we observed a loss of intensity with increased peptide mass based on our experimental data (Supplementary Figure S2), the model incorporates peptide length as part of the emission probabilities used.

Training: Derivation of Transition and Emission Probability. All transition and emission probabilities are derived from the true and false-positive examples from the training data sets. We used two independent models, PMM^+ and PMM^-, each with their own model parameters (transition and emission probabilities). PMM^+ was trained using the true-positive data, and hence should recognize patterns characteristic of real peptide fingerprint matches, whereas PMM^- was trained using the false-positive data set, to recognize the patterns present in random matches. As described below, these two models are used in tandem to score a peptide mass fingerprint, producing a log odds ratio of the PMM^+ to PMM^- score.

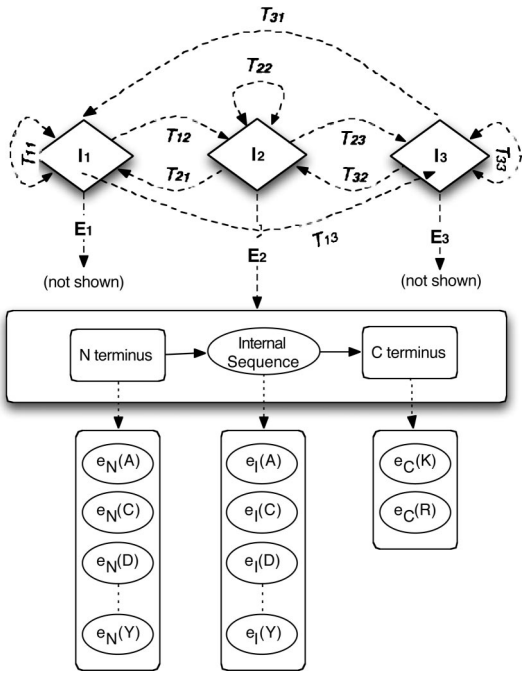

Figure 2. Model Architecture. The states I1, I2, I3 correspond to low, medium, and high-intensity peaks that were matched to a peptide sequence by mass. There is a transition probability T_{ij} between any two states *i* and *j*, as the model scans from one to the next matched peak along a sequence, though for some simulations this value was excluded from model scoring (i.e., the order of peptides in a protein was ignored, as indicated by dotted lines for transitions between states). Within each state, there is a set of emission probabilities *E* for the amino acids in the matched peptide sequence, divided into three components: the N-terminus (e_N), the C terminus (e_C), and the internal sequence (e_I). For each of these, there are independent emission probabilities for each of the amino acids. At the C terminus, only K and R are considered, corresponding to trypsin cleavage. The transition and emission probabilities are derived from training data as described in text.

Transition probabilities were determined by evaluating the frequency with which each transition occurs by scanning each of the matched PMF examples in the training set. Likewise, emission probabilities were determined by counting the frequency of occurrence for each of the amino acids in the three positions (N terminal, C terminal, and internal) for each of the three intensity states. The state length probability distributions are from a peptide length histogram derived from the training data. All probability distributions were independently derived for each of I1, I2, and I3 for each data set—the true-positive data (PMM^+) and the false-positive data (PMM^-).

Scoring a Sequence With the Models. Each of the trained models scans across a set of peptides from a matched PMF, producing an overall joint probability *P* of occurrence for the complete set of tuples **P**. The model traverses a series of states $\tau_0, \tau_1, \tau_2, \ldots \tau_i \ldots \tau_m$, one per matched peptide tuple in **P**, where $\tau_i \in \{I1, I2, I3\}$ and *m* is the total number of matched peptides. We calculate the joint probability of all peptide sequences and their measured intensities for **P** as:

$$P(\mathbf{P}|\mathscr{M}) = T_0 E(\tau_0|\mathscr{M}) \prod_{i=1}^{m-1} T_{i-1,i} E(\tau_i, S_i|\mathscr{M}) \quad (1)$$

where T_0 is the initial state frequency, τ_i denotes the state, $T_{i-1,i}$ denotes the state transition probability, and $\mathscr{M}$ is the model used (PMM^+ or PMM^-). $E(\tau_i, S_i|\mathscr{M})$ is the emission probability of the peptide sequence in state τ_i (one of I1, I2, or I3) for model $\mathscr{M}$, given by:

$$E(\tau_i, S_i|\mathscr{M}) = L(\tau_i, S_i|\mathscr{M}) e_N(\tau_i, s_{i1}|\mathscr{M})$$
$$e_C(\tau_i, s_{il}|\mathscr{M}) = \prod_{j=2}^{l-1} e_j(\tau_i, s_{ij}|\mathscr{M}) \quad (2)$$

The peptide sequence S_i of length *l* is comprised of a series of characters s_{ij}, each representing one of the 20 common amino acids. Given one of the models PMM^+ or PMM^-, the function $e_I(\tau_i, s_{ij})$ is the probability of emitting the amino acid at position *j* internally to the peptide S_i in the state τ_i, and $e_N(\tau_i, s_{i1})$ is the emission probability for the character at the N terminus of S_i in state τ_i. Finally, $e_C(\tau_i, s_{il})$ is the emission probability for the character at the C terminus of S_i in state τ_i. $L(\tau_i, S_i)$ represents the probability of emitting a peptide of length *l* for the state τ_i.

When scoring a peptide fingerprint match set **P**, eqs 1 and 2 are applied to produce a value $P(\mathbf{P}|\mathscr{M})$ for both $\mathscr{M} = PMM^+$ and $\mathscr{M} = PMM^-$. These are combined into a single log odds ratio:

$$S(\mathbf{P}) = Ln\left(\frac{P(\mathbf{P}|PMM^+)}{P(\mathbf{P}|PMM^-)}\right) \quad (3)$$

This log odds score is expected to be highest for real fingerprint matches since the numerator will be large and denominator small, and lowest (most negative) for false-positive matches where the numerator will be small and denominator large.

Performance Assessment. The system was assessed by 10-fold cross validation, whereby the true and false positive data were combined, then partitioned into 10 subsets of equal size. In turn, each of the partitions was selected as a testing set, and the remaining nine partitions combined into a training set. Model parameters were derived (trained) using the training data exclusively, then tested using the unseen test set for accuracy assessment. This was repeated, and accuracy measures were combined for all 10 partitions.

For each PMF match, a score was produced by eq 3. To assess accuracy of model prediction, it was necessary to choose a threshold *t* that divided those PMF matches to be categorized by the model as real versus those categorized as false positive. Given *t*, we then calculated the number of true positives, (T_p, the number of sequences from the real data set scoring above the threshold); true negatives (T_n, the number of sequences from the random data scoring below the threshold); false positives (F_p, the number of sequences from the random data scoring above the threshold); and false negatives (F_n, the number of sequences from the real data set scoring below the threshold).

Using these, we then calculated the accuracy:

$$\text{Accuracy} = \frac{T_p}{T_p + F_p} \quad (4)$$

the sensitivity:

$$Se = \frac{T_p}{T_p + F_n} \quad (5)$$

and the specificity:

$$Sp = \frac{T_n}{T_n + F_p} \quad (6)$$

To evaluate the overall performance independent of an arbitrary value for the threshold t, we applied receiver operating characteristic (ROC) analysis,[17] whereby we tested a series of thresholds for the score value across its whole range, calculated Sp and Se for each, and plotted them parametrically. The area under a ROC curve is a representation of the overall performance of a model, where there is generally an inverse relationship between sensitivity and specificity as the threshold value is varied. For a perfect prediction model, there is no tradeoff between sensitivity and specificity and the ROC area is 1, whereas a random prediction will produce an ROC area of 0.5.

To determine the optimal threshold for distinguishing between real and randomly matched sequences, we found the value of t producing a maximum in the Matthews correlation coefficient,

$$M = \frac{(T_n T_p) - (F_n F_p)}{\sqrt{(T_n + F_p)\ (T_n + F_n)\ (T_p + F_p)\ (T_p + F_n)}} \quad (7)$$

Validation With Ribosomal Protein Samples. To further assess the model performance on a data set never used for training, we applied it to analyze the ribosomal data set. We trained the PMM on the gel-separated protein data set described previously to generate PMM^+ parameters, and on the randomly generated data set to generate PMM^- parameters. We then used the PMM to score all 864 ribosomal matches produced by GFS, generating ROC curves, and calculating the peak Matthews correlation coefficient to find the optimal score threshold t for separating the data set according to the PMM.

Availability. Training and testing of the model was implemented in a set of PERL scripts that performed postprocessing of GFS output.

All scripts are available on our Web site at: http://bioinfo.unc.edu/Downloads/files/ModelPackage.zip.

Results and Discussion

Statistical Results. We first assessed the correlation between sequence properties of a peptide with its peak intensity when analyzed by MALDI-MS. As shown in Figure 3, we found that at the N terminus, Cysteine (C), Methionine (M), and Proline (P) had a strong negative relationship to peak intensity, with Aspartic Acid (D) and Lysine (K) having a weaker negative relationship. In contrast, Leucine (L), Tyrosine (Y), Valine (V), and Histidine (H) showed a positive correlation with the peak intensity of peptides.

Examination of the residues internal to matched peptides revealed a negative correlation between peak intensity and several amino acids, including C, D, and M (Figure 4). Though W also showed what appears to be a strong negative trend, the small number of W residues produced large confidence intervals that disallowed a firm interpretation. Arginine (R) appeared to be the only internally located residue that had a significant positive effect on intensity. C was rarely detected in the high-intensity peaks even though our protocol used iodoacetic acid to prevent disulfide bonds.[18]

At the C terminus, we found that R residues were much more abundant than K in the high-intensity peptide sequences when compared to the lower intensity peptide sequences (Table 1).

For confirmation, we examined the random false positive data, which as expected did not present such patterns.

Model Results. Having established the influence of residues on peptide peak intensity measured by MALDI-MS, we built and evaluated a machine-learning based model as described in the Methods section, determining whether it could use these correlations to improve PMF accuracy to distinguish real matches from false positive match results.

We first examined how well each component of our scoring function contributes to the model's performance, by dissecting eqs 1 and 2 above into five individual factors: state transition probability, N terminus emission probability, C terminus emission probability, length distribution probability of peptides, and internal amino acid emission probability. For the state transition factor, we took the initial state frequency and state transition into account, as given by:

$$P(\mathbf{P}|\mathcal{M}) = T_0 \prod_{i=1}^{m} T_{i-1,i}$$

To calculate the effects of the other factors, we substituted each of the following for the full emission probability as given in eq 2, i.e.: $P(\mathbf{P}|\mathcal{M}) = e_N(\tau_i, s_{i1}|\mathcal{M})$ for the N terminal emission probabilities; $P(\mathbf{P}|\mathcal{M}) = e_C(\tau_i, s_{il}|\mathcal{M})$ for the C terminal emission probabilities; $P(\mathbf{P}|\mathcal{M}) = \prod_{j=2}^{i-1} e_I(\tau_i, s_{ij}|\mathcal{M})$ for the internal sequence emission probabilities; and $P(\mathbf{P}|\mathcal{M}) = L(\tau_i, S_i|\mathcal{M})$ for the state length probability.

We then applied eq 3 to compute a log odds ratio for the scores produced using each factor individually, and performed 10-fold cross validation to produce the scores for all the testing data as shown in the ROC curves of Figure 5. By calculating the peak Matthews correlation coefficient (eq 7), we were able to define the optimal threshold for the accuracy, sensitivity, and specificity as summarized in Table 2.

The amino acid composition had the most positive effect on the overall model performance (ROC area = 0.94), closely followed by N terminus (ROC area = 0.87). The state length distribution also demonstrated a positive effect on the overall model performance (ROC area = 0.72), with the C terminus making a similarly positive contribution to the model (ROC area = 0.71). The state transition calculation, based on the matched peptide order, had a weak influence on the model performance (ROC area = 0.59). Although we had hypothesized that peptide order might contribute in cases of overlapping peptides due to missed tryptic cleavage, the results indicate that peptide order is not important. We tested 20 different combinations of all these factors and found that models both with and without transition probabilities had nearly equivalent performance, with an ROC area of 0.97 and an accuracy of 91% at the peak Matthews value. The values shown in the last line in Table 2 correspond to a model excluding transition probabilities. Although the system was originally designed as an observable Markov model (hence peptide Markov model, or PMM),[8] the transition probabilities associated with peptide order had no impact on performance, so the final system based only on emission probabilities is not actually a Markov model.

Though most factors contributed positively toward model performance, it was surprising that the C terminus made such a small contribution, because both the authors and others have observed a statistical correlation between peak intensity and the ratio of K versus R at the C terminus. We believe that the threshold we chose for the highest intensity category may have been too low to take advantage of this observation. Krause et al. analyzed the most intense peaks in a series of MALDI-MS spectra, and discovered 94% of those contained Arginine at the

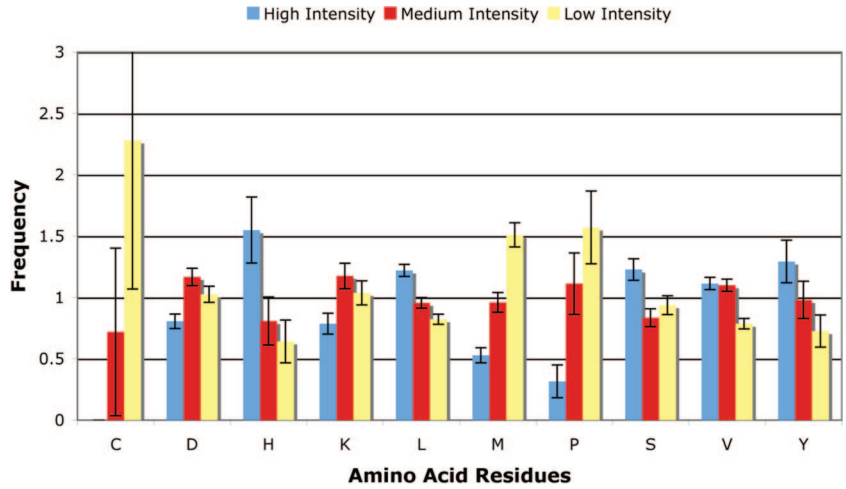

Figure 3. Frequency plot of the N-terminal residues with the greatest effect on peak intensity. The graph shows the normalized frequency distribution for each of the three intensity bins (low-intensity I1 (yellow), medium intensity I2 (red), high-intensity I3 (blue)), shown with approximated 95% confidence intervals. Normalized frequencies were obtained by taking the average frequency of each amino acid residue in I1, I2, and I3 and then dividing the original frequency by the average value. The other 10 amino acids not showing a significant correlation with intensity were omitted for clarity.

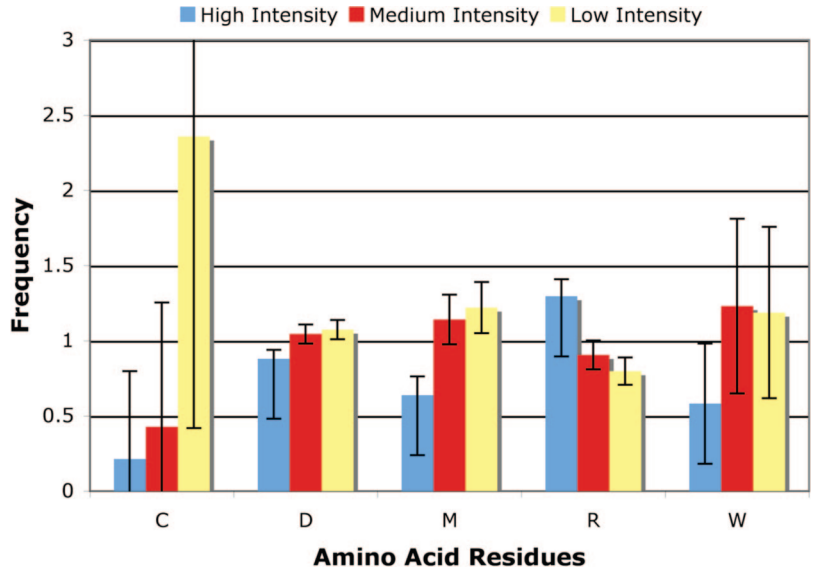

Figure 4. Frequency versus intensity plot for internal sequence residues, for each of the three intensity bins (lw-intensity I1 (yellow), medium intensity I2 (red), high-intensity I3 (blue)), shown with approximated 95% confidence intervals. Normalized frequencies were obtained by taking the average frequency of each amino acid residue in I1, I2, and I3 and then dividing the original frequency by the average value. Only the five amino acids displaying a strong correlation with intensity are shown. The other 15 amino acids were omitted for clarity.

Table 1. Arginine (R) and Lysine (K) Usage Ratio at the C Terminus for Each Intensity State

C terminus	I1	I2	I3
R/K in real data set	0.81	0.68	1.57
R/K in random data set	1.62	1.32	1.31

C terminus.[19] In our work, we only chose the top third of all the peaks as a high-intensity peak, so the percentage of C-terminal Arginine containing peaks was lower, at 62%.

Though the statistics for the internal composition of peptides did not show many strong correlations (Figure 4), the internal composition had a stronger impact than both termini on model performance. This is likely because the internal composition measure includes many more residues than the N or C termini, each of which makes a small but positive contribution to discrimination capability of the model.

Because our results were produced as the log odds ratio of the two model scores, the final scores vary from large negative numbers (false hit) to large positive numbers (true hit). To transform this somewhat arbitrary scale to a more readily understandable measure of how good a PMF match is, we applied logistic regression to transform the score values into a[0, 1] probability of the match being a correct/true match.[20] Supplementary Figure S3 shows the fitted logistic curve that relates the model's score (X-axis) to the probability of true-positive (Y-axis) for the 10-fold cross-validation results. The fit had a Chi Squared value of 62, and a minimal P value ($P < 0.0001$), indicating a very strong correlation between the model's scores and the veracity of a match. The curve indicates

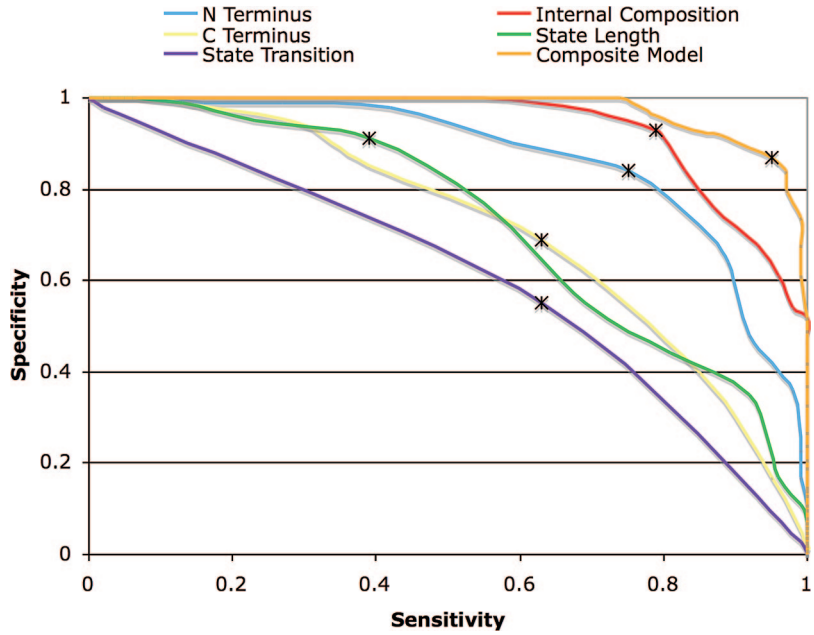

Figure 5. ROC curves for individual factors and the final combined model (all factors except transition and C terminus). See the figure legend at the top for different color lines. * on each line represents the peak Matthew Correlation Coefficient.

Table 2. Summary of the Performance of Different Factors Based on the Threshold Corresponding with the Peak Matthews Correlation Coefficient

factor	accuracy	ROC area	peak M	threshold	*Sp*	*Se*
Internal Composition	86%	0.94	0.73	6	93%	79%
N terminus	80%	0.87	0.59	1	75%	84%
State Length	65%	0.72	0.35	1	91%	39%
State Transition	59%	0.59	0.18	0	55%	63%
C terminus	66%	0.71	0.32	0	69%	63%
Final Model[a]	91%	0.97	0.82	2	87%	95%

[a] The "Final Model"used all factors except for State Transition probability.

Table 3. Top 30 Hits Ranked by PMM Score (8 Matches to the Abundant Ribosomal Protein rpsA Were Omitted to Conserve Space)[a]

ID	GFS PMF E-value	MS/MS tag score	# of tags	protein/ gene	PMM score
B17	5.6e-26	69	4	rpsA	45.4
C15	**7.9e-09**	**0**	**0**	**mopA**	**39.2**
B16	2.1e-17	60	3	rpsA	35.6
C13	**3.1e-05**	**0**	**0**	**adhE**	**25.4**
C21	**3.1e+00**	**0**	**0**	**infB**	**25.2**
B18	3.6e-08	60	3	rpsA	24.1
B19	9.5e-03	27	2	rpsA	18.8
C17	1.9e-12	0	0	rplB	18.6
C22	**9.4e+00**	**0**	**0**	**infB**	**17.7**
B23	3.1e-01	60	3	rpsA	16.9
B2	*2.1e+01*	*15*	*1*	*ypjA*	*16.0*
B22	4.1e-01	60	3	rpsA	15.8
C23	**2.5e+00**	**0**	**0**	**mopA**	**15.4**
C3	5.9e-03	0	0	rplL	13.3
A24	1.8e-02	62	4	rpsC	13.0
B14	3.3e+00	33	1	rpsA	13.0
B22	8.3e-01	19	1	rplL	12.8
C21	*2.1e+01*	*0*	*0*	*yjcX*	*12.7*
B1	1.7e-04	20	2	rpsC	12.5
A14	*3.7e-03*	*14*	*1*	*pntA*	*12.3*
A18	5.1e-08	78	6	rplB	12.0
B14	7.9e-08	18	2	rplD	11.6

[a] Light-face Roman rows represent ribosomal protein matches, bold-face Roman rows represent ribosomally associated protein matches according to Ecocyc, and light-face italic rows are matches against proteins not known to be associated with ribosomes. adhE has been found by MS/MS in our other ribosomal samples, so it was listed as ribosomally associated.

that PMM scores below –10 correspond with a chance of the PMF being true positive at less than 10%, whereas scores above 17 correspond with a chance that the PMF is true positive at over 90%.

Ribosomal Protein Tests. To assess performance on an independent validation set that was never used in training, we applied the PMM to analyze a series of peptide mass fingerprints acquired by MALDI-TOF/TOF from a reversed phase chromatographic separation of a ribosomal protein fraction. ROC analysis on this data set showed high overall model performance, with an area under the curve of 0.945. At the point of peak Matthews correlation, which corresponds to a score threshold of 4, the sensitivity was 68% and the specificity was 95%.

Because the quality of the peptide mass fingerprint spectra were low, GFS had limited specificity using only the PMF data. Using the MS/MS data in sequence tag mode provided more specificity, but limited sensitivity. Supplementary Table S4 illustrates that among the top 100 PMF matches made by GFS (ranked by E-value), 61 ribosomal proteins were identified. Using the PMM score instead to rank the GFS hits increased specificity, resulting in 72 ribosome-associated proteins occurring in the top-100 (Supplementary Table S5). Also, among the lower-half-scoring matches ranked by GFS E-value, there were 24 ribosomally associated proteins. When ranked by PMM score, there was only one occurring in the lower half, indicating that the PMM had significantly better sensitivity.

Table 3 shows the top-30 match scores ranked by the PMM scoring. Of these, 22 matched ribosomal-encoding genes, and 5 more matched genes encoding ribosomally associated proteins, either according to Ecocyc[14] annotation, or previous observations in separate MS/MS analyses of ribosomal frac-

tions. For the three that are not known to be ribosome-associated, it is possible that they copurified with our ribosomes, though further verification would be needed to confirm them.

For further comparison, we submitted the MS/MS data obtained for these PMFs to MASCOT. Although using the MS/MS data allowed Mascot to identify many of the ribosomal proteins present in 72 samples analyzed, there were 8 samples without significant hits. Our PMM model identified ribosomally associated proteins in 4 of those spots (Sample C5, C19, C21 and C22) as true positive hits (above the optimal threshold 4), 2 of them among the top 20 hits (Samples C21 and C22, Table 3). The remaining 4 spots that MASCOT did not identify were scored as negative hits by the PMM.

This analysis illustrates that the PMM performed very well on a separate validation set. It is particularly interesting that the PMM was able to identify several ribosomal proteins that neither GFS nor Mascot could identify using MS/MS data, yet the PMM used only the MS-based fingerprint data. This indicates that the peak intensities contain significant information that can be used to improve PMF search accuracy.

Conclusions

Our work revealed that the locations and types of amino acids in a peptide have a strong relationship to peak intensity for peptide analysis by MALDI-MS. We exploited this relationship by building a machine-learning based model to provide a significant improvement over standard PMF scoring, by distinguishing between correct matches and false-positive hits that occur in a PMF search.

Though PMF searching used by itself has fallen out of favor in much of the proteomics community, it is still often used in combination with MS/MS peptide analysis for identifying proteins separated by gels or other methods. The described method could be used to increase accuracy of combined PMF and MS/MS searches. More importantly, purely MS/MS-based searching relies on accurate peptide identification, which has its own limitations. Our results may be exploited to improve MS/MS searches by determining whether a putative match between a database peptide sequence and the peak intensity of the precursor ion have the expected properties.

Acknowledgment. We thank Dr. David Robinette for providing 2D gel data, analyzing spectra, and Nedyalka Dicheva of the Michael Hooker UNC/Duke Proteomics Core Facility for providing MALDI mass spectrometry analysis. We also thank Dr. Michael Wisz for his early and thoughtful input to the project. Financial support was provided by NIH K22 award from NHGRI HG00044 (M.C.G.), NIH R01 RR020823 from NCRR (M.C.G.), and NSF award MCB-0433977 to Dr. Jacek Gaertig.

Supporting Information Available: (1) Experimental procedures are described in supplementary file S1.doc, (2) Supplementary Figure S2.pdf shows the intensity histograms for different mass ranges, (3) Supplementary Figure S3.pdf shows the logistic regression curve fit, (4) Table S4.xls shows ribosomal resorts sorted by GFS E-value, and (5) Table S5.xls shows the ribosomal resorts sorted by PMM Score. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Giddings, M. C.; Shah, A. A.; Gesteland, R.; Moore, B. Genome-based peptide fingerprint scanning. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100* (1), 20–25.
(2) Mann, M.; Hojrup, P.; Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **1993**, *22* (6), 338–345.
(3) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551–3567.
(4) Wilkins, M. R.; Gasteiger, E.; Wheeler, C. H.; Lindskog, I.; Sanchez, J. C.; Bairoch, A.; Appel, R. D.; Dunn, M. J.; Hochstrasser, D. F. Multiple parameter cross-species protein identification using MultiIdent—a world-wide web accessible tool. *Electrophoresis* **1998**, *19* (18), 3199–3206.
(5) Annesley, T. M. Ion suppression in mass spectrometry. *Clin. Chem.* **2003**, *49* (7), 1041–1044.
(6) King, R.; Bonfiglio, R.; Fernandez-Metzler, C.; Miller-Stein, C.; Olah, T. Mechanistic investigation of ionization suppression in electrospray ionization. *J. Am. Soc. Mass Spectrom.* **2000**, *11* (11), 942–950.
(7) Gay, S.; Binz, P. A.; Hochstrasser, D. F.; Appel, R. D. Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics* **2002**, *2* (10), 1374–1391.
(8) Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc. IEEE* **1989**, *77* (2), 257–286.
(9) Wisz, M. S.; Suarez, M. K.; Holmes, M. R.; Giddings, M. C. GFSWeb: a web tool for genome-based identification of proteins from mass spectrometric samples. *J. Proteome Res.* **2004**, *3* (6), 1292–1295.
(10) Blattner, F. R.; Plunkett, G., III; Bloch, C. A.; Perna, N. T.; Burland, V.; Riley, M.; Collado-Vides, J.; Glasner, J. D.; Rode, C. K.; Mayhew, G. F.; Gregor, J.; Davis, N. W.; Kirkpatrick, H. A.; Goeden, M. A.; Rose, D. J.; Mau, B.; Shao, Y. The complete genome sequence of *Escherichia coli* K-12. *Science* **1997**, *277* (5331), 1453–1474.
(11) MSDB: Mass Spectrometry protein sequence DataBase. http://csc-fserve.hh.med.ic.ac.uk/msdb.html(2/15/07),
(12) Eriksson, J.; Chait, B. T.; Fenyo, D. A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* **2000**, *72* (5), 999–1005.
(13) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75* (4), 768–774.
(14) Karp, P. D.; Riley, M.; Saier, M.; Paulsen, I. T.; Collado-Vides, J.; Paley, S. M.; Pellegrini-Toole, A.; Bonavides, C.; Gama-Castro, S. The EcoCyc Database. *Nucleic Acids Res.* **2002**, *30* (1), 56–58.
(15) Agresti, A.C. B. Approximate is better than "Exact"for interval estimation of biomial proportions. *The American Statistician* **1998**, *52*, 119–126.
(16) Olsen, J. V.; Ong, S. E.; Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **2004**, *3* (6), 608–614.
(17) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143* (1), 29–36.
(18) Sechi, S.; Chait, B. T. Modification of cysteine residues by alkylation. A tool in peptide mapping and protein identification. *Anal. Chem.* **1998**, *70* (24), 5150–5158.
(19) Krause, E.; Wenschuh, H.; Jungblut, P. R. The dominance of arginine-containing peptides in MALDI-derived tryptic mass fingerprints of proteins. *Anal. Chem.* **1999**, *71* (19), 4160–4165.
(20) Hosmer D. W.; Lemeshow, S. *Applied Logistic Regression*; John Wiley & Sons: New York, 1989.

PR070088G