

PepWise: Peptide Identification Algorithms for Tandem Mass Spectrometry Based on the Weight of Pair Amino Acid Fracture

Simin Zhu, Wenya Wu, Chuanjie Yang*

China Fire and Rescue Institute, Beijing 102202, China

* Corresponding author. Email address: jianjianjiao1003@163.com

Abstract. Tandem mass spectrometry is the core of the high-throughput techniques for protein identification. Abundant of MS/MS data can be generated and need to be interpreted, although numerous of peptide identification algorithms have been proposed, most well-known algorithms have been prevalently employed to predict fragment m/z value to assign peptide sequences to spectrum, such as X!Tandem, OMSSA, Sequest, SQID and ProVerB incorporate intensity information into algorithms to assist peptide identification. Hence, we can easily know, different algorithms would use different information from the same MS data sets. Here we describe a novel protein algorithm based on the weight of pair amino acid fracture, named PepWise, compared with Mascot, Sequest at 1% False Discovery Rate (FDR), which verified the more accuracy, robustness and compatibility.

Keyword: Index-Terms Tandem Mass Spectrum, Protein Identification Algorithms, Weight of Pair Amino Acid Fracture

1. Introduction

High throughput proteomics involves the analysis of large number of peptide spectra derived, the most common method is the database search, in which the mass spectrum is scored against a corresponding database of all candidate peptides, then detect the effective matches[1,4,6-9]. In order to ascertain the peptide sequences, generally, we need to consider the following four aspects: Firstly, the digested peptides are dissociated and ionized; Secondly, the intact mass of each peptide is measured by MS/MS; Finally, the peptide is mass-selected and fragmented to produce mass spectra and propose peptide identification algorithms to process spectra[2, 3, 10, 11].

According to the description above, how to derive the peptide sequences and propose the reasonable identification algorithms is critical, although it has been improved in identification algorithms for deriving peptide sequences, anyway, a robust scoring function and consider the MS/MS feature information are still the heart of identification algorithms[5,11,13]. In the database search, which aims to evaluate the similarity between the experimental and theoretical MS/MS spectra[12]. Many peptide identification algorithms with various concepts, during the course of similarity analysis, m/z value has been the main information to be integrated into the algorithms[2, 8, 17], e.g. Mascot[7], Sequest[3] and X!Tandem[6]. Despite they are commonly and widely used in protein identification, feature information used singly and the number of peptide identifications inadequate reflect the incompleteness of algorithms which mentioned on the above. Hence, MassWiz, Dispec[9], ProVerB[10] and SQID[2] integrated diversified feature information to scoring model to improve confidence and generate better identification[11-15].

To integrate more abundant and complete feature information and maximize the universality. Firstly, we statistic the matching information of various ions type based on the partial of *S. pneumoniae* D39 data set which has been verified reliable MS/MS spectra; Secondly, define and quantify the weight of pair amino acid fracture; Finally, we integrate the feature information to score function, and propose a novel protein identification algorithm PepWise which is based on the weight of pair amino acid fracture. In order to verify effectiveness and robustness of PepWise, we use multiple MS data sets to compare with Mascot, Sequest at 1% FDR (False Discovery Rate), the results show significantly and stably higher than Mascot and Sequest.

2. Materials And Methods

2.1. MS/MS Datasets

The data sets of 18 protein standard mixtures can be download from public data sets web site ([http://regis-web.systemsbiology.net/PublicData sets/](http://regis-web.systemsbiology.net/PublicData%20sets/)), which contains four types instrument platforms: Thermo Finnigan LCQ DECA, Thermo Finnigan LTQ-FT, Thermo Finnigan LTQ and Micromass/Waters QTOF Ultima, in order to describe conveniently, the instrument names are abbreviated as follows: LCQ, FT, LTQ, QTOF, respectively. The data sets of *S.pneumoniae* D39 and *E. Coli* were obtained from LTQ-Orbitrap, and can be downloaded from the following web sites : [http://bioinformatics.jnu.edu.cn /software/proverb/](http://bioinformatics.jnu.edu.cn/software/proverb/) and <http://marcottelab.org/MSdata/Data03/>, respectively. The partial of *S. pneumoniae* D39 proteome which has been verified valid by Mascot, Sequest and ProVerB are served as training dataset for the feature information of the algorithm model.

2.2. Peaks Selecting

Noise peaks are inevitable in each MS/MS spectra, we need to select reliable peaks to improve SNR (Signal-To-Noise Ratio). Various peptide identification algorithms have different methods to select peaks, such as Sequest[3] selects the highest 200 peaks from mass spectrum, Mascot[7] selects one peak from 14 Da, X!Tandem[6] selects the maximum of 50 peaks from all fragment spectrum and OMMSA[4] selects the top 5 peaks in each 100 Da window. Here, in order to improve SNR, we firstly remove isotope, which peaks closer to $1 \pm 0.25\text{Da}$, secondly, divide the m/z range into ten parts, then select the highest of 20 peaks in each parts.

2.3. Generate the Theoretical Spectra

The core of the database search approach is to evaluate the similarity between the experimental and theoretical spectra. Therefore, generating the theoretical spectrum is critical for the peptide identification algorithm, generated rules as follows:

Rule 1: Loss of H_2O . If b or y ions type involved S, T, E, D residue.

Rule 2: Loss of NH_3 . If b or y ions type involved R, K, Q, N residue.

Rule 3: $+1/+2$ fragment ions. If the parent ions charge were not less than 2 and contained one of R, K, H residue.

2.4. False Discovery Rate (FDR)

All the highest rank candidate peptides are exported to calculate the FDR threshold[19], and the formula is as follows:

$$FDR = \frac{\text{Number of decoy PSMs above threshold}}{\text{Number of target PSMs above threshold}}$$

2.5. Scoring Function

The scoring function is the heart of MS/MS peptide identification algorithms. In this paper, we firstly define the weight of pair amino acid fracture, then construct the scoring function from three aspects: fragment ion matches, consecutive fragment ion matches and b/y fragment ion matches.

2.5.1 Define the Weight of Pair Amino Acid Fracture

The dataset of *S. pneumoniae* D39 is served as training dataset for parameters of the algorithm model, specific methods as follows:

We consider six types of fragment ions, which are b , $b - NH_3$, $b - H_2O$, y , $y - NH_3$, $y - H_2O$, respectively. Count the number of different ions type appear under various pair amino acid fracture, then calculate the probability of different ions type in various pair acid fracture, and the formula is as follows:

$$P_{i,j} = p(AA_j | T_i)$$

Where:

AA is constructed by the set $AA = \{AA, AC, AD, \dots\}$.

j is the element of set $\{1, 2, 3, \dots, 400\}$.

AA_j denotes the j -th pair amino acid fracture.

T is constructed by the set

$T = \{b, b - NH_3, b - H_2O, y, y - NH_3, y - H_2O\}$

i is the element of the set $\{1, 2, \dots, 6\}$.

T_i denotes the i -th ions type.

$P_{i,j}$ is the probability of the j -th pair amino acid fracture AA_j under the i -th ions type T_i .

Here, we need to define the weight of different pair amino acid fractures which under various ions type, the mathematical formula is as follows:

$$W_{AA_j|T_i} = -\log(1 - p(AA_j | T_i))$$

Where:

$W_{AA_j|T_i}$ is the weight of the pair j -th amino acid fracture.

AA_j is the amino acid under the i -th ions type.

T_i is the higher weight of the greater probability.

2.5.2 Scoring Function for MS/MS Spectra

How to evaluate the similarity between experimental spectra and theoretical spectra is crucial for peptide algorithms. For matching a spectrum against a candidate peptide Pep , the score of peptide matching is calculated as follows:

$$Score(Pep) = S(Pep) \times \sqrt{\frac{\sum_{i=1}^{k_0} I_i}{\sum_{i=1}^N I_i}}$$

Where:

Pep is denoted as candidate peptide.

k_0 is the number of experimental peaks matched.

N is the number of experimental peaks which have been selected.

$Score(Pep)$ is the score for candidate peptide against experimental spectra.

$S(Pep)$ is the primary score function and described as follows:

$$S(Pep) = S_0 + S_1 + S_2$$

Where:

S_0 is the score for fragment ion matches: When the distance of experimental spectrum and theoretical spectrum less or equal than fragment error tolerance, they matched.

$$S_0 = \frac{k_0}{0.1406n_0} \sum_l W_l$$

k_0 is the number of experimental peaks matched

n_0 is the number of peaks in the theoretical spectrum

$W_l = W_{AA_j|T_i}$ is the weight of l -th matching, which is denoted as W_l

0.1406 is the random parameter which had been reported by ref 9, calculated by the following formula:

$$\frac{\text{sum of the random peptide matching peaks number}}{\text{sum of the random peptide theoretical peaks number}}$$

S_1 is the score for consecutive fragment ion matches: If $Ions_{-1}$ and $Ions_{-2}$ are consecutive fragment ions match, must satisfy three conditions: matched, belong to the same ions type and only differ of a residue.

$$S_1 = \frac{k_1}{0.0279n_1} \sum_i (W_m + W_q)$$

k_1 is the number of experimental peaks consecutive matches in MS/MS spectra.

n_1 is the number of theoretical peaks consecutive matches in MS/MS spectra.

W_m is the weight of the m -th peak which had been matched against theoretical spectra.

W_q is the weight of the q -th peak which had been matched against theoretical spectra, here, peak m and peak q is a consecutive matches

0.0279 is the random parameter which has been reported by ref 9, calculated by the following formula:

$$\frac{\text{sum of the random peptide consecutive matching number}}{\text{sum of the random peptide theoretical consecutive matching number}}$$

S_3 is the score for b/y ions match: In a typical CID experiment, peptide bond is easy to fracture and generate C-terminal y ions and N-terminal b ions, the content of y ions and b ions reflects the degree of similarity between experimental spectra and theoretical spectra.

$$S_2 = \frac{k_2(\sum_i Wb_i + \sum_i Wy_i)}{0.0706n_2}$$

k_2 is the sum of number of experimental peaks matching to b and y ions in spectrum.

n_2 is sum of the number of theoretical peaks matching to b and y ions in spectrum.

$Wb_i = W_{AA_j|Tb_ions}$, which is the weight of b ions type which based on the j -th pair amino acid fracture AA_j and ions type T_{b_ions} had been matched.

$Wy_i = W_{AA_j|Ty_ions}$, which is the weight of y ions type which based on the j -th pair amino acid fracture AA_j and ions type T_{y_ions} .

0.0706 is the random parameter which has been reported by ref 9, calculated by the following formula:

$$\frac{\text{sum of the random peptide b / y - ions matching number}}{\text{sum of the random peptide b / y - ion theoretical peak number}}$$

3. Results

All peptide identification algorithms need to be compared after FDR calculation. In this paper, we compare PepWise with two widely-used MS identification algorithms Mascot and Sequest, Six different data sets from ISB standard mixture of 18 proteins, the compared results as follows figures and table:

S. pneumoniae D39 data set is simultaneous searched by Msacot, Sequest and PepWise, the number of identified peptides of the three algorithms that mentioned above are more than 3000 at 1% FDR (Fig.1), and higher overlap between Mascot and PepWise; In addition, the number of spectra which PepWise identified is the most highest (Fig.2), and also had higher overlap with others.

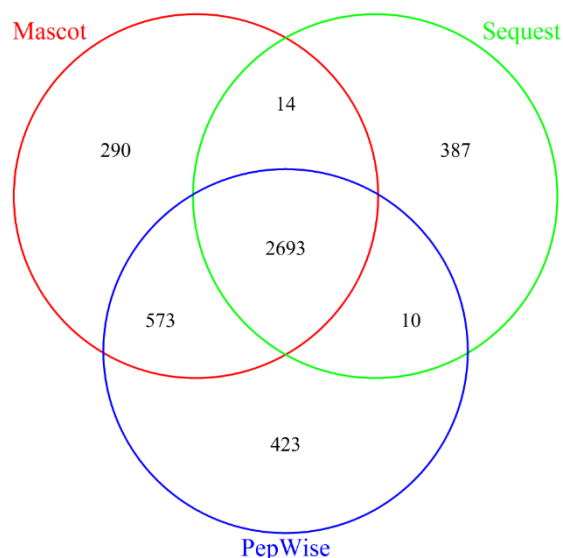


Fig.1. Comparison of Mascot, Sequest and PepWise using D39 data set: number of peptides

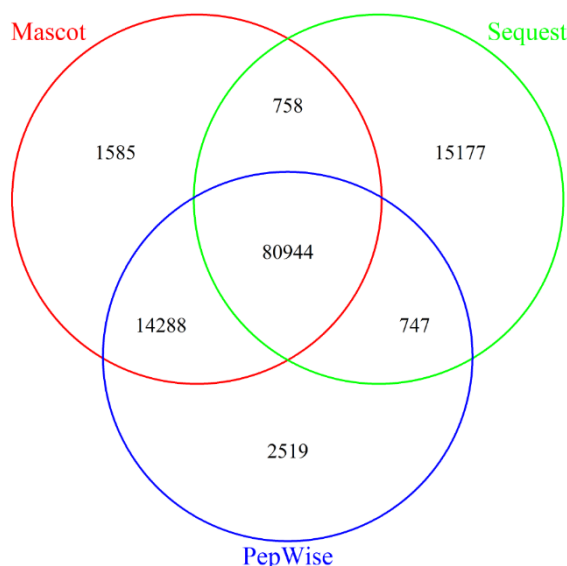


Fig.2. Comparison of Mascot, Sequest and PepWise using D39 data set: number of spectra

In terms of the publicly available standard 18 proteins dataset and E.coli (including E.Coli1, E.Coli2 and E.Coli3) are test under PSMs-Level $FDR \leq 0.01$, the histogram of the number of identified peptides and spectra are showed by Fig.3 and Fig.4. PepWise identified more peptides than Mascot in almost all MS/MS data, the result shows its robustness and high stability.

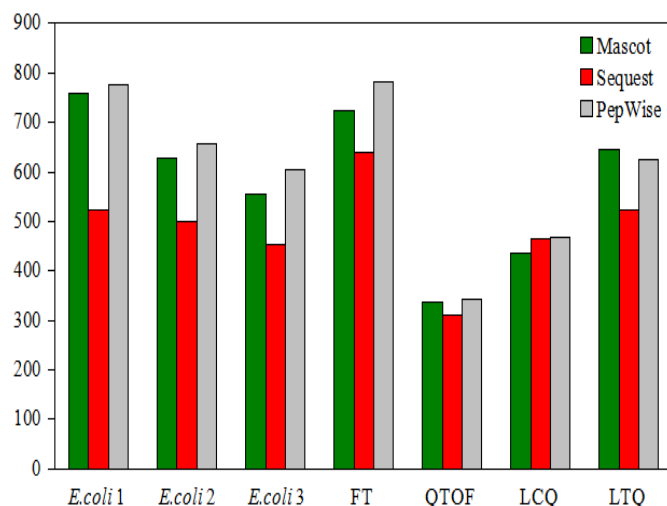


Fig.3. Comparison of Mascot, Sequest and PepWise using various data sets, the height of histogram shows the number of peptides

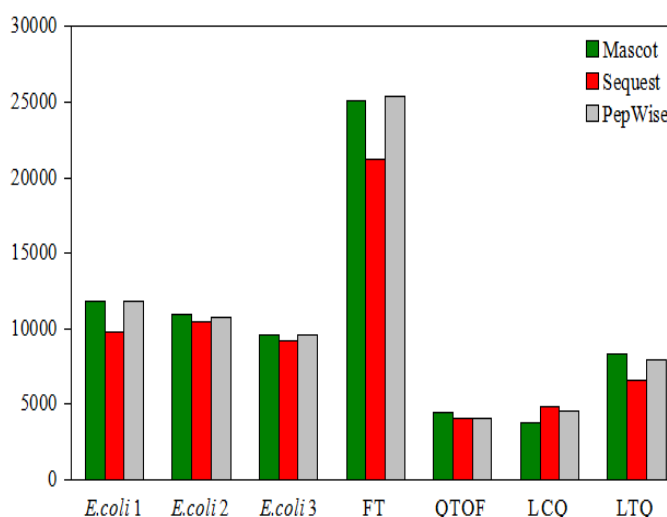


Fig.4. Comparison of Mascot, Sequest and PepWise using various data sets, the height of histogram shows the number of spectra

Overlap of targets PSMs based on the various data sets which are described on the above, PepWise explains large fractions of PSMs also identified by Mascot and Sequest, table 1 showing the overlap results of Mascot, Sequest and PepWise, abbreviated M, S and P, respectively.

Table 1 Overlap among Mscot, Sequest and PepWise

	SUM	M&S	S&P	P&M	M&S&P
D39	3290	2707	2703	3266	2693
		0.82	0.82	0.99	0.82
FT	702	573	575	690	568
		0.82	0.82	0.98	0.81
QTOF	311	272	272	303	268
		0.87	0.87	0.97	0.86
E.Coli 1	663	429	431	653	425
		0.65	0.65	0.98	0.64
E.Coli 2	552	390	389	541	384
		0.71	0.70	0.98	0.70
E.Coli 3	505	345	346	490	338
		0.68	0.69	0.97	0.67
LCQ	444	379	393	408	368
		0.85	0.89	0.92	0.83
LTQ	592	476	464	570	459
		0.80	0.78	0.96	0.78

4. Conclusion

In this paper, we propose a new algorithm called PepWise based on the weight of pair amino acid fraction model, the detail value of the weight can be obtained support info table. According the analysis on the above, we validate its accuracy, robustness, and compatibility. Although PepWise has not been tested the data which is generated by HCD, it reflects the physicochemical attribute of peptide, we suppose the peptide identification algorithm can also support HCD.

5. Conflict Of Interest

One supplementary table can be found in supporting information, which is the weight of all pair amino acid fracture.

Acknowledgments

We are grateful to Beijing scriptless comprehensive emergency drill technology research service for the help of the algorithm design and proposals.

References

- [1] Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 Da. *Anal. Chem.* 1988, 60(20), 2299-2301.
- [2] Li W, Ji L., Goya J, Tan G, Wysocki, VH. SQUID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. *J. Proteome Res.* 2011, 10(4): 1593-1602.
- [3] Eng JK., McCormack AL, Yates JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, 5(11): 976-989.
- [4] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, 20(9): 1466-1467.
- [5] Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, Maynard DM, Yang X, Shi W, Bryant SH. Open mass spectrometry search algorithm. *J. Proteome Res.* 2004, 3(5): 958-964.
- [6] Li D, Fu Y, Sun R. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* 2005. 21(13): 3049-3050.
- [7] Sadygov RG, Yates JR. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. *Analytical chemistry* 2003. 75(15): 3792-3798.
- [8] Cottrell JS, London U. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 1999. 20(18): 3551-3567.
- [9] Yadav AK, Kumar D, Dash D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. *Journal of Proteome Research* 2011, 10(5): 2154-2160.
- [10] Xiao CL, Du YL, Li ZF, Wei L, Zhang G, He QY. Dispec: A Novel Peptide Scoring Algorithm Based on Peptide Matching Discriminability for Candidate Peptides. *PLoS ONE* 2013, 8(5): p. e62724-.
- [11] Xiao CL, Chen XZh, Du YL, Sun XS, Zhang G, He QY. Binomial probability distribution model-based protein identification algorithm for tandem mass spectrometry utilizing peak intensity information. *Journal of Proteome Research* 2013, 12(1): 328-335.
- [12] Fenn JB, Mann M, Meng CK, Wong SF. Electrospray ionization for mass spectrometry of large biomolecules. *Science* 1989, 246(4926): 64-71.
- [13] Khatun J, Ramkissoon K, Giddings MC. Fragmentation characteristics of collision-induced dissociation in MALDI TOF/TOF mass spectrometry. *Anal. Chem.* 2007, 79(8): 3032-3240.
- [14] Chalkley RJ, Hansen KC, Baldwin MA. Bioinformatic methods to exploit mass spectrometric data for proteomic applications. *Methods Enzymol* 2005, 402: 289-312.
- [15] Matthiesen R, Jensen ON. Analysis of mass spectrometry data in proteomics. *Methods Mol Biol* 2008, 453: 105-122.

- [16] Cox J, Mann M, MaxQuant enables high peptide identification rates, individualized p.p.b.- range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 2008, 26(12): 1367-1372.
- [17] Craig R, Beavis RC, TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 2004, 20(9): 1466-1467.
- [18] Gay S, Binz PA, Hochstrasser DF, Appel RD. Peptide mass fingerprinting peak intensity prediction: extracting knowledge from spectra. *Proteomics* 2015, 2(10): 1374-1391.
- [19] Fu, Y, Xiu LY, Jia W, Ye D, Sun RX, Qian XH, He SM. DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. *Mol. Cell. Proteomics* 2011, 10(5): M110.000455.
- [20] Kall L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.* 2008, 7(1): 29-34.