Zhao Song[1]
Luonan Chen[1, 2, 6]
Ashwin Ganapathy[1]
Xiu-Feng Wan[1*]
Laurent Brechenmacher[3]
Nengbing Tao[4]
David Emerich[5]
Gary Stacey[3]
Dong Xu[1]

[1]Computer Science Department and Christopher S. Bond Life Sciences Center, University of Missouri-Columbia, Columbia, MO, USA
[2]Department of Electrical Engineering and Electronics, Osaka Sangyo University, Osaka, Japan
[3]Divisions of Plant Sciences and Biochemistry, Department of Microbiology and Molecular Immunology, Christopher S. Bond Life Sciences Center, University of Missouri-Columbia, Columbia, MO, USA
[4]Monsanto, U2G, Animal Genomics and Breeding, Creve Coeur, MO, USA
[5]Department of Biochemistry, University of Missouri-Columbia, Columbia, MO, USA
[6]Institute of Systems Biology, Shanghai University, Shanghai, P.R. China

Received May 17, 2006
Revised August 25, 2006
Accepted September 25, 2006

Research Article

Development and assessment of scoring functions for protein identification using PMF data

PMF is one of the major methods for protein identification using the MS technology. It is faster and cheaper than MS/MS. Although PMF does not differentiate trypsin-digested peptides of identical mass, which makes it less informative than MS/MS, current computational methods for PMF have the potential to improve its detection accuracy by better use of the information content in PMF spectra. We developed a number of new probability-based scoring functions for PMF protein identification based on the MOWSE algorithm. We considered a detailed distribution of matching masses in a protein database and peak intensity, as well as the likelihood of peptide matches to be close to each other in a protein sequence. Our computational methods are assessed and compared with other methods using PMF data of 52 gel spots of known protein standards. The comparison shows that our new scoring schemes have higher or comparable accuracies for protein identification in comparison to the existing methods. Our software is freely available upon request. The scoring functions can be easily incorporated into other proteomics software packages.

Keywords:
MOWSE / PMF protein identification / Scoring function / Statistical distribution
DOI 10.1002/elps.200600305

1 Introduction

The general approach for MS protein identification is by matching the features derived from the mass spectra of a protein sample against a protein sequence database that contains the sequences of the proteins in the sample [1]. It involves protein digestion using an enzyme (for example, trypsin, glu-C, *etc.*) and chromatographic separation, followed by PMF [2] or MS/MS analysis [3]. PMF protein identification compares the masses of peptides derived from the experimental spectral peaks with each of the possible peptides generated by computationally digesting proteins in the sequence database. The MS/MS method further breaks each digested peptide into smaller fragments, whose spectra provide effective signatures of individual amino acids in the peptide for protein identification. While the MS/MS method is more accurate in defining peptides, it is much more expensive and time-consuming than PMF. PMF provides an economic method for protein identification, and it can serve as an effective filter for selecting some proteins on which to conduct MS/MS analysis. Currently, PMF using MALDI-TOF MS is still widely used.

The basic idea for protein identification from PMF spectral peaks is illustrated in Fig. 1. At the experimental side, a protein is digested into peptide fragments using enzymes that recognize specific sites. The most common enzyme is trypsin, which cleaves very specifically at R−X and K−*X* bonds except when X = P (the rule on X = P is sometimes not used as a hard constraint). A PMF experiment provides the spectra of the mass-to-charge ratios (m/z) for the digested

Correspondence: Professor Dong Xu, Computer Science Department and Christopher S. Bond Life Sciences Center, 1201 East Rollins Road, University of Missouri-Columbia, Columbia, MO 65211-2060, USA
E-mail: xudong@missouri.edu
Fax: +1-573-884-9676

Abbreviations: MPBSF, modified probability-based scoring function; **NDSF,** normal distribution-based scoring function; **NMOWSE,** neighbor-MOWSE scoring function; **PBSF,** probability-based scoring function

* Current address: Department of Microbiology, Miami University, Oxford, OH 45056, USA

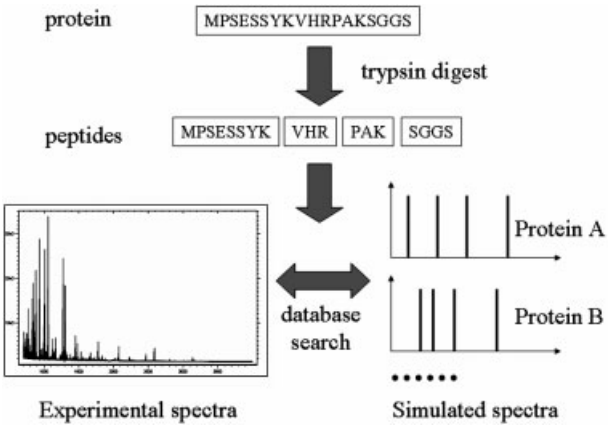


Figure 1. PMF protein identification. The protein in the gel sample is digested into peptides, whose mass-to-charge ratios are shown in the PMF spectra. The PMF spectra are used to compare with simulated spectra of each protein in a database for protein identification.

peptides. The peak intensity relates to the abundance of the peptide, but the relationship is complicated [4]. At the computational side, a database is prepared for all possible protein sequences derived from the genomic sequence of the organism in the gel sample or all the sequences collected in a comprehensive protein database such as UniProtKB/Swiss-Prot [5]. Each protein in the database is then computationally digested into peptide sequences according to the type of digestion. The common PMF protein identification is carried out through two steps: (i) the experimental PMF spectral peaks are compared with simulated ones from each of the possible peptides generated by computationally digested proteins in the sequence database, and (ii) the proteins in the sequence database with many peptide matches are considered as the top candidates for the proteins in the experimental sample.

Several computational tools have been developed for PMF protein identification. MOWSE [6] was an earlier software package for PMF protein identification, and EMOWSE (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/emowse.html) is the latest implementation of the MOWSE algorithm. MS-Fit in the ProteinProspector package (http://prospector.ucsf.edu/) [7] uses a variant of the MOWSE scoring scheme. It incorporates several new features, including constraints on the minimum number of peptides to be matched for a possible hit, the number of missed cleavages, and the target protein's molecular weight range. MASCOT (Matrix Science Inc., http://www.matrixscience.com/) [8] is an extension of the MOWSE algorithm. It incorporates the same scoring scheme, but provides a probability-based score. ProFound (http://prowl.rockefeller.edu/) [9] uses the Bayesian probability theory and an Expert System for protein identification, with a generalized probability score. OLAV-PMF [10] applies a probabilistic model to estimate the ratio of two likelihoods between a list of experimental peptide masses and the corresponding list of expected ones. Probity [11] analytically calculates the risk of random matching between experimental masses and theoretical masses of a protein in a search database. ChemApplex considered peak intensity and the accuracy of the match between the experimental mass and the theoretical mass in the scoring function [12]. Ossipova *et al.* [13] developed a method to optimize the parameters for PMF protein identification in database search.

There are several limitations for the current computational methods of PMF protein identification, which may result in under-utilization of the available information content in PMF spectra for protein identification. A general issue is the scoring function, which assesses the match between an experimental spectrum and a simulated spectrum in a protein database. Current scoring functions use simple statistics, which typically do not consider peak intensity information, distribution model of matching m/z in a database, or the distribution of m/z matches along a protein sequence. When multiple proteins in the database can fit the PMF spectra, some of the existing scoring functions may not filter out false-positive results effectively. If these limitations can be addressed effectively, many gel spots can be identified confidently without using MS/MS experiments for further validation.

In this paper, we present a number of new scoring functions and compare them with MOWSE. The methods described in this paper have been implemented in a software package, which is available upon request.

2 Materials and methods

2.1 Database search

The database used for protein identification in this paper is sprot45 from UniProtKB/Swiss-Prot (last updated in January 2005), together with the 40 proteins from soybean (generated after January 2005) that we have identified but not included in the database. The database has 163 275 proteins in total, including eight fields for each entry: accession number, peptide number, peptide sequences, peptide masses, peptide lengths, protein sequence, protein name, and protein molecular weight. The molecular weight of a peptide of N residues is calculated as

$$\sum_{i=1}^{N} \text{residue\_mass}_i + 18.015 \quad (1)$$

Equation (1) takes into account an amino-terminal hydrogen and a carboxy-terminal hydroxyl group, which sum up to 18.015.

In this study, we only consider complete trypsin digestion of a protein and peptide without including any missed cleavage. In addition, we assume that the charge state of all the peptides is 1 and no post-translational modification exists in any peptide. We use only monoisotopic peaks.

2.2 Scoring schema

Here we first describe briefly the widely used MOWSE scoring function. Then we will illustrate the four novel scoring schemes (score schemes 2–5 in the following) that we developed.

2.2.1 MOWSE scoring function

MOWSE [6] is one of the earliest scoring schemes in protein identification using PMF data, which is still widely used. The scheme is based on the number of possible matches within a target protein and the frequency of occurrence of the molecular weight of each peptide. A frequency table, as indicated in Fig. 2, is constructed for all peptide entries in the database. Each column in the frequency table represents the MW of the protein and is divided into 10 kDa intervals. Rows represent the MW of peptides and are divided into 100 Da intervals. Proteins found in the database are entered into the table based on their molecular weights and the weights of peptides found in each protein. Each cell thus comprises the number of occurrences of peptides within a specific molecular weight range in a protein of certain intact molecular weight. The frequency table is constructed by normalizing the value in each cell with the largest number found in each column. Specifically, the frequency f_{ij} in cell-(i,j) is $f_{ij} = N_{ij}/N_{jmax}$, where $N_{jmax} = \max\{N_{1j}, N_{2j}, \ldots\}$ is the largest number in column j. For protein identification, each protein in the target database is scored by multiplying the frequency value of the matched peptide, whose molecular weight differs from the experimental spectral peak within a cutoff value (typically 1 Da). This product is scaled with the protein molecular weight and then inverted. The final score Score = 50 000/$(p_n \times w_p)$, where p_n is the product of matched distribution scores and w_p the "hit" protein molecular weight in the database [4]. $p_n \propto \prod_{i=R(l),\, l\in H} f_{ij}$, where $R(l)$ represents the row number of the table for the lth fragment of the mass spectra, and H is the set of the matched fragments of the mass spectra with the protein.

2.2.2 Normal distribution-based scoring function (NDSF)

To make use of the peak intensity and the quantitative difference between the experimental mass values of selected peaks and matched mass values in the protein database, we developed the following energy function based on Eq. (1) in ref. [9]:

$$\text{Score} = \sum_i \left[Int_i \times \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[-\frac{(mt_i - me_i)^2}{2\sigma_i^2} \right] \right] \qquad (2)$$

In Eq. (2), mt_i is the mass of theoretical peptide in the database, me_i is the mass of matched experimental peak, Int_i is the corresponding intensity value for the experimental peptide, and $\sigma_i = (1/3)\text{tolerance} \times me_i$(typically tolerance = 100 ppm = 0.01%). This form of the scoring scheme assumes that all mass matches between theoretical peptides and experimental peptide follow a normal distribution with the mass of experimental value as mean.

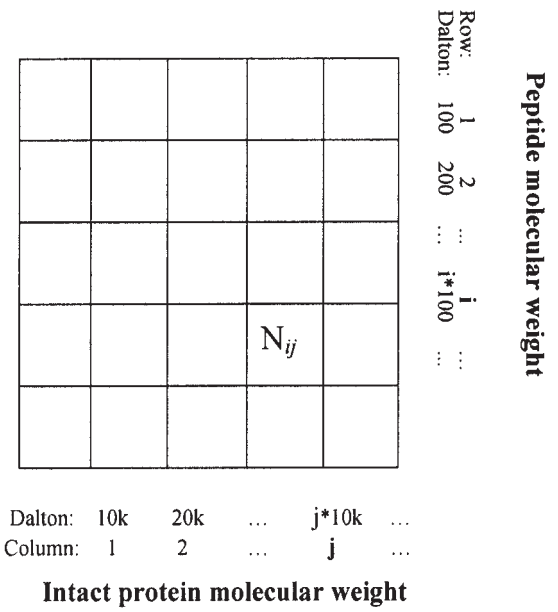


Figure 2. MOWSE occurrence table of peptides based on the database. Each column in the table represents the molecular weight of the protein and is divided into 10 kDa intervals. Rows represent the molecular weights of peptides and are divided into 100 Da intervals. N_{ij} in cell *i,j* is the total number of occurrence of digested peptides for all the proteins in the database with MW ranging from $(j-1)\times 10$ to $j\times 10$ kDa.

2.2.3 Neighbor-MOWSE scoring function (NMOWSE)

When different MS peaks match multiple peptides in a protein, the distribution of these peptides on the protein reflects the likelihood for the protein to represent the MS spectra. For example, if these peptides are contiguous on a protein sequence, it represents a better chance to be the true protein for the spectra than the case if they are distributed randomly on the sequence. We modified the MOWSE scoring scheme to reflect such a statistical relationship. The frequency table is constructed similarly as that of MOWSE. The difference is that the NMOWSE algorithm adds weight to neighboring matching peptides. The weight is obtained from the frequency table, as proportional to the sum of the frequencies of both matching neighbors, *i.e.*,

$$\text{factor}_{i,\,i+1} = \lambda \times (\text{factor}_i + \text{factor}_{i+1}) \qquad (3)$$

where λ is a parameter and factor_i is the value (*i.e.*, f_{kj} where $k = R(i)$) in the frequency table for the ith peptide of the experimental protein. Then the factors for all the contiguous

matches from PMF spectra to a protein sequence are multiplied, *i.e.*,

$$\text{Weight} = \prod_{i+1=j} \text{factor}_{i,j} \tag{4}$$

The "Weight" will be multiplied to the regular MOWSE score as the final score.

2.2.4 Probability-based scoring function (PBSF)

To handle the statistical properties in PMF protein identification more systematically, we developed a new scoring scheme based on the MOWSE occurrence table. In this case, based on Fig. 2, when comparing a mass distribution of peptides (n fragment molecular weights in the spectra) with the database entry of molecular weights (protein k in the column j), $R(l)$ represents the row number of the table for the lth fragment of the mass spectra. When the difference in two peptide weights is within a tolerance value, it is a hit or match. Otherwise it is nonmatching. The probability for a match between a mass distribution of peptides and a protein k in the database is computed *via*

$$Pr(P_k) = \prod_{i=R(l),l\in H_k} \left[1 - \left(1 - \frac{m_{ij}}{M_j}\right)^{n_{ij}^k}\right] \tag{5}$$

where $Pr(P_k)$ represents a probability or ratio for protein k matching with the fragment peptides of the experimental mass spectra. H_k is the set of the matched fragments of the mass spectra with protein k, and n_{ij}^k is the number of peptides in cell-(i,j) of protein k. Let M_j be the total number of proteins in the j column of Fig. 2 among the databases. m_{ij} represents the average number of occurrences of peptides in cell i,j for one protein of the database, *i.e.*, $m_{ij} = N_{ij}/M_j$, and M_j is the total number of occurrences of peptides in the jth column of the database, *i.e.*, $M_j = \sum_{i=1}^{n_r} m_{ij}$, where n_r is the total number of rows in the table. Clearly, m_{ij}/M_j is the frequency in the cell i,j for the column j. Note that such a frequency is different from f_{ij} of MOWSE.

In mass spectra, high-abundance peaks are more likely to be the peaks representing true peptides, whereas low-abundance peaks are more likely to be noisy. To account for the peak intensity effect we modify Eq. (5) as

$$\text{Pr}(P_k) = \prod_{i=R(l),l\in H_k} \left[\left[1 - \left(1 - \frac{m_{ij}}{M_{ij}}\right)^{n_{ij}^k}\right](1 - I_l)\right] \tag{6}$$

where I_l is the normalized intensity ([0,1]) of the lth spectrum, *i.e.*,

$$I_l = \frac{1}{1 + e^{-\alpha(\hat{I}_l - \bar{I})}} \tag{7}$$

In Eq. (7), $\hat{I}_l$ is the original intensity, $\bar{I}$ is the average intensity for all selected peaks, and α is a constant. To achieve a good prevision in computing, we adopt $-\log \text{Pr}(P_k)$ as the score function for protein identification.

2.2.5 Modified probability-based scoring function (MPBSF)

We further developed another scoring scheme by integrating the information for the neighboring matching peptides into an MPBSF. The score utilizes the average distance of matched peptides and is defined as

$$\text{ADMP} = \frac{\sum_{i=1}^{n_m-1} Dis_{i,i+1}}{n_s/n_p} \tag{8}$$

where the numerator represents the sum of the distances between two adjacent matching peptides, while the denominator represents the total number of possible digested segments in the protein divided by the number of matching peptides. Specifically, n_m, n_s, n_p are the number of the matching peptides in the spectra, the number of the digested segments, and the number of the matching peptides in the protein, respectively. The final MPBSF score is calculated as $-\log Pr(P_k) - \log(\text{ADMP})$.

2.3 Samples

In order to provide benchmark data for the computational studies, we used seven protein standards, which yielded 12 gel spots. In-gel trypsin digests were performed for the Coomassie-stained 2-D gel plugs. The digests were dried on a centrifugal evaporator, reconstituted, and desalted on C18 ZipTips. The desalted digests were analyzed by MALDI-TOF MS with CHCA in positive ion delayed extraction reflector mode. The sample spots were washed on target with diammonium citrate to reduce the interference from matrix ion clusters, and reanalyzed by MALDI-TOF MS in positive ion delayed extraction reflector mode. The 12 spots on 2-D gel and their corresponding proteins and isoforms are shown in Table 1.

We also used 40 proteins (600 μg) extracted from soybean (cv Williams 82) and root hair, which were separated by 2-DE (24 cm IPG strip, linear pH 4–7), according to the method

Table 1. Gel spots of protein standards

Gel spot	Protein name	Species	Swiss-Prot ID
Plug1	Glucoamylases	*Aspergillus niger*	P04064
Plug2	Trypsin (trypsinogen)	Bovine	P00760
Plug3	Lentil lectin	*Lens culinaris* (Lentil)	P02870
Plug4	Myoglobin	Horse	P68082
Plug5	β-Lactoglobulin	Bovine	P02754
Plug6	Carbonic anhydrase I	Human	P00915
Plug7	Trypsin inhibitor	Soybean	P01070
Plug8	Trypsin inhibitor	Soybean	P01070
Plug9	β-Lactoglobulin	Bovine	P02754
Plug10	β-Lactoglobulin	Bovine	P02754
Plug11	Myoglobin	Horse	P68082
Plug12	Trypsin (trypsinogen)	Bovine	P00760

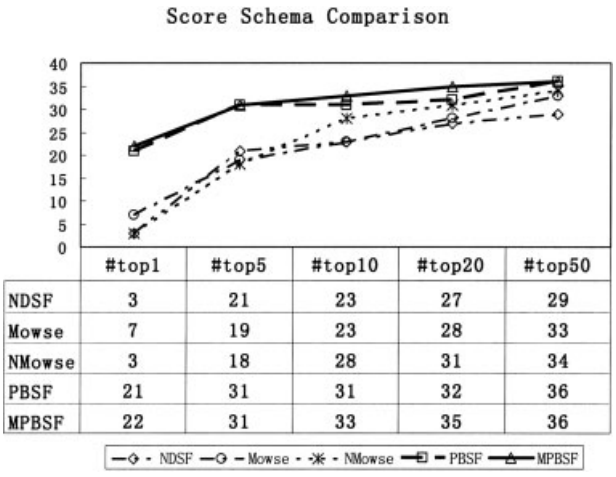
described in ref. [14]. Four replicates were performed and gel pictures were analyzed using Phoretix (nonlinear dynamics, v2005). Spots identified in at least three out of the four replicates were excised using a spot picker and their molecular weights and p*I*s were determined. The gel plugs were then digested using sequencing-grade modified trypsin (Promega, Madison, Wisconsin USA). Tryptic peptides were lyophilized, reconstituted in 10 μL of 700:290:10 by volume ACN/water/formic acid (HCOOH w/v in water $\geq$ 88%) and 0.5 μL of the solution was mixed with the same volume of α-cyano-4-hydroxycinnamic acid (Fluka MS-grade, Sigma–Aldrich, St. Louis, USA) solution (5 mg/mL in 500:380:20:100 ACN/water/10% TFA/100 mM ammonium dihydrogen phosphate). The sample/matrix (0.3 μL) mix was deposited on a stainless-steel plate (ABI01-192-6-AB). The tryptic peptides were analyzed on an Applied Biosystems Inc. 4700 MALDI TOF/TOF MS in positive ion delayed extraction reflector mode with a 355 nm (200 Hz) laser. The instrument was calibrated with ABI peptide standards (4700 Mass standards kit, 4333604). Spectra were analyzed using the GPS Explorer software (v. 3.0) (Applied Biosystems) and the Matrix Science's MASCOT search engine (www.matrixscience.com) against the NCBI Viridiplantae protein database. Search parameters included, a maximum of 150 ions *per* MS spectrum with an S/N>20, a mass error of 0.1 Da for the monoisotopic precursor ions, a maximum of one allowed miscleavage by trypsin, an exclusion of peptide masses corresponding to the autolysis of the trypsin, carbamidomethylation of cysteines and methionine oxidation, respectively as fixed and variable modifications.

We found 40 proteins that were identified confidently with the MS/MS mode, and we used their corresponding MS fingerprinting data as the inputs for our tests of scoring schemes. We assume that a test protein identification is correct if our search using the fingerprinting data matches the protein identified from the MS/MS data. The 40 spots on 2-D gel and their corresponding proteins and isoforms are shown in Supplementary Table 1.

3 Results

3.1 Score schema comparison

We compared the performance of our newly developed schemes with the MOWSE score function, using the same experimental datasets as described in Section 2.3 (12 standards together with 40 soybean proteins). For each protein identification, we manually selected a set of peaks from a spectrum provided by the Proteomics Center, University of Missouri-Columbia. Matched peptides should cover at least 25% of a protein sequence in order to be listed as a candidate of correct result. Figure 3 shows the comparison results of the five scoring functions (NDSF, MOWSE, NMOWSE, PBSF, and MPBSF) in terms of ranking correct proteins among top hits. It indicates that PBSF and MPBSF performed significantly better than the other three methods (especially in the "#top 1" category). The performance between PBSF and MPBSF is similar, while MPBSF slightly outperforms PBSF.



	#top1	#top5	#top10	#top20	#top50
NDSF	3	21	23	27	29
Mowse	7	19	23	28	33
NMowse	3	18	28	31	34
PBSF	21	31	31	32	36
MPBSF	22	31	33	35	36

Figure 3. Scoring function comparison, numbers in which the expected protein ranks top 1 to top 500 *versus* top ranking field.

To provide a more robust comparison, we randomly sampled selected peaks for each gel spot. For a fixed percentage of selected peaks, we randomly picked peaks ten times, and used each set of generated spectra for protein identification with the five scoring schemes. Figure 4 compares different methods, where the percentages for the expected proteins to rank top 1 or top 10 among the 52 sets of PMF spectra are plotted against the percentage of selected peaks in the generated spectra. The result is consistent with Fig. 3. The figure clearly indicates that PBSF and MPBSF are significantly better than the other three methods, while MPBSF is slightly better than PBSF. The overall performance of the algorithms can be sorted as MPBSF>PBSF>NMOWSE>MOWSE>NDSF. It is worth mentioning that NDSF performed worse than MOWSE, although NDSF considers the match more quantitatively using the difference between the theoretical mass in the database and the mass measured from the spectrum, as well as the peak intensity information. This may be due to the fact that the NDSF scheme is too sensitive to noise, and as a result does not perform well. Considering the hit distribution on a protein systematically improves the protein identification accuracy, as NMOWSE outperformed MOWSE and MPBSF outperformed PBSF. However, such an improvement is less than the one from the rigorous statistical model in PBSF.

We have developed a capacity to handle missed cleavages. We tested the missed cleavage feature using the dataset of 52 gel spots. We found that the result was similar to the one without the missed cleavage feature. We also tested our dataset on MASCOT, and the results between allowing and ignoring missed cleavage were also quite similar. This suggests that our dataset may contain few missed cleavages. Nevertheless, the capacity to handle missed cleavages, implemented in our package, is useful in general.

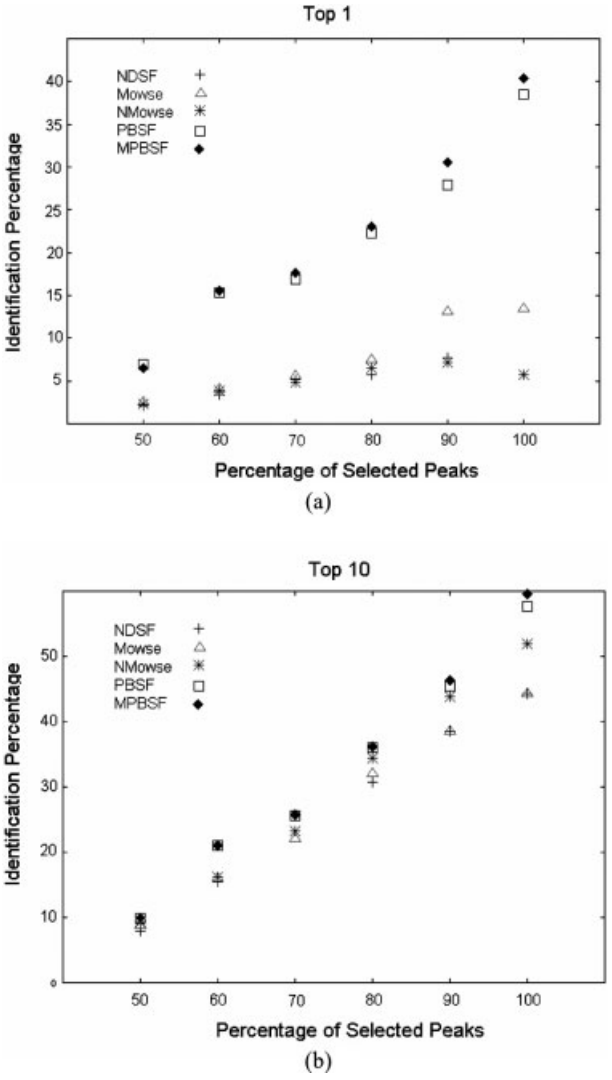


Figure 4. Percentages in which the expected protein ranks top 1 (a) and top 10 (b) *versus* percentage of selected peaks from PMF spectra.

3.2 Comparison with MASCOT and ProteinProspector

We compared our software with two widely used software packages, *i.e.*, MASCOT and ProteinProspector. The performance of our software is similar to or slightly worse than MASCOT. As the online version of MASCOT or ProteinProspector does not allow a user to add a search database, we limited our comparison on the 12 known spots, whose proteins can be found in a search database (UniProtKB/Swiss-Prot) of MASCOT or ProteinProspector. Among the 12 testing data points, our software ranked the correct protein the 1st for four cases, in top 15 for six cases; while MASCOT ranked the correct protein the 1st for five cases and in top 15 for seven cases. This is probably due to special treatment and features beyond the MOWSE score that MASCOT uses. For example, MASCOT uses a confidence assessment on top of the scoring function. Our software performs better than ProteinProspector. When we set a threshold for the molecular weight, the result of MPBSF did not change much, which indicated that the performance of MPBSF is basically invariant against the molecular weight of a protein. This is not the case in ProteinProspector. Without limitation of protein's molecular weight, ProteinProspector could find none ranked top in 50 for the 12 proteins; with the limitation setting to 1000–100 000 Da, ProteinProspector could find 1 ranked the 1st, 3 ranked the 2nd, and 1 ranked the 37th. The testing result is listed in Table 2.

4 Discussion

Although MS/MS provides more information for protein identification, PMF will still be useful for fast and inexpensive protein identification. One of the problems is that by applying PMF in protein identification, the multi-to-one relationship between gels and protein still exists. This is mainly a result of the information content for protein identification in PMF being much less than MS/MS. Nevertheless, the information content in PMF spectra may not be fully utilized by current computational methods for protein identification. The work described in this paper represents an effort to explore more effective scoring schemes by better using the information content in PMF spectra to improve protein identification accuracy. The scoring schemes developed (especially PBSF and MPBSF) are novel to the best of our knowledge. Our results using experimental PMF spectra demonstrate that the new scoring schemes can yield higher accuracy in protein identification. The new scoring schemes are generally applicable to any PMF protein identification software.

Table 2. Comparison for ranking of correct hit between our software and MASCOT/ProteinProspector

Spots/ranks	MPBSF	MASCOT	ProteinProspector without MW limitation	ProteinProspector with MW limitation
Spot1	1	>20	>50	2
Spot2	8	>20	>50	>20
Spot3	>20	>20	>50	>20
Spot4	1	1	>50	1
Spot5	>20	8	>50	>20
Spot6	1	1	>50	2
Spot7	>20	2	>50	>20
Spot8	>20	>20	>50	>20
Spot9	>20	1	>50	>20
Spot10	>20	1	>50	>20
Spot11	1	1	>50	2
Spot12	13	>20	>50	37

The better performance of our new scoring schemes is mainly due to more rigorous formulations and consideration of peptide hit distribution on a protein. Although MOWSE applied the propensity of molecular weights for proteins and peptides in protein identification, it does not have a comprehensive consideration for the underlying statistical distribution. In contrast, PBSF builds on the MOWSE table, but has a rigorous treatment for the statistical distribution in matching PMF spectra to the proteins in a search database. The test result using the experimental PMF spectra showed that major improvement in protein identification accuracy came from PBSF, or the rigorous statistical treatment, as PBSF alone outperformed MOWSE, NMOWSE, and NDSF by a large margin. To a less extent, the improvement came from the consideration of peptide hit distribution on a protein, as MPBSF slightly outperformed PBSF. The peptide hit distribution represents a useful, independent source of information for protein identification, and it has not been explored by any other PMF identification. By combining the PBSF statistical model and the peptide hit distribution information on a protein, MPBSF achieved the best performance among all the scoring schemes. To test the effect of adding peak intensity in scoring function, we removed the last item $(1 - I_l)$ in Eq. (6). We compared the computational result with MPBSF and found the result was not as good as the original PBSF but still consistently better than MOWSE, NMOWSE and NDSF (data not shown). This indicates that the peak intensity helps protein identification, but it is not the main source of accuracy improvement.

There are some limitations for the validations of our methods. In particular, the number of test cases is limited. Furthermore, although the scoring functions shown in this paper incorporated some factors for protein identification, such as missed cleavage and protein molecular weight range setting, other factors still need to be incorporated, in particular, post-translational modification to handle mass increments, neutral losses, or diagnostic fragment ions in peptide mass spectra [15, 16]. In addition, we will follow some developed strategies to handle the issue that most of the gel spots are mixtures of multiple proteins [17–19]. We are incorporating these factors into our software package and will test it using more PMF data. In addition, we are exploring more systematic handling of the statistics for the scoring schemes, in particular, using the Dirichlet distribution [20] for the treatment of the peptide hit distribution on a protein and combining this distribution with the PBSF statistical model.

This work has been supported by the MU-Monsanto Program. Work in the GS laboratory was also supported by a grant (DBI-0421620) from the National Science Foundation, Plant Genome Program. The authors like to acknowledge the Proteomics Center at the University of Missouri-Columbia for MS services. We would like to thank Beverly DaGue, Brian Mooney, Jay Thelen, Chi-Ren Shyu, Guohui Lin, Yu Chen, and Zhihai He for helpful discussions.

5 References

[1] Gevaert, K., Vandekerckhove, J., *Electrophoresis* 2000, *21*, 1145–1154.

[2] Cottrell, J. S., *Pept. Res.* 1994, *7*, 115–124.

[3] Yates III, J. R., McCormack, A. L., Link, A. J., Schieltz, D. *et al.*, *Analyst* 1996, *121*, 65R–76R.

[4] Gay, S., Binz, P. A., Hochstrasser, D. F., Appel, R. D., *Proteomics* 2002, *2*, 1374–1391.

[5] Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C. *et al.*, *Nucleic Acids Res.* 2003, *31*, 365–370.

[6] Pappin, D. J., Hojrup, P., Bleasby, A. J., *Curr. Biol.* 1993, *3*, 327–332.

[7] Clauser, K. R., Baker, P. R., Burlingame, A. L., *Anal. Chem.* 1999, *71*, 2871–2882.

[8] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[9] Zhang, W. Z., Chait, B. T., *Anal. Chem.* 2000, *72*, 2482–2489.

[10] Margnin, J., Masselot, A., Menzel, C., Colinge, J., *J. Proteome Res.* 2004, *3*, 55–60.

[11] Eriksson, J., Fenyo, D., *J. Proteome Res.* 2004, *3*, 32–36.

[12] Parker, K. C., *J. Am. Soc. Mass Spectrom.* 2002, *13*, 22–39.

[13] Ossipova, E., Fenyo, D., Eriksson, J., *Proteomics* 2006, *6*, 2079–2085.

[14] Wan, J., Torres, M., Ganapathy, A., Thelen, J. *et al.*, *Mol. Plant Microbe Interact.* 2005, *18*, 458–467.

[15] Liebler, D. C., *Introduction to Proteomics: Tools for the New Biology*, Humana Press, Totowa, NJ 2001.

[16] Matthiesen, R., Trelle, M. B, Højrup, P., Bunkenborg, J., Jensen, O. N., *J. Proteome Res.* 2005, *4*, 2338–2347.

[17] Jensen, O. N., Podtelejnikov, A. V., Mann, M., *Anal. Chem.* 1997, *69*, 4741–4750.

[18] Ossipova, E., Nord, LI, Kenne, L., Eriksson, J., *Rapid Commun. Mass Spectrom.* 2004, *18*, 2053–2058.

[19] Eriksson, J., Fenyo, D., *J. Proteome Res.* 2005, *4*, 387–393.

[20] Berger, J., *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York 1985.

