International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

Database Search Sequencing Algorithms for peptides

Huamei Li, Simin Zhu, Kai Zheng, Xiaozhou Chen*

Key Laboratory of IOT Application Technology of Universities in Yunnan Province，Yunnan Minzu University，Kunming 650031, China

Abstract

Database search sequencing a new method to interpret tandem mass spectrum data. It provides us to research proteomics ions. This mothod is composed of four major parts, namely to remove isotope and select peaks, product theoretical mass spectrum, score for matches between theoretical with experimental spectra. Scoring methods is the heart of database search sequencing algorithms. In this paper we mainly developed a new scoring algorithm, and made a comparison of widely used database algorithms. The results show that our way is effective.

Introduction

Tandem mass spectrometry (MS/MS) emerge in the wake of completed proteome projects [1,3]. In a typically proteomics experiment, large-scale fragment spectra will be generated. How to interpret the spectra rapidly and accurately is still difficult to solve [2,5,10]. Many protein identification algorithms have been proposed, including Mascot [6], Sequest [12], X!Tandem [8], OMMSA [7] and MassWiz [9].

Database search sequencing is the most popular approach to peptide identification, which is pioneered by Yates in the early 1990s. In this approach [2,4], the experimental mass spectrum is scored against theoretical mass spectrum which product by virtual enzymatic to detect significant matches. Additionally[5,9,13], database algorithms implicitly assume the genome is accurately sequenced and all protein coding genes are annotated[5,6,11,14].

1. Database algorithms methods

1.1 Remove isotope and select peaks

Isotope and noise peaks can reduce the SNR (signal to noise ratio), remove isotope and select peaks are necessity for subsequent analysis. In ref 10, peaks closer than 1 ± 0.25 Da are considered as isotope peaks and will be filtered.

Various algorithms proposed different methods to select peaks, Sequest and SQID select the most strongest 200 and 80 peaks from all fragment spectra, respectively. ProverB and MassWiz divide the spectrum dynamically and take top six and five peaks from each window, respectively. OMMSSA select the 50 most intensive peaks by default. Mascot selects the highest peak in each 14 Da mass interval.

1.2 Product theoretical mass spectrum

Database searching approach, which core principle is to evaluate the similarity between the experimental and theoretical spectra. Therefore, generated the theoretical spectrum is critical for the peptide identification algorithm, generated rules as follows:

Rule 1: Loss of H2O. If the b-, y-fragment ions involved S, T, E, D ions.

Rule 2: Loss of NH3. If the b-, y-fragment ions involved R, K, Q, N ions.

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

1.3 Scoring function

Scoring function is the heart of database algorithms,

ISSN:2319-7900

improvements of which are mainly made form developing new scoring algorithms. Sequest is based on empirical scoring model, then computed the cross-correlation between experimental and theoretical spectra. Sequest scoring model is composed of two steps, Preliminary scoring with experimental mass spectrum against theoretical mass spectrum, the formula as follows:

$$S_p = (\sum i_m) n_i (1+\beta)(1+\rho)/n_t \quad (1)$$

The second step is Cross-correlation test.

$$XCorr = [R_0 - (\sum_{\tau-75}^{75} R_\tau)/151]/10^4 \quad (2)$$

Where

$$R_\tau = \sum_{i=0}^{n-1} x[i]y[i+\tau] \quad (3)$$

The meaning of each parameter represented by on the above fomula is given in ref 12.

ProverB is based on probability model, which given by binomial probability distribution. The scoring function is composed of three aspects.

Scoring Function for Simple Fragment Matches

$$\begin{cases} p = p_0 + f \\ p = p(k|n,p) = \binom{n}{k} p^k (1-p)^{n-k} \end{cases} \quad (4)$$

Scoring Function for Consecutive Ion Matches

$$\begin{cases} p_1 = \dfrac{r \cdot k}{n} \\ p_1 = \binom{n_1}{k_1} p_1^{k_1} (1-p_1)^{n_1-k_1} \end{cases} \quad (5)$$

Scoring Function for Spectrum Intensity of b/y-Ion Peaks.

$$\begin{cases} p_2 = \binom{n_2}{k_2} p_2^{k_2} (1-p_2)^{n_2-k_2} \\ p_2 = \dfrac{1+c}{1+T}(0.02+f) \end{cases} \quad (6)$$

The detail of the parameter defined is given in ref 10. SQID as an intensity-incorporated protein identification algorithms, which make use of the coarse intensity from a statistical analysis, the score function as follows:

$$Score = (m+n) \times \frac{1+\sum_{i=1}^{K} \Pr_i}{1+K \times 0.155} \quad (7)$$

Where

m = the number of matched peaks;

n = the number of consecutive ions pairs;

$\Pr$ = the probability for a certain AA pair to have strong peaks;

K = the number of most intense peaks used to calculate the intensity score;

OMMSA and X!tandem are based on Poisson scoring model and hypergeometric scoring model, respectively.

2. Comparison of widely used database algorithms

2.1 Mass spectrum data sets

Standard mixtures of 18 proteins two types of instruments: Thermo Finnigan LTQ-FT and Micromass/Waters QTOF Ultima, abbreviated FT and QTOF, respectively. The data sets based on the two types of instruments which mentioned on the above could download from the following web site: https://regis-web.systemsbiology.net//PublicData sets/. The

ISSN:2319-7900

data sets of the *E.coli* proteome spectra were download from http://marcottelab.org/MSdata/Data_03/. S. pneumoniae D39 data as training dataset that contains more than 270,000 spectra obtained from http://bioinformatics.jnu.edu.cn/software/proverb/.

2.2 Comparison of searching results by widely used algorithms

All peptide identification algorithms need to be compared after 1% FDR calculation, the searching results are given by following table.

Table.1 Comparison of searching results by algorithms

	Mascot	Sequest	OMMSA	SQID	Dispec	ProVerB
D39	3570	3104	3437	3521	3651	3626
FT	725	640	626	682	734	765
QTOF	338	310	277	340	338	357
*E.coli*1	758	522	698	714	819	834
E.coli2	627	501	635	584	687	725
E.coli3	556	452	564	509	606	658

Acknowledgment

This research is supported by Foundations of Educational Committee of Yunnan Province (Grant No. 2013J119C)

Reference

[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. Electrophoresis1999, 20 (18),3551−67.

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

[8] 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

[9] 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.

[10] 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.

ISSN:2319-7900

[11] 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.

[12] 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.

[13] Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J.V.; Mann, M. Andromeda: a peptide search engine integrated into the MaxQuant environment. J. Proteome Res. 2011, 10 (4), 1794−805.

[14] Kapp, E. A.; Schutz, F.; Connolly, L. M.; Chakel, J. A.; Meza, J. E.; Miller, C. A.; Fenyo, D.; Eng, J. K.; Adkins, J. N.; Omenn, G. S. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 2005, 5 (13), 3475 −3490.

Biographies

H.M L is Studying for a master's degree at Yunnan Minzu University, Kunming, Yunnan province. Currently, his research areas include protein identification algorithms and protein microarray data analysis.