Statistical Assessment for Mass-spec Protein Identification Using Peptide Fingerprinting Approach

Ashwin Ganapathy[1], Xiu-Feng Wan[1], Jinrong Wan[2],
Jay Thelen[3], David W. Emerich[4], Gary Stacey[2], and Dong Xu[1,*]

[1]Digital Biology Lab, Department of Computer Science,
[2]National Center for Soybean Biotechnology, Department of Plant Microbiology and Pathology,
[3]Proteomics Center, [4]Department of Biochemistry, University of Missouri –Columbia
*Correspondence: xudong@missouri.edu

Abstract—We derive and validate a novel statistical model for confidence assessment of protein identification results using peptide mass fingerprint data. We simulate the digestion of the proteins and compare each peptide mass with the input mass. We compute scores from this matching of peptide and compute the distribution of scores for all the proteins in the database. Based on the distribution, we can provide the expectation value for a protein match in the database. We conclude that, given the complexity and noise of the data, the best method for effective confidence matching is using one scoring scheme for matching and another scoring scheme for confidence assessment.

Keywords—proteomics, mass spectrometry, peptide mass fingerprint, confidence assessment.

I. INTRODUCTION

Protein identification using mass spectroscopy (MS) is the dominant technology for proteomics (characterizing proteins at the genome scale). This technology is now playing a major role in studies of biological systems at the molecular level during the post-genomic era [1], with myriad applications, including functional annotation of genes, identification of genes associated with a pathway or a disease, determination of protein interactions, molecular machines, or gene regulation. The general approach for protein identification is through matching the features derived from the mass spectra of peptides against a protein sequence database [2]. It involves protein digestion using a trypsin enzyme, which cleaves with high specificity at the carboxyl side of lysine and arginine residues, chromatographic separation, followed by peptide mass fingerprinting (PMF) [3] or tandem mass (MS/MS) spectrometry analysis [4]. PMF uses intact masses of digested peptides for protein identification whereas MS/MS is based on peptide fragments produced by collision-induced dissociation. Compared to MS/MS, PMF is more economic and efficient although it may be relatively less accurate for peptide identification.

Computational identification of peptides using PMF from MS spectral peaks depends on various features, such as the type of digestion used for the preparation of gel samples, calculation of mass-to-charge ratio of the peptides, and matching protein sequences using peptides. The basic idea of the computational methods is first match the MS spectral peak with possible peptides theoretically digested from proteins in the search database, and then the proteins in the search database with a number of peptide hits are considered as likely candidates from the experimental sample. Several tools have been developed for PMF, among which ProteinProspector [5] and Mascot (Matrix Science Inc., http://www.matrixscience.com/) are the most popular. However, existing analysis tools often give too many matches for a given biological sample and provide no confidence assessment for choosing the best identification of peptides or proteins [5-7]. They assume that the target protein is in the search database and use the best hit from the raw score ranking as the prospective target. This may lead to false positive results since the top hit may not be the query protein. In particular, the raw scores are not normalized based on the protein length, the number of redundant hits for a spectral peak in the search database, etc. The ranking based on raw scores may be misleading. On the other hand, some peptides may be missed due to the noise of the PMF data, and this will cause false negative prediction, i.e., the correct protein was not ranked among top hits selected by computational methods. Given the potential inaccurate data analysis, it is very important to develop a confidence assessment for the PMF data analysis results (1) to get an idea to what extent a user can trust the protein identification result and (2) to re-rank the protein hits based on the confidence assessment instead of raw scores. Such a capacity may significantly improve the computational analysis of PMF data.

In this paper, we describe a novel PMF data analysis method with a statistical assessment approach. To our knowledge, this is the first computational method that gives a comprehensive statistical assessment for PMF data analysis.

II. DATASETS

The PMF application data used for testing were obtained from MU Proteomics Center. The protein samples were digested using modified porcine trypsin (Promega, Madison, WI). The mass/charge (m/z) ions were obtained from a

Voyager DE-Pro MALDI-TOF Mass Spectrometry workstation (Applied Biosystem, Foster City, CA). The monoisotopic m/z values and associated intensities were extracted, which are the input data for our approach. The SWISS-PROT database [8] was utilized as the query database.
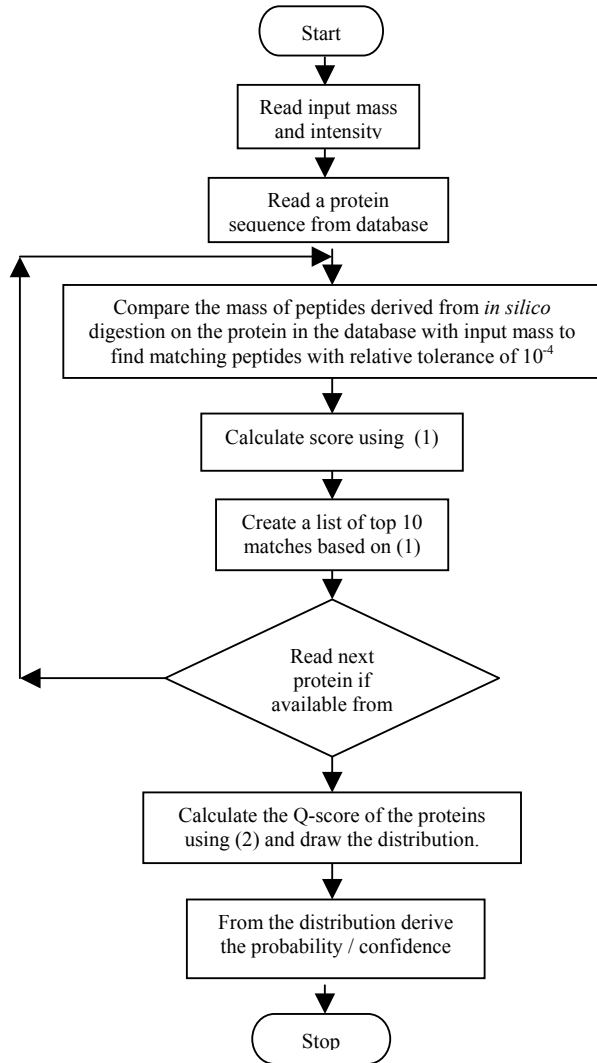
III. METHODS

Fig. 1. Flowchart of the algorithm.

Figure 1 describes the workflow of our algorithm. The proteins in the search database are preprocessed into theoretical peptides with cleavages from a defined enzyme digestion as the constraint. For a set of mass/intensity reading values from PMF spectra, we utilized the ratio (|tm − Mp|)/Mp to determine whether the target peptide is a candidate, where Mp is the observed mass from the PMF spectra and tm is the mass of a theoretical peptide. If (|tm − Mp|)/MP ≤ 10^{-4}, the target peptide will be picked as a candidate peptide. The hit score S, which is a raw score, for a protein j will be calculated using the following equation:

$$S = \left(\sum_i Int_i \middle/ \left| Mp_i - tm_i + d \right| \right) * \left[TP /(1 + FP) \right] \dots \quad (1)$$

where Int_i is the intensity of the query peptide i, Mp_i is the molecular weight for the query peptide i, tm_j is the molecular weight for the theoretical peptide in the database that is within the threshold, and d is a constant to be optimized. TP represents True Positives for peptides hits in query protein (number of peaks in the identified proteins) while FP False Positives (number of peaks not in the identified proteins). The higher this score, the higher possibility to be the match.

To evaluate the statistical significance of identified protein, we treat all of the proteins in the database as the statistical background. The Q-score$_1$ is defined as the ratio between the number of residues with peptide matches and the protein length. The Q-score was normalized by the ratio of the total number of residues with peptide matches in the database vs. the total number of residues in the database:

$$Q\text{-}score_1 = \frac{N_i}{L_i} \middle/ \frac{N}{L} \dots\dots\dots\dots\dots\dots\dots(2)$$

where N_i denotes the number of residues in the peptides matched for protein i in the database, L_i the length of the protein i in the database, N the total number of residues in the peptides matched in the entire database, and L the sum of residues for all of the proteins in the database.

The histogram of transformed Q-score for query proteins will be plotted and fitted as a Gaussian distribution with observed mean μ and standard deviation σ. The proteins in the database with Q-score larger than (μ+2σ) will be treated as significant protein hits. The protein hits are ranked by the generated probability.

IV. RESULTS

To test our algorithm, we also applied our algorithm for the PMF data generated for several known proteins by the Proteomics Center at the University of Missouri.

Table 1 is the prediction for sample 1, where the target protein was known to be horse myoglobin (P02188 Swiss Prot Accession Number). For the horse myoglobin, the top hit has a score S of 201, and we denote the raw score in Equation (2) as $Q\text{-}score_1$ = 62. For better statistical assessment, we perform the following transformations:

Transformation of Qscore :
  Qscore$_2$ = Qscore$_1$ / mean(Qscore$_1$)
  Qscore$_3$ = ln (Qscore$_2$)
  Qscore = (Qscore$_3$ – min(Qscore$_3$))*50

We get the Transformed *Q-score* = 341 and given a Gaussian distribution, we obtain the probability of matching the input m/z values with the mass of the peptide in this protein by chance (expectation value) as $1.9 *10^{-5}$. This means that we have a high confidence that our hit is correct. For this protein, the total number of input m/z, intensity was 11 and we are able to match 9 of the 11 inputs.

Figure 2 shows the Gaussian distribution of transformed Q-scores of all the proteins in the SWISS-PROT database. The regression was plotted using Sigma-Plot 8.0 (SPSS Inc., Chicago, IL). The curve can be fit as the following:

$$y = 2.59 + 44272 * \exp\{-0.5 * ((x - 18867) / 32.38)^2\} \quad (3)$$

With normalized Gaussian distribution, we calculated the probability for true identification given the Q-score. The basic idea is that score lies after 2σ from center ($>\mu+2\sigma$), the result is significant, as there is less chance such a result is achieved by chance; if the score is close to the peak of the Gaussian distribution ($\sim\mu$), the result is insignificant, since the result can be easily achieved by chance without any connection between the PMF data and the matched peptide; if the score shifts from center from the left by σ ($<\mu-\sigma$) or more of the Gaussian distribution, it is very unlikely that the matched peptide in the database represents the true protein in the experimental sample.

Figure 3 is the Gaussian distribution for Sample 2 where the target protein was "membrane-bound acyl-CoA binding protein" (Arabidopsis thaliana, Genbank accession AF320561). Our algorithm found the top hit in the database as "Myosin VIIa, MOUSE species". The distribution tells us that the Q-score lies inside ($\mu+\sigma$) making the confidence level of this prediction is very low. Indeed the top hit was incorrect. For this sample, the score S was 288, and the statistical assessment was with raw *Q-score₁* = 4.03; transformed *Q-score* = 205 and based on Gaussian distribution, the expectation value is 0.29, which mean a low confidence.
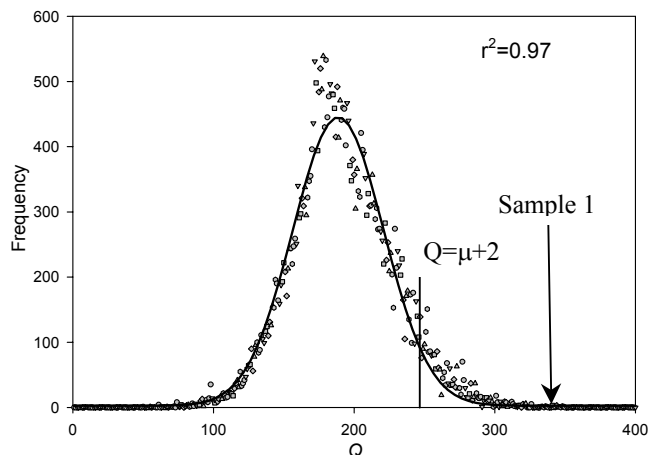

Fig. 2. Prediction for Sample 1. A normal curve is fit to the graph with regression coefficient r^2=0.97
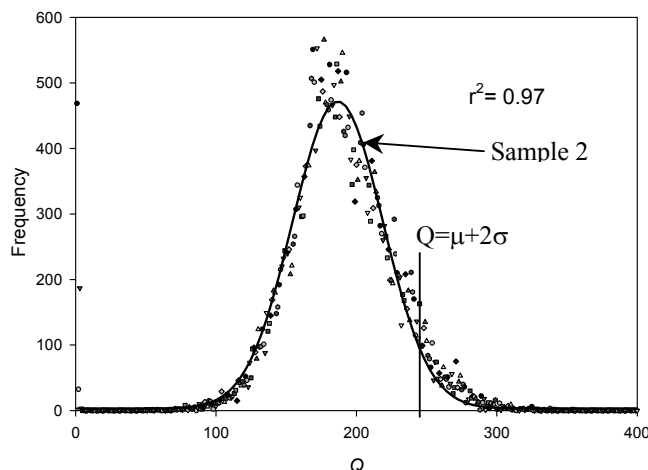

Fig. 3. Prediction for Sample 2. A normal curve is fit to the graph with regression coefficient r^2=0.97.

Table 1. Prediction results for Horse Myoglobin

Start..End	Peptides	Theoretical mass (dalton)	Mass Matched (dalton)	True prediction?	Intensity
1..16	GLSDGEWQQVLNVWGK	1815.903	1815.8389	Yes	25.57
17..31	VEADIAGHGQEVLIR	1606.8554	1606.8289	Yes	100
32..42	LFTGHPETLEK	1271.6636	1271.6260	Yes	10.64
64..77	HGTVVLTALGGILK	1378.8423	1378.8032	Yes	13.24
79..96	KGHHEAELKPLAQSHATK	1982.0572	1982.0261	Yes	5.85
80..96	GHHEAELKPLAQSHATK	1853.9623	1853.9027	Yes	10.58
103..118	YLEFISDAIIHVLHSK	1885.0224	1884.978	Yes	22.39
119..133	HPGDFGADAQGAMTK	1502.6699	1502.6249	Yes	14.71
134..139	ALELFR	748.4358	748.4176	Yes	13.17

The computational mining for the protein was mainly run on a Pentium 4 – 2.8ghz machine with 2GB ram. Mining for a protein with 10 m/z, intensity data points using SWISS-PROT database took 0.12 second. Large-scale validation is ongoing and performed on supercomputers at Oak Ridge National Lab.

V. DISCUSSION

In this paper, we combined a computational approach with statistical assessment for protein identification using PMF data. The final result gives a expectation value indicating the confidence for the protein identification to be true. We demonstrated our algorithm is effective using both positive and negative examples.

One limitation for our algorithm is that currently it does not address the problem of post-translational modification, which has been a great challenge for any MS data analysis. For many cases, the peptides cannot be identified solely by measure the molecular weight based on the protein sequences. Another limitation is that the computing time is long for confidence assessment. We will develop an improved method using a similar idea in BLAST [9]. In this case, the distribution does not have to be produced explicitly for each protein identification. We will also test and improve our methods using more PMF data on known protein samples.

VI. CONCLUSION

In summary, we have tried to find a good statistical model to describe the matching tendency of each protein towards an input mass. We transform the protein hits as Gaussian distribution, by which we can derive the statistical confidence for each protein hit for that database. Our results demonstrate this approach is effective in PMF data analysis.

REFERENCES

1. Ferguson PL, Smith RD. Proteome analysis by mass spectrometry. Annu Rev Biophys Biomol Struct. 2003, 32:399-424.
2. Gevaert K, Vandekerckhove J. Protein identification methods in proteomics. Electrophoresis. 2000, 21:1145-1154.
3. Cottrell JS. Protein identification by peptide mass fingerprinting. Pept Res. 1994, 7:115-124.
4. Yates JR 3rd, McCormack AL, Link AJ, Schieltz D, Eng J, Hays L. Future prospects for the analysis of complex biological systems using micro-column liquid chromatography-electrospray tandem mass spectrometry. Analyst. 1996, 121:65R-76R.
5. Clauser KR, Baker PR, Burlingame AL, Role of accurate mass measurement (+/-10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Analytical Chemistry. 1999, 71:2871- 2882.
6. Eriksson J and Fenyo D, A model of random mass-matching and its use for automated significance testing in mass spectrometric proteome analysis. Proteomics, 2002, 2, 262-270.
7. Helen IF, David Fenyo, Ronald C Beavis, RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimizes protein identification, and archives data in a relational database. Proteomics, 2002, 2,36-47.
8. Bairoch A, Apweiler R, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Research, 1999, 27:49-54.
9. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997, 1;25:3389-402.