BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 5 2008, pages 674–681
doi:10.1093/bioinformatics/btn011

Gene expression

Incorporating sequence information into the scoring function: a hidden Markov model for improved peptide identification

Jainab Khatun[1], Eric Hamlett[1] and Morgan C. Giddings[1,2,3,*]

[1]Department of Microbiology and Immunology, [2]Department of Biomedical Engineering and [3]Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Received on July 27, 2007; revised and accepted on January 4, 2008

Advance Access publication January 10, 2008

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: The identification of peptides by tandem mass spectrometry (MS/MS) is a central method of proteomics research, but due to the complexity of MS/MS data and the large databases searched, the accuracy of peptide identification algorithms remains limited. To improve the accuracy of identification we applied a machine-learning approach using a hidden Markov model (HMM) to capture the complex and often subtle links between a peptide sequence and its MS/MS spectrum.

Model: Our model, HMM_Score, represents ion types as HMM states and calculates the maximum joint probability for a peptide/spectrum pair using emission probabilities from three factors: the amino acids adjacent to each fragmentation site, the mass dependence of ion types and the intensity dependence of ion types. The Viterbi algorithm is used to calculate the most probable assignment between ion types in a spectrum and a peptide sequence, then a correction factor is added to account for the propensity of the model to favor longer peptides. An expectation value is calculated based on the model score to assess the significance of each peptide/spectrum match.

Results: We trained and tested HMM_Score on three data sets generated by two different mass spectrometer types. For a reference data set recently reported in the literature and validated using seven identification algorithms, HMM_Score produced 43% more positive identification results at a 1% false positive rate than the best of two other commonly used algorithms, Mascot and X!Tandem. HMM_Score is a highly accurate platform for peptide identification that works well for a variety of mass spectrometer and biological sample types.

Availability: The program is freely available on ProteomeCommons via an OpenSource license. See http://bioinfo.unc.edu/downloads/ for the download link.

Contact: giddings@unc.edu, giddings@med.unc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

One of the foundations of systems biology research is protein identification via mass spectrometry. A central method for protein identification in a proteomic experiment is the analysis of short peptides typically comprised of ~3–20 amino acids. Peptides are derived from enzymatic digestion of either a purified protein sample or of a cellular fraction containing complex mixtures of proteins. In the latter case, the number of peptides generated is high, which requires one or more stages of separation, typically using liquid chromatography as exemplified by MudPIT (Washburn *et al.*, 2001). Tandem mass spectrometry (MS/MS) can then be used to analyze the resultant peptides and hence identify the corresponding proteins (Wysocki *et al.*, 2005).

In MS/MS, individual peptides are chosen for fragmentation, usually by collision induced dissociation (CID), after which the masses of the fragmentation products are measured. MS/MS peptide analysis produces a fragmentation pattern for the peptide that corresponds to its primary sequence (Supplementary Fig. 1). An MS/MS spectrum can be used either to search a database of proteins to find the closest matching amino acid sequence to the observed spectrum (Bafna and Edwards, 2001; Eng *et al.*, 1994; Falkner and Andrews, 2005; Geer *et al.*, 2004; LeDuc *et al.*, 2004; Narasimhan *et al.*, 2005; Perkins *et al.*, 1999; Sadygov and Yates, 2003; Zhang *et al.*, 2002), or to derive a set of *de novo* amino acid sequences compatible with the spectrum (Bandeira *et al.*, 2004; Dancik *et al.*, 1999; Frank and Pevzner, 2005; Horn *et al.*, 2000; Ma *et al.*, 2003; Taylor and Johnson, 2001). In either case, if a sufficient number of peptide MS/MS spectra or their derived sequences are matched to a single gene or protein entry in a database, it will confirm that the protein was present in the sample.

A peptide identification algorithm must take an MS/MS spectrum, comprised of a series of mass and intensity values for its peaks, and determine how each peak corresponds to an underlying peptide sequence that may have generated it. The CID peptide fragmentation process produces spectra containing many different ion types, and when the spectrum is obtained, the ion type for each peak is not known. The goal of an MS/MS database search program is to find the peptide sequence P that best matches the properties of the observed spectrum S, which includes assigning ion types to each matched peak in S.

MS/MS spectra exhibit several traits that make peak assignment difficult, which is the primary reason for the limitations of existing MS/MS search algorithms. Common spectral

*To whom correspondence should be addressed.

interpretation challenges include the widely-varying intensity profiles of the peaks produced by fragmentation, along with the presence of many unpredicted peaks due to the breakage of bonds besides the primary peptide backbone bond (i.e. CO–NH_2). Also, internal rearrangement of the peptide's chemical structure can occur, producing peaks that remain unidentified (Yague *et al*., 2003).

Peak intensity variability is due to several factors. First, not all sites along a peptide are equally susceptible to breakage—ours and other studies have shown a complex dependency of peptide fragmentation upon the amino acid sequence (Khatun *et al*., 2007; Tabb *et al*., 2004). The resulting variability of peak intensity within a spectrum has a complex and subtle link to the underlying peptide sequence. Further, fragment ions of different ion types have a propensity towards certain intensity and/or mass ranges within a spectrum, producing intensity variability depending on both ion type and position in a spectrum (Khatun *et al*., 2007; Tabb *et al*., 2003).

While there are many successful peptide identification algorithms for MS/MS data, small differences in the accuracy of the methods can have significant impacts on the success of an experiment, due to the large data sets used and large databases searched. Finding a single, correct peptide entry from a database of 10^7 or more possible peptides is challenging given the factors mentioned above. This challenge becomes even more extreme as other researchers and we undertake new efforts to match each MS/MS spectrum against an entire genome sequence in order to locate the original locus responsible for encoding a given peptide. While this strategy has great potential both for enhancing genome annotation and bypassing limitations of existing annotations (Giddings *et al*., 2003; Kuster *et al*., 2001), it is hampered by the difficulties encountered in matching a spectrum uniquely to plant and mammal genomes comprised of $>10^9$ theoretical peptides.

Motivated by these challenges, we developed a highly accurate algorithm for scoring spectra against peptide sequence databases, called HMM_Score. It uses statistically-derived features of MS/MS, including the probability of each fragment ion type, the effects on amino acids on peptide bond breakage, the correlation between mass and the various ion types, and the correlation of various ion types with the intensity of the observed peak, to improve peptide recognition. HMM_Score applies a hidden Markov model (HMM) to extract probabilistic rules directly from a 'training' set of real spectra, to maximize the information available to the system (Rabiner, 1989).

There have been a few other recent efforts to apply HMMs and machine learning to MS/MS spectral analysis. Fischer *et al*. used HMMs for *de novo* peptide sequence interpretation (Fischer *et al*., 2005), modeling the peak intensities as the emission probabilities. They reported improved performance over competing *de novo* sequencing algorithms. Wan *et al*. developed a database-matching algorithm that considered two primary features of spectra, specifically the distribution of peak intensities and the match tolerance of ion types (Wan *et al*., 2006). They demonstrated improved accuracy for peptide identification compared to standard tools such as Mascot Cite (Perkins *et al*., 1999) and SEQUEST Cite (Eng *et al*., 1994).

HMM_Score is different from these previous efforts because it takes advantage of not only peak intensity information, but also the mass distribution of ion types, as well as the effects of different amino acids on fragmentation of a peptide (hence the observation of associated peaks in the spectrum). It jointly uses information from the peptide sequence and the MS/MS spectrum to score a peptide, resulting in a versatile and highly-accurate scoring system.

We comprehensively evaluated HMM_Score on three data sets generated from two types of mass spectrometers, each with a different ionization technique. Data sets consist of MS/MS spectra assigned by existing popular algorithms followed by manual validation. We also tested the program on a dataset of MS/MS spectra from known protein standards. We applied a variety of analyses to evaluate the performance of HMM_Score, including sensitivity analysis, Reciever Operating Characteristic (ROC) curve analysis, Precision-Recall (PR) curve analysis, and box-plot analysis, with HMM_Score showing very favorable performance in each measure compared to existing approaches.

2 METHODS

2.1 Data sets

For evaluation of HMM_Score, we chose several data sets where the identities of the peptides were already known by combining the application of existing programs and manual validation. The modest size of these data sets does not reflect a limitation of HMM_score, but of the need to have data where the answer is known beforehand in order to benchmark the program, as further addressed in the Section 4.

Dataset 1 was used to assess our model's applicability to spectra from a matrix-assisted laser desorption ionization (MALDI) dual time of flight (TOF/TOF) mass spectrometer. It consists of spectra obtained on an Applied Biosystems 4700 TOF/TOF, that were derived from 300 *E.coli* protein fractions, as we previously described (Khatun *et al*., 2007). The analysis generated 3000 spectra, initially analyzed by Mascot to identify them, then further validated by hand to confirm Mascot's assignments, as described in (Khatun *et al*., 2007). This procedure produced 579 validated MS/MS spectra whose identities were known, for assessment of HMM_Score performance. This data set is deposited in Proteome Commons and is available using the GetFileTool from http://www.proteomecommons.org/dev/dfs/GetFileTool.jnlp with the hash key given in Supplementary file Hash_Keys.txt, labeled as '*Hash for Data set 1*'.

Data set 2 was used to assess our model's applicability to data from ion trap mass spectrometers. The data set was downloaded from http://www.ludwig.edu.au/archive/, and consists of MS/MS spectra from the Human Plasma Proteome Project, generated on an LCQ Deca XP ion trap mass spectrometer (Thermo-Finigan, San Jose, CA, USA) using an electrospray ionization (ESI) source. The analyzed peptides were from tryptic digests of human plasma and/or serum proteins subsequently separated by LC, and were of relatively high complexity. The analysis generated 6000 MS/MS spectra, 677 of which were identified by seven algorithms and manually validated, as described in Kapp *et al*. (2005). We used those 677 validated spectra for assessing the performance of HMM_Score, since their corresponding peptide identities were known beforehand.

Data set 3 (known protein standards) was derived from two sources. Subset 3a is comprised of 1986 MS/MS spectra obtained on MALDI TOF/TOF from over 300 known, purified proteins, published on Proteome Commons by Strahler *et al*. This data set was downloaded from http://www.proteomecommons.org/archive/1117680671827/index.html.

We also generated our own data set, subset 3b, comprised of 51 MS/MS spectra generated on the ABI 4700 MALDI TOF/TOF from four purified proteins: Apomyoglobin, Bovine Serum Albumin (BSA), β-galactosidase, Cytochrome C (Sigma, St. Louis, MO) and one his-tag purified protein, CheZ. These spectra and the corresponding masslists are downloadable from http://www.proteomecommons.org/dev/dfs/GetFileTool.jnlp using the hash given in the Supplementary file Hash_Keys.txt labeled as '*Hash for Data set 3'*.

2.2 Databases

For all searches with HMM_Score, Mascot and X!Tandem, we used the UniProt KB/***Swiss-Prot*** Release 54.1 (277 883 annotated protein sequences, Swiss Institute of Bioinformatics, The European Bioinformatics Institute and Protein Information Resource). To assess the effect of database size on search specificity, we also used a database of all annotated proteins in *E.coli* (4279 annotated proteins) (Blattner *et al.*, 1997) for data set 1.

We then generated databases of peptides by performing an *in silico* digest of all proteins present in each database, using digestion rules for the enzyme trypsin, cleaving after lysine (K) and arginine (R) residues. This produced 19 373 678 putative peptides from ***Swiss-Prot*** with up to one missed tryptic cleavage allowed. We also created a 'decoy' database by digesting all reversed protein sequences in each of the databases in order to assess false positive match rates, using the same trypsin rule.

2.3 Search strategy and parameters

For scoring a spectrum/peptide pair with HMM_Score, the peaks in the spectrum were filtered so that only the 100 most intense peaks were used to score a peptide. Searches were carried out against the databases using following parameters: 2 Da precursor ion mass tolerance, 0.5 Da fragment ion mass tolerance using monoisotopic masses and up to one missed tryptic cleavage site. While the TOF/TOF data would allow tighter tolerances to be used, we used these settings for consistency with the ion trap data set results.

2.4 Model

The goal of the model is to score a peptide sequence P against a spectrum S to determine how well they match. It first predicts a set of possible masses for the 11 most common ion types that could be produced by the peptide P, giving a mass list M_p. It then compares M_p against the actual set of masses M_s observed in the spectrum S, using the HMM pictured in Figure 1. It proceeds one-by-one through the masses from the list M_s, for each one visiting a state in the model (circles in Fig. 1). The states correspond to ion types, and the ion type is chosen depending upon the matching mass in the predicted set M_p, for which the ion type is already known. The 'internal' ions displayed in the figure are those resulting from secondary fragmentation of b and y ions, thus having two bond breakages.

To illustrate how matching occurs: if the first mass in M_s is 125.4 Da, and there is a matching mass in M_p of 125.1 Da corresponding to an immonium ion from peptide P, then the first state visited would be the immonium state. If there is no matching mass in the list M_p, then the 'unassigned' state is visited. If there is more than one ion mass in M_p that matches a mass from M_s, then there is more than one possible path through the HMM, and the best one is chosen using the Viterbi algorithm, described in Section 2.6.

The primary utility of the HMM comes from probabilistic scoring of how well the properties of each peak from spectrum S match the predicted properties for the corresponding ion type(s) visited, as described in Section 2.6. For example, y ions typically have the highest intensities. Therefore, if we are visiting the 'y' state due to a match between a predicted y ion from M_P to a spectrum mass from M_s, if the intensity of the actual peak from S is low, it will score poorly, whereas if it is high intensity, it will score well. This is done for each peak in S.

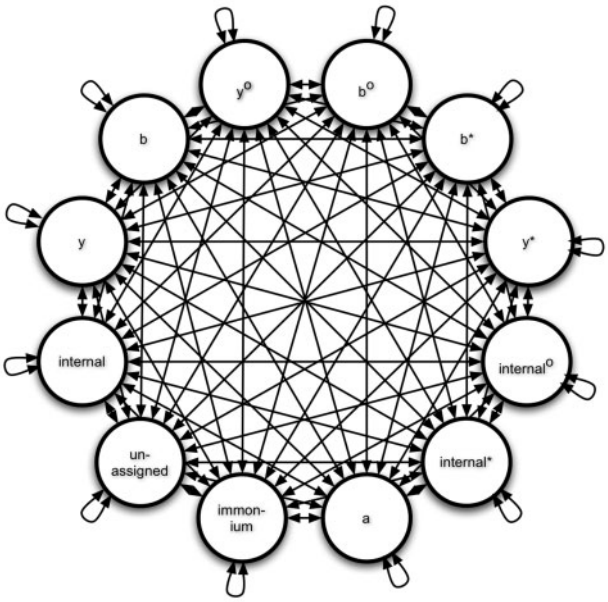

Fig. 1. HMM architecture. The states (circles) correspond to 11 common ion types. The considered ion types are y, b, a and internal ion along with their neutral losses, with ^{0} denoting the neutral loss of water, and * denoting the neutral loss of ammonia from the specified ion type. The unassigned state is included to model peaks that do not match to any of the given ion types. Each state has a set of emission parameters that describe what properties the associated peaks in the spectrum should have for an ideal match to the given peptide sequence. For example, y-ions typically have the highest intensities in a spectrum, whereas internal ions have low intensities.

Additionally, the HMM uses transition probabilities to determine whether adjacent peaks in S and their ion state assignments in the model occur in probable combinations. For example, it might be expected to commonly see a y ion peak preceded by its neutral loss products, y^0 (water) and y* (ammonia). In Figure 1, the arrows leading from state to state represent transitions.

2.5 Model parameters

The emission and transition probabilities are parameters derived by posterior estimation from training data consisting of pre-determined spectrum/peptide assignments taken from one of the data sets, as described under 'Section 3.1'. Masses and relative intensities were divided into ten discrete bins for each of the 11 ion types, and the frequency of emission for ions within each mass and intensity bin were calculated for each ion type. For parameter estimation from the training data, we extracted the following factors: (1) the probability of emitting each ion type within a spectrum; (2) the probability of emitting an ion of the given type occurring within the associated mass bin; (3) the probability of emitting an ion of the given type occurring within the associated intensity and (4) the sequence dependent probability for each observed peak, i.e. the probability of peptide bond cleavage producing the peak, given the amino acids in the peptide sequence adjacent to the site. We also extracted the probability of co-occurrence of ion-types to determine the transition probabilities between states. Further details of parameters estimation can be found in Khatun *et al.* (Khatun *et al.*, 2007). The training set used for parameter estimation was either derived from 10-fold cross validation (i.e. 9/10-th used for training, 1/10-th for testing, repeated 10 times for each unique partition), or from a separate

data set (i.e. training on data set 1 and testing on data sets 2 and 3), as further described in the Results section.

2.6 Scoring function

Given a peptide sequence P, comprised of amino acids a_j ($j=1, L$), and an MS/MS spectrum S comprising N peaks, HMM_Score scores them jointly proceeding from the begin state to a first state π_1 corresponding to the first spectral peak, then the state π_2 for the second spectral peak, and so on to produce a path $\pi=\{\pi_1, \pi_2, \ldots, \pi_N\}$. Each of the HMM states π_i corresponds to one of the 12 ion types considered, q_j ($j=1, \ldots, 12$), and is determined at each step by the ion type(s) any matching mass(es) found in M_p. The joint probability of observing a series of masses M_S, their associated intensities I_S, a peptide P, and a series of states π based on matches between M_S and M_P, is given by:

$$p(M_s, I_s, P, \pi) = T_{b\to\pi_1} e_{\pi_1} \prod_{i=1}^{N-1} T_{\pi_i\to\pi_{i+1}} e_{\pi_{i+1}}. \quad (1)$$

$T_{b\to\pi_1}$ is the transition probability from 'begin' states to the first observed state (peak), which is equal to the probability of observing each ion type as the first one in the spectrum. $T_{\pi_{i\to\pi_{i+1}}}$ is the transition probability between states representing the ion types, and e_{π_i} is the probability of emission from the state π_i and is given by the product of two terms as,

$$e_{\pi_i} = e\left(\pi_i, M_S^i, I_S^i\right) e\left(M_S^i | a_{j-1}, a_j\right). \quad (2)$$

The first term considers the probability of emitting a peak of the given mass and intensity corresponding to the current state's ion type, and can be written as:

$$e\left(\pi_i, M_S^i, I_S^i\right) = e^M\left(M_S^i|\pi_i\right) e^I\left(M_S^i|\pi_i\right), \quad (3)$$

where the term $e^M(M_S^i|\pi_i)$ computes the probability of observing the mass corresponding to the present ion type (state), while the term $e^I(M_S^i|\pi_i)$ computes the probability of observing the intensity value corresponding to the current ion type.

The second emission term, $e(M_S^i|a_{j-1},a_j)$, calculates the probability of observing the present peak, given the particular amino acids pair adjacent to the cleavage site in the peptide sequence P that generated the predicted peak M_P^i that is being matched to the spectrum peak M_S^i. This is given by

$$e\left(M_S^i | a_{j-1}, a_j\right) = p_C\left(M_S^i | a_{j-1}\right) p_N\left(M_S^i | a_j\right). \quad (4)$$

The term $p_C(M_S^i|a_{j-1})$ is the conditional probability of observing the mass due to the C terminal fragmentation after the given amino acid a_{j-1} and $p_N(M_S^i|a_j)$ is the conditional probability of observing the mass due to the N terminal fragmentation before given amino acid a_j. In the case of internal ions where two cleavage events must occur, this value was averaged from both pairs of amino acids surrounding the two sites. This accounts for the effect that certain amino acids have on fragmentation of the backbone bonds in their vicinity. For example, the presence of the amino acid proline promotes frequent breakage of the peptide bond on its N terminal side while suppressing it on the C terminal side.

Multiple paths may be possible due to bifurcation at points where there are more than one theoretical ion produced by the peptide sequence P that match observed peaks in S. In such cases, the Viterbi dynamic programming algorithm (Durbin *et al.*, 1998) is used to determine the path with maximal probability π^* given by:

$$\pi^* = \arg\max_\pi p(M_S, I_S, P, \pi). \quad (5)$$

The described calculations are performed using the logarithm (Log) of the probability values to avoid numerical underflow, so Equation 5 produces $\mathrm{Log}(p(M, I, P, \pi^*))$. The result is a score with a larger (absolute) numeric value corresponding to lower probability values. It is counterintuitive to the potential user of the program to have a larger numeric score value correspond to worse matches, particularly since nearly all other MS/MS scoring algorithms produce larger positive numerical values for better matches. To avoid this confusion, we convert the probability to a score value that produces larger positive numerical values for the better matches. Dividing the probability by 1000 then negating and exponentiating gives:

$$\mathrm{Score}(P, S) = \exp\left(\frac{-Ln(p(M, I.P, \pi^*))}{1000}\right). \quad (6)$$

While the resulting score is no longer a probability, this transformation preserves the direct relationship of the path probability with the goodness of fit for the peptide match.

Initial experiments with this formulation revealed an issue that its results were skewed towards longer peptides by the occurrence of many false positive matches to the theoretical internal ions predicted for the peptide sequence—they are usually far more numerous than predicted b or y ions. However, the opposite is true in real spectra: an analysis of data set 1 showed that 70% of the predicted y ions were observed, 50% of predicted b ions observed, and only 10% of predicted internal ions observed. To correct this, we introduced an additional factor that weighs the actual and predicted occurrence frequency for each ion type, as given by:

$$w(P, S) = \sum_{q_{j=1,12}} E_f^{qj} * O_f^{qj}. \quad (7)$$

The first term E_f^{qj} is the expected frequency of each of the 11 ion types, enumerated by the index q_j, and the second term O_f^{qj} is the observed frequency for each of the corresponding ion types within the spectrum S. We used the observed frequency instead of the number of matching ions for each ion type, because the number of matching ions varies with the length of the peptide. This gives the highest score correction when both the predicted and observed ion frequency is high, and lowest when prediction and observation are both low. Another way of considering it is that for those ion types that *should* be observed more frequently, there is more effect on the score when they actually are observed.

The combination of all scoring factors thus produce a final score for each peptide/spectrum pair given by:

$$F\mathrm{Score}(P, S) = w(P, S) * \mathrm{Score}(P, S). \quad (8)$$

Equation 8 provides relative score value representing the goodness of fit between each peptide P in the database and the spectrum S, which increases in a positive fashion for better-fit between peptide sequence and spectral features. The score value is relative to the spectrum and peptide database searched, and is therefore not meaningful for comparisons of search results for different spectra or databases. To compare search results from run to run, the score distribution must be considered, i.e. the difference in score between the best match and the second-best, third-best, etc. This is not unique to HMM_Score, but is a common feature of many types of database search programs. To provide a fixed, probabilistic reference for the score values that allows their comparison from one result to another, we convert them to E-values, as described in the next section.

2.7 Expectation value calculation

Expectation values, commonly used by many database search methods [e.g. BLAST, (Altschul *et al.*, 1990)], convert a score expressed on an arbitrary value range into a measure of the relative uniqueness of a given match score. As described by Fenyö and Beavis for application to MS/MS search (Fenyo and Beavis, 2003), this provides a value that can be directly compared between different searches or even different algorithms.

We converted the score values produced by Equation 8 into E-values using the procedure given in Supplementary Section 2.7.1.

2.8 Performance evaluation

We applied several common performance evaluation measures to HMM_Score, including sensitivity and specificity measurements, along with ROC curve analysis (Hanley and McNeil, 1982). We also performed PR curve analysis (Davis and Goadrich, 2006) and box-plot analysis of the E-value distributions. The detailed procedures for calculating these measures are given in Supplementary Section 2.7.2. All results reported below were obtained using the E-values produced from the score in Equation 8.

We also compared HMM_Score performance to X!Tandem and Mascot. We downloaded X!tandem (version 2007.07.01.2) and ran Mascot standalone version (version 2.0.05) with the kind permission of the UNC-Duke Michael Hooker Proteomics Core facility. Searches were performed for all programs using the same version of the ***Swiss-Prot*** database. The parameters used for all programs were: peptides within 2 Da from the precursor ion were scored (no modification search), there was a 0.5 Da mass tolerance for product ion matching, and monoisotopic masses were used.

We assessed the sensitivity of each algorithm for its top scoring match results for each spectrum, and also calculated the sensitivity of each algorithm with the false positive rate limited to 1%. The number of correct identifications at a 1% false positive was determined by varying E-value threshold for X!Tandem and Mascot's reported—$\mathrm{Log}_{10}(P)$ score, until the target rate of false positives achieved, then counting the number of correct matches above that threshold. The same procedures were used for HMM_Score as described in the Section 3.1.

3 RESULTS

To give a comprehensive view of HMM_Score's performance, we tested it using three different data sets obtained on two different instruments and from entirely different biological samples (*E.coli* bacterial proteins and human blood plasma proteins). Our testing was focused upon spectra for which the matching peptide was previously known and validated, to provide a firm basis for comparison.

3.1 Sensitivity

Sensitivity is a general measure of the ability for an algorithm to search a large database and find a correct match for the target. In the case of MS/MS search, sensitivity implies the ability to identify the correct peptide P from a protein sequence database that matches the spectrum S. Sensitivity results for HMM_Score are shown in Table 1. For data set 1, comprised of 579 MS/MS spectra from *E.coli* proteins, HMM_Score was trained and tested by 10-fold cross validation for sensitivity assessment. HMM_Score identified 546 correct peptides in the first place when searched against the complete ***Swiss-Prot*** database, and 576 correct peptides in the first place for a search against the smaller *E.coli* database, representing a 0.6% error rate.

We performed similar testing for data set 2 and 3, though in this case the HMM_Score model was trained (i.e. parameters estimated) from data set 1, and then applied directly to score all of the spectra for both data sets 2 and 3. We did this due to the discovery that training with the high-quality data from set 1 resulted in better overall performance, despite the fact that it was generated on a different instrument than data set 2 used in testing. When data set 2 was searched against the complete ***Swiss-Prot*** version 54.1, HMM_Score (trained on data set 1) identified 469 correct peptides in the first place, and for data set 3 identified 1352 correct peptides in the first place (Table 1).

Table 1. The number of correct peptide hits produced by HMM_Score, searching against the database named in the second column

Dataset	Database	Total spectra	Correct identifications	
			a	b
Dataset 1	***Swiss-prot***	579	546	513
	E.Coli protein	579	576	576
Dataset 2	***Swiss-Prot***	677	469	416
Dataset 3a	***Swiss-Prot***	1986	1352	1284

The 'Correct Identifications' column (a) gives the number of correct peptide hits that were ranked first place when searched against the databases. The 'Correct Identifications' column (b) gives the number of correct peptides found at a 1% false positive rate.

We also examined how sensitive HMM_Score is when it produces false positives (FP) at a rate of 1%. The sensitivity at a specified false positive rate was calculated using forward sequence search, with the E-value threshold varied until the target error rate achieved. As shown in Table 1, column b, for data set 1 (*E.coli*, MALDI TOF/TOF), this resulted in 513 correct positive identifications when searching all of ***Swiss-Prot***, and 576 when searching only the *E.coli* protein database. For this set, performance was not substantially affected by the threshold limitation, since the false-positive rate was already very low. The same procedure applied to data set 2 produced 416 correct positive identifications, and data set 3 produced 1284 correct positive identifications.

In many projects, it is desirable to know the predicted false-positive error rate for a given identification score value when matching spectrum to a database. The procedure to calculate predicted false positive error rate is given in Supplementary Section 2.7.3. and the predicted E-value threshold are given in Supplementary Table 1.

3.2 Discriminative capability

We determined the discrimination ability of HMM_Score for real hits versus false-positive hits for each of the 3 data sets, as shown in Figure 2. We plotted the distribution of scores comparing spectral searches against the ***Swiss-Prot*** database (green bars) versus a reversed decoy database where all results are false-positives (red bars). For all three data sets there is a clear distinction in the range of expectation values between real (green) and false (red) hits. For data sets 1 and 3, the separation is almost complete. There is some overlap for data set 2 between the bottom 10% of scores for the real database search with the top-range of the decoy database search. It is notable that the false positive results were similar for all data sets with E-values between $\sim$0.1 and $\sim$5, whereas the true-positive results had a much greater difference in their distributions depending on the instrument and experimental conditions used.

3.3 Precision-recall and receiver operating characteristic analysis

Precision-recall analysis is used to assess the fidelity of database search across a range of different score cutoff thresholds to examine the tradeoffs between loose and strict scoring

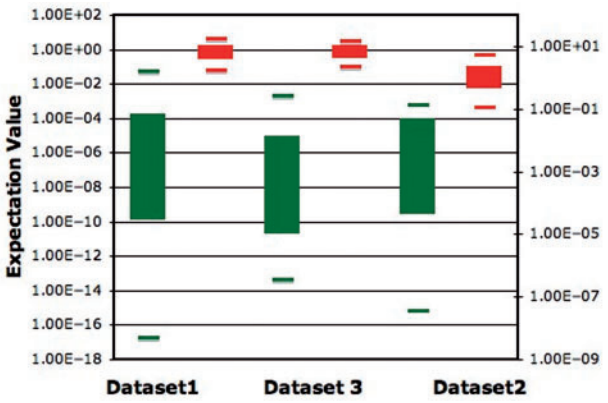

Fig. 2. Distribution of E-values obtained for MS/MS searches against forward (green) and reverse (red) sequence databases for all data sets. The solid boxes represent the ranges between lower 25% and higher 75% quartiles, and the horizontal lines represent the 5 and 95% ranges, respectively. The left y-axis range corresponds to data sets 1 and 3, while the right y-axis corresponds to data set 2. The difference in ranges is due to the differing instrumentation and experimental conditions used to produce the data for set 2 compared to sets 1 and 3.

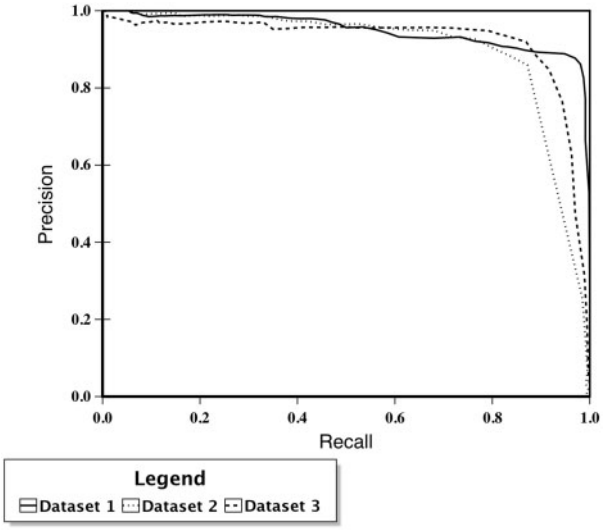

Fig. 3. The Precision-Recall curves for all data sets. To generate PR curves we considered false positive matches using the forward sequence of ***Swiss-Prot*** database. The solid circle curve is for data set 1 and the area this curve is 0.95. The triangle and asterisks are for data sets 2 and 3, respectively and areas under these curves are 0.90 and 0.93, respectively.

criteria. As the score threshold that distinguishes a 'positive' from 'negative' match is varied, precision and recall are calculated (Equations 13 and 14, Supplementary_Section 2). The PR curves calculated by varying the E-value threshold for positive identification by HMM_Score against ***Swiss-Prot*** are shown for each of the data sets in Figure 3, using the same training/testing procedures as for Section 3.2. The area under the PR curve, which gives an overall assessment of performance across a range of thresholds, was 0.95 for data set 1, 0.90 for data set 2, and 0.93 for data set 3. HMM_Score performed very well on all data sets, retaining high precision and recall throughout most of the range, with precision dropping substantially only at recall rates >0.8.

Receiver operating characteristic analysis is another method used to assess search fidelity by plotting specificity versus sensitivity as the score threshold is varied. However, for large databases its calculation is made problematic because it relies on assessment of 'true negatives' which are quite numerous for searching a large database. For comparison with the results reported by Kapp *et al*. we performed ROC analysis using the E-value scores for searches against the forward sequence database, with sensitivity and specificity calculated as discussed in the Supplementary Section 2. The resulting curves are shown in Supplementary Figure 2 for all data sets. The areas under the curve are 0.99, 0.95 and 0.98 for data set 1, 2 and 3, respectively for searches against the complete ***Swiss-Prot*** database.

3.4 Comparison of HMM_Score with Mascot and X!Tandem

Mascot is one of the most widely used database search algorithms (Perkins *et al*., 1999), and X!Tandem is an increasingly popular open source MS/MS search engine (Craig *et al*., 2004). To provide a practical basis for the comparison of HMM_Score performance, we applied the same sensitivity assessments reported in Section 3.1 for HMM_Score, to Mascot and X!Tandem for each of the three data sets. The results are shown in Table 2 (HMM_Score results are shown in Table 1). For each data set, HMM_Score outperformed X!Tandem and Mascot, particularly when the program results are limited to a 1% false positive rate. For data set 2, comprised of 677 validated human blood plasma MS/MS spectra and matching peptides (Kapp *et al*.), X!Tandem identified 289 correct peptides and Mascot identified 241 correct peptides at the 1% false positive rate in a search against ***Swiss-Prot***.

For comparison, HMM_Score identified 416 peptides at the 1% false positive rate for data set 2, which represents a 44% increase over X!Tandem and 73% increase over Mascot. The comparison using data set 2 may be the most representative

Table 2. The number of correct peptide hits produced by Mascot and X!Tandem when searching against ***Swiss-Prot*** protein database

Dataset	Total spectra	Algorithm	Correct identifications	
			a	b
1	579	X!Tandem	526	460
		Mascot	543	444
2	677	X!Tandem	419	289
		Mascot	372	241
3a	1986	X!Tandem	1297	1113
		Mascot	1333	1148

The 'Correct identifications' column (a) gives the number of correct peptide hits that were ranked with the top score when searched against the databases. The 'Correct identification' column (b) gives the number of correct peptides found when the false positive rate was limited to 1%.

of real-world performance, because it was obtained by Kapp *et al*. using the combined results of 7 different search algorithms, then hand verified. Data sets 1 and 3 are biased in favor of Mascot because they were obtained using an initial search by Mascot, followed by manual validation. In other words, for data sets 1 and 3a, when Mascot produced a false negative search result during their initial validation, that peptide was never to be seen by the other programs in their subsequent assessment. This artificially reduced both HMM_Score and X!Tandem's apparent performance on those data sets. However, even with this caveat, HMM_score performed better than Mascot in each case when searching a large database like ***Swiss-Prot***.

The number of correct peptides identified here by Mascot and X!Tandem differ from the results reported by Kapp *et al*. because they used a different (and smaller) database (International Protein Index (IPI), Human database, version 2.21). While our limited access to Mascot prevented us from testing it with IPI v2.21, we did assess X!Tandem and HMM_Score with this database. The version of X!Tandem we used identified 342 correct peptides at a 1% false positive rate, whereas Kapp *et al*. reported 311 peptides at the 1% false positive rate for the program. The discrepancy may be attributed to slightly different parameters used, such as a 3 Da precursor tolerance for Kapp's experiments versus 2 Da for ours.

For the same IPI v2.21 database, HMM_Score produced 425 correct hits at a 1% false positive rate. It is notable that the performance of HMM_score dropped only slightly to 416 correct peptides identified with the much larger ***Swiss-Prot*** database. Representing only a 2% drop in identifications at a 1% false positive rate, with a 441% larger database, this is perhaps the most significant result of our tests. It indicates that HMM_score specificity degrades very little with increasing database size, which is particularly important when searching putative peptide digests of whole eukaryotic genomes that may represent 10^9 peptides or more.

To provide an assessment of HMM_Score performance that was entirely independent of the results using another search algorithm, we collected spectra from known purified proteins, producing data set 3b. For this data set Mascot was run from their website (http://www.matrixscience.com/cgi/search_form.pl?FORMVER=2&SEARCH=MIS). The sensitivity of the three programs are shown in Supplementary Table 2, again with HMM_Score correctly identifying 12% more peptides than Mascot and 19% more peptides than X!Tandem. This result gives an assessment of performance when all spectra (including those of poor quality) are considered.

4 DISCUSSION

Assessing algorithm performance in a domain such as MS/MS spectrum identification is a significant challenge, because it requires the availability of standard data sets where the answer is already known to determine how well the program is working. While there are large numbers of MS/MS spectra now available in public data sets, the only way of knowing whether search results for an algorithm are correct is by manual validation, which is infeasible when tens of thousands of spectra are involved. We therefore focused in this work on validating HMM_Score using three data sets where the answer was already known with some reasonable confidence. Even in the case of those sets (the best presently available to our knowledge), their validation depended on other programs. This introduces a conundrum: when programs disagree, which one is correct? For this reason, Kapp *et al*. started with 6000 human blood plasma derived MS/MS spectra, and analyzed them using 7 different programs, then applying a voting procedure to determine which ones could be confidently identified by more than one program. This resulted in a set of 677 spectra for which the matching peptide sequence assignment had good confidence due to agreement of multiple programs and subsequent manual validation. We used this as data set 2. It is likely there are other peptides we could identify using HMM_Score in the remainder of the unidentified spectra from that data set (5129 were unidentified), but there would be no firm way to assure that our results were correct for those.

Because of these issues, data set 3b is likely the most representative of true performance. Its validation had no prior dependence upon other algorithms—it was based on a set of known protein standards, and it was small enough that we were able to confirm each result by hand. The second best is Kapp's data set (data set 2), because the validation was by multiple programs. For both of these data sets, HMM_Score's performance was very good. While no algorithm can presently identify every spectrum produced by a typical experiment, HMM_Score makes a significant improvement in identification capability.

While initially we performed intra-set 10-fold cross-validation to assess performance, it was suggested by a colleague that we attempt cross-instrument validation. We tested this suggestion, and realized that when HMM_Score was trained on the MALDI TOF/TOF data (data set 1), it produced better search results when applied during testing to the other data sets, including the ion trap data used in data set 2. While this was a surprise, we determined it was because data set 1 was a cleaner data set, which was more informative for parameter estimation (training). A fortunate side effect is that the validation described for data sets 2 and 3 above was performed with a model trained on data set 1, there is minimal concern about cross validation issues such as how the data set is partitioned. The training set was on a different instrument and from a different organism than the testing set for all reported experiments on data sets 2 and 3.

HMM_Score has computational complexity that scales linearly with the size of the spectrum and the number of peptides in the database searched. The program presently takes about 1–4 ms to calculate the most probable path using Viterbi for a mass list of 100 fragment ions. Thus, for a database like ***Swiss-Prot*** where the average number of peptides scored for each precursor within 2 Da is 25 000, HMM_Score takes about a minute per spectrum. We have further optimized this in our GFS software (Giddings *et al*., 2003) by pre-filtering using short sequence tags (Mann and Wilm, 1994) and then only applying HMM_Score to the pre-filtered spectra.

5 CONCLUSION

We have developed a scoring function using a HMM as its core, which uniquely employs information accessible from both the

MS/MS spectrum and the peptide sequence to provide excellent peptide identification performance. The use of machine learning, whereby additional fragmentation information was provided by the peptide sequence, resulted in a very robust algorithm. HMM_Score demonstrates the power of using machine learning methods to take advantage of the statistics present in large data sets. Further improvement may be possible by incorporating additional features of spectral data, such as the correlation between all amino acids present in a peptide and its likelihood of observation by MS/MS.

Preliminary application of the method to whole human genome search using Human Plasma Proteome data (results not shown) indicates that the improved accuracy of protein identification will aid with the accurate matching of MS/MS spectra directly to the whole human genome, in order to facilitate the identification of proteins that aren't represented by available annotation and/or to facilitate proteome-based re-annotation.

We have incorporated HMM_Score in our GFS software version 2.1, presently an in-house beta version scheduled for public release in late December 2007 (downloadable from http://gfs.unc.edu). To broaden the application of the method, we are also incorporating HMM_Score as a pluggable scoring function into X!Tandem (Craig *et al*., 2004). In addition, a standalone JAVA version of the program is available for download from Proteome Commons, which is linked from our website at http://bioinfo.med.unc.edu/Downloads.

ACKNOWLEDGEMENTS

We thank the UNC/Duke Michael Hooker Mass Spectrometry Core Facility the MALDI TOF/TOF spectra. We are also thankful to Dr. Ruth Silversmith for providing purified CheZ protein, and Mark Holmes, Jameson Miller and Stuart Jefferys for critically reading the manuscript. We are grateful for the work of Kapp *et al*, and making their testing data available, and we thank Dr. Martin McIntosh for the suggestion for cross-instrument validation. The work was supported by NIH R01HG003700 and NIH R01RR020823 to MCG.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.

Bafna,V. and Edwards,N. (2001) SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, **17** (Suppl. 1), S13–S21.

Bandeira,N. *et al*. (2004) Shotgun protein sequencing by tandem mass spectra assembly. *Anal. Chem*., **76**, 7221–7233.

Blattner,F.R. *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**, 1453–1474.

Craig,R. *et al*. (2004) Open source system for analyzing validating, and storing protein identification data. *J. Proteome Res*., **3**, 1234–1242.

Dancik,V. *et al*. (1999) De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol*., **6**, 327–342.

Davis,J. and Goadrich,M. (2006) The relationship between precision-recall and ROC curves. In *Proceeding of 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania.

Durbin,R. *et al*. (1998) *Biological Sequence Analysis: Probabilistic Model of Proteins and Nucleic Acids*. Cambridge University Press, United Kingdom.

Eng,J.K. *et al*. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*., **5**, 976–989.

Falkner,J. and Andrews,P. (2005) Fast tandem mass spectra-based protein identification regardless of the number of spectra or potential modifications examined. *Bioinformatics*, **21**, 2177–2184.

Fenyo,D. and Beavis,R.C. (2003) A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem*., **75**, 768–774.

Fischer,B. *et al*. (2005) NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem*., **77**, 7265–7273.

Frank,A. and Pevzner,P. (2005) PepNovo: *de novo* peptide sequencing via probabilistic network modeling. *Anal. Chem*., **77**, 964–973.

Geer,L.Y. *et al*. (2004) Open mass spectrometry search algorithm. *J. Proteome Res*., **3**, 958–964.

Giddings,M.C. *et al*. (2003) Genome-based peptide fingerprint scanning, *Proc. Natl Acad. Sci. USA*, **100**, 20–25.

Hanley,J.A. and McNeil,B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Horn,D.M. *et al*. (2000) Automated *de novo* sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl Acad. Sci. USA*, **97**, 10313–10317.

Kapp,E.A. *et al*. (2005) An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. *Proteomics*, **5**, 3475–3490.

Khatun,J. *et al*. (2007) Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Anal. Chem*., **79**, 3032–3040.

Kuster,B. *et al*. (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, **1**, 641–650.

LeDuc,R.D. *et al*. (2004) ProSight PTM: an integrated environment for protein identification and characterization by top-down mass spectrometry. *Nucleic Acids Res*., **32**, W340–W345.

Ma,B. *et al*. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom*., **17**, 2337–2342.

Mann,M. and Wilm,M. (1994) ErroraTolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem*., **66**, 4390–4399.

Narasimhan,C. *et al*. (2005) MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Anal. Chem*., **77**, 7581–7593.

Perkins,D.N. *et al*. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Rabiner,L.R. (1989) A tutorial on hidden markov-models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**, 257–286.

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

Tabb,D.L. *et al*. (2004) Influence of basic residue content on fragment ion peak intensities in low-energy collision-induced dissociation spectra of peptides. *Anal. Chem*., **76**, 1243–1248.

Tabb,D.L. *et al*. (2003) Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem*., **75**, 1155–1163.

Taylor,J.A. and Johnson,R.S. (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal. Chem*., **73**, 2594–2604.

Wan,Y.H. *et al*. (2006) PepHMM: a hidden Markov model based scoring function for mass spectrometry database search. *Anal. Chem*., **78**, 432–437.

Washburn,M.P. *et al*. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology, *Nature Biotechnology*, **19**, 242–247.

Wysocki,V.H. *et al*. (2005) Mass spectrometry of peptides and proteins. *Methods*, **35**, 211–222.

Yague,J. *et al*. (2003) Peptide rearrangement during quadrupole ion trap fragmentation: added complexity to MS/MS spectra. *Anal. Chem*., **75**, 1524–1535.

Zhang,N. *et al*. (2002) ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, **2**, 1406–1412.