Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Design and Implementation of Probability-Based Scoring Function for Peptide Mass Fingerprinting Protein Identification

Zhao Song, Luonan Chen, Chao Zhang, Dong Xu

Abstract—**Protein identification through high-throughput mass spectrum data is an important domain in proteomics. Peptide Mass Fingerprinting (PMF) is one of the major methods for protein identification using the mass-spec technology. We developed a software package called "ProteinDecision" for PMF protein identification, together with a user-friendly graphical interface. "ProteinDecision" can handle the issues of selecting peaks from mass spectrum, transforming database format, displaying the top ranks of identification result, and detailed information for each ranking. We used a novel scoring function by considering the distribution of matching a mass-to-charge and peak intensity in a database based on the MOWSE table. Our new scoring function is assessed better than existing ones by comparing the computational results using experimental PMF data. A standalone version of "ProteinDecision" is freely available upon request.**

I. Introduction

The general approach for MS protein identification is through 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, pepsin, glu-C, etc.) and chromatographic separation, followed by peptide mass fingerprinting (PMF) [2] or tandem mass (MS/MS) spectrometry 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.

Several computational tools have been developed for PMF protein identification. MOWSE [4] 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/) [5] 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/) [6] is an extension of the MOWSE algorithm. It incorporates the same scoring scheme, but provides a probability-based score. ProFound (http://prowl.rockefeller.edu/) [7] uses the Bayesian probability theory and an Expert System for protein identification, with a generalized probability score.

There are several limitations in existing computer systems for PMF protein identification. First, although some research has been done [8], the current scoring functions generally under-utilize the available information content in PMF spectra for protein identification. Second, PMF protein identification need reliable free standalone software to handle large-scale PMF spectra, as current users typically either apply expensive commercial software packages or use free Web servers to handle one spectrum at a time. In this paper, we present a new probability-based scoring function and compare it with MOWSE. We will also introduce the software package "ProteinDecision", which is freely available upon request.

II. Methods and Results

Here we first describe briefly the widely used MOWSE scoring function. Then we will illustrate the novel scoring scheme that we developed.

A. MOWSE Scoring Function

MOWSE [4] 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 Figure 1, is constructed for every peptide entry in the database. Each column in the frequency table represents the molecular weight of the protein and is divided in 10 kDa intervals. Rows represent the molecular weight of peptides and are divided in 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 occurrence 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

Zhao Song, Computer Science Department and Christopher S. Bond Life Sciences Center, 1201 East Rollins Road, University of Missouri-Columbia, Columbia, MO 65211-2060, USA
Luonan Chen, Department of Electrical Engineering and Electronics, Osaka Sangyo University Nakagaito 3-1-1, Daito, Osaka 574-8530, Japan
Chao Zhang, Computer Science Department and Christopher S. Bond Life Sciences Center, 1201 East Rollins Road, University of Missouri-Columbia, Columbia, MO 65211-2060, USA
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

1-4244-0033-3/06/$20.00 ©2006 IEEE.

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 1.0 Da). This product is scaled with the protein molecular weight and then inverted. The final score Score = 50000 / (p_n * w_p), where p_n is the product of matched distribution scores and w_p the 'hit' protein molecular weight in the database [4].

$$p_n \propto \prod_{i=R(l),\, l\in H} f_{ij} \qquad \text{Equation 1}$$

where R(l) represents the row number of the table for the l-th fragment of the mass spectra, and H is the set of the matched fragments of the mass spectra with the protein.

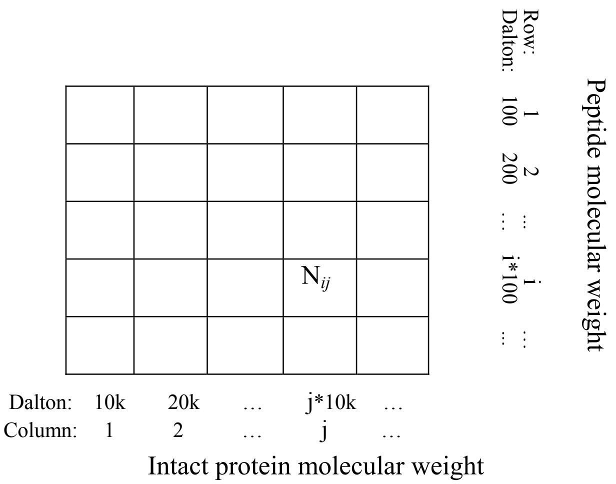

Peptide molecular weight
Row: 1 2 … i …
Dalton: 100 200 … i*100 …
N_{ij}
Dalton: 10k 20k … j*10k …
Column: 1 2 … j …
Intact protein molecular weight

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

B. 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 Figure 1, when comparing a mass distribution of peptides (n fragment molecular weights in the spectra) with the database entry molecular weights (protein k in the column j), R(l) represents the row number of the table for the l-th fragment of the mass spectra. When the difference of two peptide weights is within a tolerance value, it is a "hit" or match. Otherwise it is non-matching. The probability for a match between a mass distribution of peptides and a protein k in the database is computed via

$$\Pr(P_k) = \prod_{i=R(l),\, l\in H_k} \left[1-\left(1-\frac{m_{ij}}{M_j}\right)^{n_{ij}^k}\right] \qquad \text{Equation 2.}$$

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) for protein k. Let W_j be the total number of proteins in the column-j of Figure 1 among the database. m_{ij} represents the average number of occurrences of peptides in cell-(i,j) for one protein of the database, i.e., $m_{ij}=N_{ij}/W_j$, and M_j is the total number of occurrence of peptides in the j-th 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-intensity peaks are more likely to be peaks representing true peptides, whereas low- intensity peaks are more likely to be noise. To account for the peak intensity effect we modify the Equation 2 as

$$\Pr(P_k) = \prod_{i=R(l),\, l\in H_k} \left\{\left[1-\left(1-\frac{m_{ij}}{M_{ij}}\right)^{n_{ij}^k}\right](1-I_l)\right\} \qquad \text{Equation 3}$$

where I_l is the normalized intensity ([0,1])of the l-th spectrum, i.e.,

$$I_l = \frac{1}{1+e^{-\alpha(\hat{I}_l-\bar{I})}} \qquad \text{Equation 4.}$$

In Equation 4, $\hat{I}_l$ is the original intensity, $\bar{I}$ is the average intensity for all selected peaks, and α is an constant. To achieve good prevision in computing, we adopt -logPr(P_k) as the score function for protein identification.

C. Results

We compared the performance of our newly developed scheme (PBSF) with the MOWSE score function [9]. The result shows that PBSF performed significantly better than MOWSE.

Figure 2 shows the details of the comparison of the two scoring functions using 20 sets of PMF spectra, whose proteins were known. The data were provided by the Proteomics Center, University of Missouri-Columbia. For protein identification, we used a set of manually annotated peaks provided by the Proteomics Center. Matched peptides must cover 25% of a protein sequence.

By searching against the database of sprot45 from Swissprot (last updated in January 2005), we obtained 5 top ranks out of 20 samples with the MOWSE scores, while 9 top ranks with the PBSF scores. At other level of ranking field from top 5 to top 50, PBSF performs consistently better than MOWSE, as shown in Figure2.

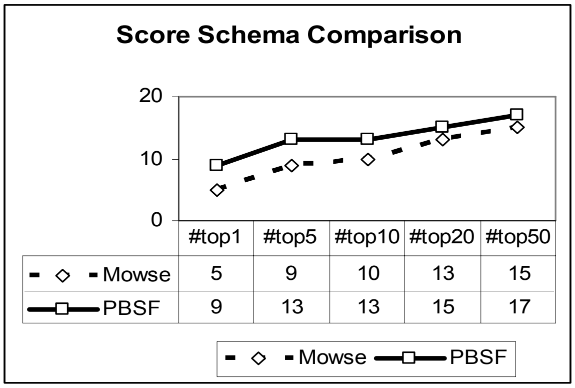


Figure 2. Scoring function comparison

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 10 times, and used each set of generated spectra for protein identification with the 2 scoring schemes, as shown in Figure 3. The result is consistent with Figure 2. Again, PBSF outperforms MOWSE

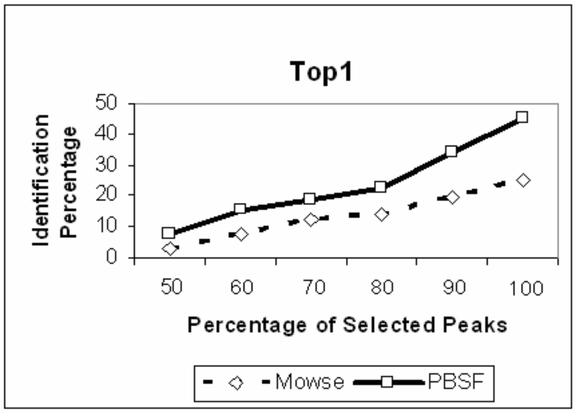


Figure 3. Percentages in which the expected protein ranks top versus percentage of selected peaks from PMF spectra

III. Software

Based on the novel scoring function which we have tested, we developed a standalone software package in Java, "ProteinDecision".

A. Installation and Usage

The software requires JRE 5.0 running environment, which can be freely download from website: http://java.sun.com/j2se/1.5.0/jre/download.jsp . An input file of PMF peak list is required to run "ProteinDecision". The input file contains 2 columns separated by "tab" character. The first column gives centroid value of mass/charge ratio, and the second column provides corresponding peak intensity of that ratio. The data should be sorted by intensity in the descending order, with the following data as an example:

```
568.149208  4865
863.507145  4703
1967.0083823679
1983.9525753183
550.627291  3161
…
```

Generally a fingerprint mass spectrum includes thousands of centroid m/z and intensity. As it is in plain text, users can manually edit the file.

B. Database transformation

The default database used for protein identification in "ProteinDecision" is sprot45 from Swissprot. A user can also incorporate other databases of protein sequences in the FASTA format. For an uploaded database, theoretical enzyme (especially trypsin) digestion of protein is performed and a transformed new data file with the digested proteins is used for matching a PMF spectrum. The transformed file contains 8 fields for each entry (protein): accession number, number of peptides, peptide sequences, peptide masses, peptide lengths, protein sequence, protein name, and protein molecular weight. Each of the field takes one line and contents in each line are separated by "tab". The molecular weight of a peptide of N residues is calculated as:

$$\sum_{i=1}^{N} residue\_mass_i + 18.015 \qquad \text{Equation 5}$$

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

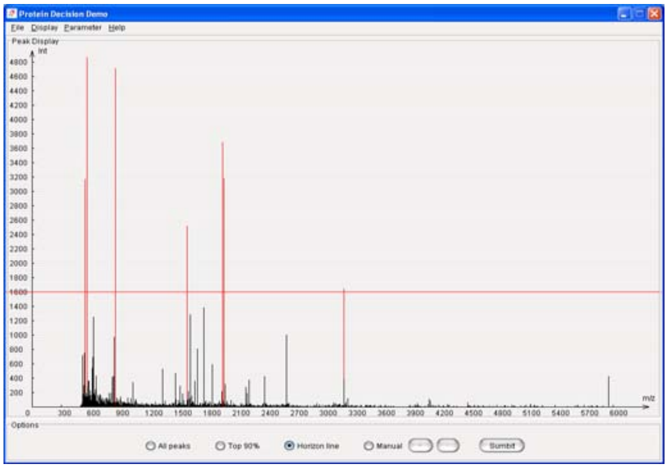


Figure 4. The interface for selecting peaks in the input file. The buttons in the bottom allow a user to choose among four methods of peak picking.

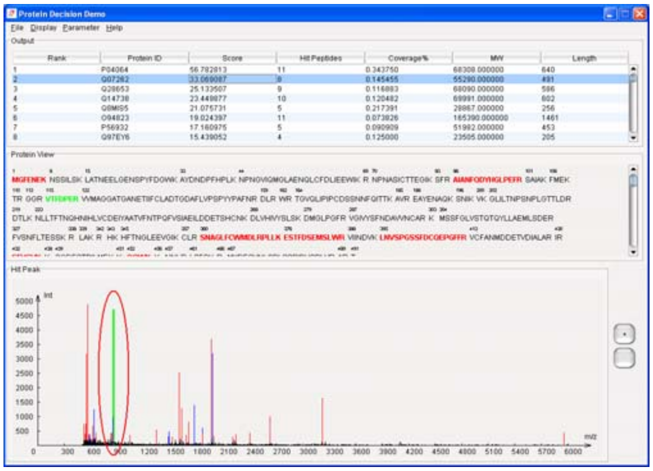
C. Peak Selection

The software supports various ways of peak selections. Four methods are provided as following:

1) "All peaks": Select all of the input peaks in the input file. This option is recommended when the input file is already filtered or all the peaks in the file are likely to represent true peptides.
2) "Top n%": The software provides a user interface to input a number n between 0 and 100, selecting the peaks whose intensity value is larger than n% of the highest intensity of all.
3) "Horizon line": The software provides a user interface to manually select peaks through adjusting the position of a red line. When the red line moves, the peaks whose intensity above the line are selected and those below the line are filtered out (see Figure 4).
4) "Manual": The software provides an option for users to manually pick peaks one by one. When the top of a peak is clicked, it switches the state of "selected" or "not selected" with different colors. Users could click the "+" or "-" button to zoom in or out the view for convenience.

D. Result Analysis

When the selected peaks are submitted, "ProteinDecision" will start automatically to do the computation. Result will be shown in the main panel as shown in Figure 5.



Rank	Protein ID	Score	Hit Peptides	Coverage%	MW	Length
1	P04064	56.782813	11	0.343750	68308.000000	640
2	Q07262	33.069087	8	0.145455	55290.000000	491
3	Q28653	25.133507	9	0.116883	68090.000000	586
4	Q14738	23.448877	10	0.120482	69991.000000	602
5	Q8MIS5	21.075731	5	0.217391	28867.000000	256
6	O94823	19.024397	11	0.073826	165390.000000	1461
7	P56932	17.160975	5	0.090909	51992.000000	453
8	Q97EY6	15.439052	4	0.125000	23505.000000	205

MGFENEK NSSILSK LATNEELGENSPYFDGWK AYDNDPFHPLK NPNGVIQMGLAENQLCFDLIEEWIK R NPNASICTTEGIK SFR AIANFQDYHGLPEFR SAIAK FMEK
TR GGR VTFDPER VVMAGGATGANETIIFCLADTGDAFLVPSPYYPAFNR DLR WR TGVQLIPIPCDSSNNFQITTK AVR EAYENAQK SNIK VK GLILTNPSNPLGTTLDR
DTLK NLLTFTNQHNIHLVCDEIYAATVFNTPQFVSIAEILDDETSHCNK DLVHIVYSLSK DMGLPGFR VGIVYSFNDAVVNCAR K MSSFGLVSTQTQYLLAEMLSDER
FVSNFLTESSK R LAK R HK HFTNGLEEVGIK CLR SNAGLFCWMDLRPLLK ESTFDSEMSLWR VIINDVK LNVSPGSSFDCQEPGFFR VCFANMDDETVDIALAR IR

Figure 5. Visualization of protein identification results. The protein that ranks the 2^{nd} place is displayed, with red labeling matched peptides. A

Results are visualized in three sub-panels. The first sub-panel is the ranking list of top 50 hits including 50 rows and 7 columns. Each row illustrates a protein hit in database sorted by scores in the descending order. The columns represent the properties of entry, i.e., the rank of the entries from 1 to 50, protein ID in the database, the score assigned to the protein using the probability-based scoring function, the number of matched peptides between the spectrum and the theoretical peptides, the coverage of matched peptides in the protein in terms of percentage of residues, the molecular weight of the protein, and the length of the protein, respectively.

The second sub-panel is the sequence information. When a protein displayed in the ranking list is clicked, its detailed information will be shown, as shown in Figure 5. The sequence of the protein, with spaces indicating the digestion boundaries and colored segments of matched peptides.

The third sub-panel is the peaks information. For a special entry in the list, the corresponding peaks information displays. Peaks labeled in red are selected peaks in the input file and peaks in blue represent matched peptides. If the user clicks the matched sequence that is labeling in red in the second sub-panel, the corresponding peak in the third sub-panel and its matched peptide sequence in the second sub-panel will change color to green.

E. Other Features

The software also provides a number of other options, such as setting their own interface style.

IV. Discussion

In our work, we explore a novel scoring scheme by better using the information content in PMF spectra to improve protein identification accuracy. The result illustrates that because of the rigorous statistical treatment, our new scoring function outperformed MOWSE significantly. With the probability-based scoring function, our software performs well in protein identification accuracy, and provides many functions for users to know details of a matching.

There are also some limitations for the current software. We have not yet incorporated other factors for protein identifications, such as missed cleavage, post-translational modification, etc. We are incorporating these factors into our software package in future work and test it on more PMF data.

Acknowledgment

This work has been support by the MU-Monsanto Program. We would like to thank Gary Stacey, David Emerich, Nengbing Tao, Xiufeng Wan, Beverly DaGue, Jay Thelen, Chi-Ren Shyu, and Guohui Lin for helpful discussions. We also like to thank the Proteomics Center at the University of Missouri-Columbia for providing the data.

References

[1] Gevaert, K., Vandekerckhove, J., Electrophoresis. 2000, 21, 1145-1154.
[2] Cottrell, J. S., Pept Res. 1994, 7, 115-124.
[3] Yates 3rd, J. R., McCormack, A. L., Link, A. J., Schieltz, D., Eng, J., and Hays. L., 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., Estreicher, A., Gasteiger E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., Schneider, M, 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., Analytical Chemistry. 1999, 71, 2871- 2882.
[8] Parker, K. C. J Am Soc Mass Spectrom. 2002, 13, 22-39.
[9] Song, Z., Chen, L., Ganapathy, A., Wan, X., Tao, N., Emerich, D., Stacey, G., Xu, D. Electrophoresis. Submitted.