IDFraIP:A Novel Protein Identification Algorithm Based on Fragment Intensity Patterns

Simin Zhu[1, a], Huamei Li[2, b], Kai Zheng[3, c] and *Xiaozhou Chen[4, d]

[1] Key Laboratory of IOT Application Technology, Yunnan Minzu University, Kunming,650500, Yunnan, China.

[2] Key Laboratory of IOT Application Technology, Yunnan Minzu University, Kunming,650500, Yunnan, China.

[3] Key Laboratory of IOT Application Technology, Yunnan Minzu University, Kunming,650500, Yunnan, China.

[4] Key Laboratory of IOT Application Technology, Yunnan Minzu University, Kunming,650500, Yunnan, China.

[a]email: zhusimin2013@126.com, [b]email: li_hua_mei@163.com, [c]email: zhengkai9012@sina.com, [d]email: ch_xiaozhou@163.com

Keywords: Protein Identification Algorithm; Intensity Patterns; MS/MS; Database Search

Abstract. A Identifying peptides for their fragmentation spectra by database search sequencing method is crucial to interpret LC-MS/MS data, widely used algorithms had not been fully exploited the intensity patterns in fragment spectra, SQID incorporated intensity information and identified peptides significantly more peptides than Sequest and X!Tandem. Although SQID adopted various datasets which based on different platforms to show its robustness and effectiveness, many other characterizes were not considered. This article utilized intensity pattern modeling which had been reported by SQID, proposed a novel scoring model to identify fragment spectra. Compared with SQID and Sequest at 1% False Discovery Rate (FDR), IDFraIP identified more confident peptides and spectra.

Introduction

Tandem mass spectrometry (MS/MS) represented a pioneer role for examining the activities and functional states of proteins[1]. In proteomics experiments, large numbers of MS/MS fragment spectra generated, how to interpret and extract high confidence peptides for experimental spectra is crucial to proteomics studies [2, 5, 10-11]. Hence, identifying large-scale spectra by virtue of protein identification algorithms are necessary [5, 7].

Most of the identification algorithms reported in the literature used database search sequencing method, the most important of the above algorithms is to determine similarity between experiment spectra and theoretical spectra [1, 2, 3]. Currently protein identification algorithms primarily utilize predicted fragment m/z value to assign peptide sequences for MS/MS spectra [5, 9, 11], including Sequest [12], X!Tandem [8], Mascot [6]. Intensity information was rarely considered, SQID [5] demonstrated that intensity pattern modeling could improve the number of credible identified peptides and spectra. At the same time, SQID showed us an effective ideology to establish algorithm model [1, 4-5, 9].

Scoring function is the nucleus of peptide identification algorithms [9]. We accorded to the intensity pattern model reported by SQID [5], furtherly rebuilt a novel protein identification algorithm, named IDFraIP. In order to validate the accuracy and robustness of IDFraIP, we compared with SQID and Sequest via various datasets which produced from different platforms at 1% FDR, showing its higher identification and accuracy.

Materials and Methods

MS/MS Datasets. Standard mixtures of 18 proteins from two types of instruments: Thermo Finnigan LTQ-FT and Micromass/Waters QTOF Ultima, abbreviated FT and QTOF, respectively, the datasets could been downloaded from the following web site: https://regis-web.systemsbiology.net//PublicData sets/. The data sets of the *E.coli* proteome spectra downloaded from http://marcottelab.org/MSdata/Data_03/. *S. pneumoniae* D39 data as training dataset that contains more than 270,000 spectra which obtained from http://bioinformatics.jnu.edu.cn/software/proverb/ .

Data Preprocessing. The raw format files of *S.pneumoniae* D39 and *E.coli* needed to convert to dta format files by Bioworks 3.31. when utilized Mascot software to search, the dta format files needed to merge Mascot generic format (MGF) by merge.pl program. The dta format files as the input files of this article method and Sequest software.

Peaks Selecting. Isotope peaks could increase the false positive rate (FPR), removing isotope peaks was needful, the method of removing isotope peaks in this article was as follows: if two peaks closer than $1 \pm 0.25\ Da$ are considered as isotope peaks, the weaker intensity of the peak would be removed.

Meantime, various algorithms provided diverse methods to select effective peaks, SQID and Sequest selected the strongest 80 and 200 peaks from all fragment spectra respectively. While OMMSA select the 50 most peaks from the spectra. Here, we divided the spectra into several bins by 100 Da length and then selected the top six ion peaks in each bin.

False Discovery Rate (FDR). The identified peptides which scores with rank1 PSMs of all spectra needs to be calculated false discovery rate by Kall's method. And the specific formula as follows:

$$FDR = \frac{no.\ of\ \ decoy\ PSMs\ above\ threshold}{no.\ of\ \ t\arg et\ PSMs\ above\ threshold} \tag{1}$$

Scoring Model. Experimental spectra are assigned peptides by scoring against a list of candidate peptides. In protein identification scoring model, the essential aspect is how to evaluate the match level of experimental spectra against theoretical spectra. In order to put forward a reasonable scoring model, we utilized various characterizes to evaluate matching effect, applied Poisson distribution model and considered three aspects: consecutive ions pairs match and b/y ions match:

$$Score(Pep) = \frac{K}{N} \cdot S(Pep) \cdot (1 \pm \rho) \cdot \sqrt{\frac{\sum_{i=1}^{K} I_i}{\sum_{i=1}^{N} I_i}} \tag{2}$$

Where:

Pep = candidate peptide

$Score(Pep)$ = final score for candidate peptide

K = the number of matched peaks in the experimental spectra

N = the number of theoretical fragment peaks

I_i = intensity of the $i\text{-}th$ peak

ρ = the penalty point for consecutive ion matches, which given by ref 12

$S(Pep)$ =primary score for peptide Pep, and defined as follows:

$$S(Pep) = \sum_{i=1}^{k} P(T_i \mid j_i) \cdot e^{P_{r_i}} \tag{3}$$

T_i = the $i-th$ ions type, i.e. $T_i \in \{b, b-H_2O, b-NH_3, y, y-H_2O, y-NH_3\}$

j_i = divided the mass of the $i\text{-}th$ peptide into five section, i.e. $j_i \in \{[0,\frac{1}{5}),[\frac{1}{5},\frac{2}{5}),[\frac{2}{5},\frac{3}{5}),[\frac{3}{5},\frac{4}{5}),[\frac{4}{5},1]\}$

$P(T_i \mid j_i)$ = under prerequisite j, the probability of fragment peak which the $i\text{-}th$ ion type is T_i

P_{r_i} = the $i\text{-}th$ intensity pattern, only depicted b and y ions type, the detailed of statistical model is given by ref 5, and the primarily calculation formula as follows:

$$P_r = \frac{total\ number\ of\ strong\ peaks}{total\ number\ of\ \exp ected\ peaks} \tag{4}$$

Test Results

In this paper, we compare IDFraIP with SQID, Mascot and Sequest at 1% FDR, showing more superiority and higher identification peptides. The following table show the test results.

Table 1. Searching results of various software

	Mascot	**Sequest**	**SQID**	**IDFraIP**
D39	3570	3104	3521	3584
FT	725	640	682	777
QTOF	338	310	340	353
*E.coli*1	758	522	714	774
E.coli2	627	501	584	665
E.coli3	556	452	509	602

Additional, we need to calculate the number of high-confidence peptides, which is the overlap of each two algorithms. Here we only utilized Mascot, Squest and IDFraIP to calculate.

Table 2. High Confidence Peptides

	Mascot & Squest	Sequest & IDFraIP	IDFraIP & Mascot	Mascot, Squest & IDFraIP
D39	2707	2708	3201	2697
FT	573	579	693	570
QTOF	272	270	306	267
*E.coli*1	429	425	619	416
E.coli2	390	391	533	384
E.coli3	345	346	480	338

Summary

We develop a novel algorithm named IDFraIP. Then compared with two software Mascot and Sequest with diverse platforms and experimental datasets at 1% FDR, showing its robustness and versatility.

Acknowledgement

This research is supported by Yunnan Minzu University Graduate InnovationFund(Grant No. 2015YJCXY283)

References

[1] Elias, J.E., et al., Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol, 2004. 22(2): p. 214-9.

[2] Yadav, A.K., D. Kumar, and D. Dash, MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J Proteome Res, 2011. 10(5): p. 2154-60.

[3] Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. Nature2003,422(6928), 198−207.

[4] Washburn, M. P.; Wolters, D.; Yates, J. R., 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol.2001, 19(3), 242−7.

[5] Li, W.; Ji, L.; Goya, J.; Tan, G.; Wysocki, V. H. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry.J. Proteome Res. 2011, 10(4), 1593−602.

[6] Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 1999, 20 (18), 3551−67.

[7] Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm.J. Proteome Res. 2004, 3(5), 958−64.

[8] Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. Bioinformatics2004, 20(9), 1466−7.

[9] Yadav, A. K.; Kumar, D.; Dash, D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J. Proteome Res. 2011, 10(5), 2154−60

[10] Chuan-Le Xiao, Xiao-Zhou Chen, Yang-Li Du, Xuesong Sun, Gong Zhang, Qing-Yu He, Binomial probability distribution model-based protein identification algorithm for tandem mass spectrometry utilizing peak intensity information. Journal of Proteome Research, 12, pp. 328−335, 2013.

[11] Chuan-Le Xiao, Xiao-Zhou Chen, Yang-Li Du, Zhe-Fu Li, Li Wei, Gong Zhang, Qing-Yu He, Dispec: a novel peptide scoring algorithm based on peptide matching discriminability. PLoS ONE, 8(5), p. e62724.

[12] Eng, J. K.; McCormack, A. L.; Yates, J. R. 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.