RESEARCH ARTICLE | JANUARY 10 2019

PITDI: A novel protein identification algorithm for tandem mass spectrometry based on target-decoy matching information FREE

Xiangyu Lu ✉; Simin Zhu

AIP Conf. Proc. 2058, 020009 (2019)
https://doi.org/10.1063/1.5085522

PITDI: A Novel Protein Identification Algorithm for Tandem Mass Spectrometry Based on Target-Decoy Matching Information

Xiangyu Lu[1, a)], Simin Zhu[1)]

[1]*Nanjing University of Science and Technology, Nanjing 210094, China.*

[a)]Corresponding author email: luxiangyu@njust.edu.cn

Abstract. Database search has become one of the most important method to interpret LC-MS/MS data in proteomics, Many algorithms to identify peptides in MS/MS data have been proposed, Including Mascot, Sequest, X!Tandem, ProVerB and SQID. However, many of them mainly based on peak matches or peak intensity, not on peak intensity identifiability. On the basis of this matter, this article provides a new perspective and characteristics reference information, scilicet PROPIN, which is a novel peptide identification algorithm based on peak intensity identifiability. Compared with Mascot, Sequest, PROPIN identified significantly more peptides from LC-MS/MS data sets at 1% False Discovery Rate (FDR), and also showing its robustness and versatility on various platforms and experimental data sets.

Key words: Protein identification algorithm; peak intensity identifiability; tandem mass spectrometry.

INTRODUCTION

Tandem mass spectrometry (LC-MS/MS) technology's development has been provided an important platform for identifying complex proteome, especially with the advent of soft ionization techniques like electrospray ionization (ESI) and matrix-assisted laser desorption (MALDI)[1-3]. Due to its high sensitivity, rapid detection and the capability to provide accurate information about mass and structure of peptides, hence gradually replaced Edman degradation (Edman) sequencing techniques to become the preferred choice in proteomic studies. In a typical LC-MS/MS experiment, in order to determine peptide sequences, generally, we need to consider from the following four aspects: 1) the digested peptides are dissociated and ionized; 2) the intact mass of each peptide is measured by MS; 3) the peptide is mass-selected and fragmented to produce mass spectra;4) propose peptide identification algorithms to process spectra [1]. Because of enormous spectra generated in proteomic experiments, a sensitivity and accuracy of the peptide identification algorithm is crucial for downstream analyses [1,4,5].

A robust scoring function is the heart of any peptide identification algorithm, the aim is to evaluate the similarity between the experimental and theoretical spectra, and the assigns the best match within the peptide error windows as the result [4,6]. Generally speaking, scoring models mainly consider three aspects: peak matches, peak consecutive matches and the intensities of matched peaks [1,5]. During the course of similarity analysis, m/z value has been the main information to be integrated into the algorithms [1], e.g. Mascot [7], Sequest [8] and X!Tandem [9]. Despite they are commonly adapted search tools and widely used commercial software in protein identification algorithms [10,11]. Feature information is used singly and the number of identifications inadequate which above-algorithms unavoidable. Thence integrate diversified characteristic information in peptide identification algorithms to improve confidence and generate more identifications [4,5,12,13]. e.g. SQID [1], ProVerB [5], OMSSA [14], MassWiz [4]. In order to evaluate the similarity between the experimental and theoretical spectra, above-algorithms apply kinds of ways to select non-noise peaks from spectra. SQID selects the top 80 of the most abundant peaks from the simplified spectrum for scoring, and considering the m/z, consecutive peak match and intensity information in scoring function [1]. ProVerB is based

International Conference on Frontiers of Biological Sciences and Engineering (FBSE 2018)
AIP Conf. Proc. 2058, 020009-1–020009-8; https://doi.org/10.1063/1.5085522
Published by AIP Publishing. 978-0-7354-1786-1/$30.00

on binomial probability distribution model, it selects the top six ion peaks in the 100 Da window, since the authors consider the matching condition of six types of fragment ions, namely b, y, b-NH3, y-NH3, b-H2O and y-H2O [5]. OMSSA is based on a Poisson scoring model and selected the 50 most intensive peaks by default [14]. MassWiz is based on empirical scoring model, considering the weights of major ions, continuity of b-y ions intensities. it divides the spectrum dynamically, then selects the maximum of 5 most intense peaks from each bin [4]. Above-algorithms (i.e. SQID, ProVerB, OMSSA, MassWiz) are integrated more abundant feature information than Mascot and Sequest, and have enhanced the accuracy, completeness, and robustness [1,4,5,14]. However, most peptide identifying algorithms consider the peak intensity not, but only peak matches between experimental and theoretical spectra [7,9,11,14]. Though in many cases the m/z information alone is enough to provide reliable identification, intensity can potentially improve the confidence and generate more identifications because it is also highly dependent on the sequence of peptide and the amino acid residue composition [1]. e.g. ProVerB has introduced the strength probability of the pair-wise amino acid fragments into the scoring function, increase the number of peptide identification [5,15]. However, these algorithms don't involve the entire feature information of peak intensity identifiability.

In order to integrate more abundant and complete feature information into peptide identification algorithm, enhance the ability of identification. We propose a novel peptide identification algorithm which based on peak intensity identifiability. and compare with Mascot and Sequest using multiple MS data sets at 1% FDR, showing its higher capability and sensitivity from the MS.

MATERIALS AND METHODS

Ms Data Set

The data sets (Mix 3) of 18 protein standard mixtures were obtained from four types of instruments (Thermo Finnigan LCQ DECA, Thermo Finnigan LTQ-FT, Thermo Finnigan LTQ and Micromass /Waters QTOF Ultima, abbreviated as LCQ, FT, LTQ, QTOF, respectively), and are downloaded from http: // regis-web. systemsbiology.net/PublicData sets/ to test the accuracy and robustness of the algorithms [16, 17]. The data sets of S.pneumoniae D39 (contained more than 270,000 spectra) proteome and E.coli proteome are obtained from Thermo Finnigan LTQ-Orbitrap (http://bioinformatics.jnu.edu.cn/software/proverb/) and http://marcottelab.org/MSdata/Data 03/ respectively [18], and the dataset of S. pneumoniae D39 proteome is served as training dataset for parameters of the algorithms model.

Data Preprocessing

The raw format files of the S.pneumoniae D39 and E.coli datasets are converted to dta format file by Bioworks 3.31 (Thermo Finnigan, San Jose, CA), Furthermore, the dta format files are merged to Mascot generic format (mgf) via using the merge pl program (http://www.Matrixscience. com/downloads/merge.zip). The downloaded dta format files of the 18 proteins datasets are merged to Mascot generic format (mgf) by the merge.pl program. The dta format files were as the input files of the method and Sequest software.

Mass Spectrometry (Ms\Ms) Database Search

Using the forward and reverse databases to build three datasets for the target-decoy based FDR calculation, specific data of target-decoy S.pneumoniae D39 database, 18 proteins database and E.coli database respectively contain 3828,3644 and 8558 protein sequences. Mascot 2.3 (Matrix Science, London, U.K.), Sequest 28.13 (Thermo Fisher Scientific, Waltham, MA) and PROPIN are used to search input files, Mascot 2.3 search mgf files and Squest 28.13 search dta files. Searches are performed with full tryptic specificity, trypsin digestion with two missed cleavage and Cys (+57.021464 Da, Carbamidomethylation) and Met (+15.994915 Da, oxidation) are considered as fixed and variable modification respectively. According to MS/MS characteristics, we set the values of precursor ion mass tolerance and fragment ion mass tolerance as in Table 1. The fragment ion tolerance of Sequest is considered as 1.0 Da, due to it requires an integer value for m/z [21, 22, and 23].

TABLE 1. The parameters of precursor and fragment ion tolerance settings

Instrument Type	PROIDIN and Mascot		Sequest	
	Precursor ion tolerance	Fragment ion tolerance	Precursor ion tolerance	Fragment ion tolerance
LCQ_Deca	3.0 Da	0.5 Da	3.0 Da	1.0 Da
LTQ	3.0 Da	0.5 Da	3.0 Da	1.0 Da
LTQ-FT	10 ppm	0.5 Da	10 ppm	1.0 Da
QTOF	0.2 Da	0.2 Da	10 ppm	1.0 Da
LTQ-Orbitrap	10 ppm	0.5 Da	10 ppm	1.0 Da

False Discovery Rate

For calculating FDR threshold, the peptide spectrum matches (abbreviated below as PSMs) which are extracted from the Mascot's results format files (. dat) and Sequest output results format files (. out) with the highest rank are exported. PROPIN's results and the extracted results of Mascot and Sequest are written to csv format files. To calculate FDR values by Kall's method, all target and decoy scores with rank 1 PSMs are sorted in ascending order [19,20]. We can get the value of FDR from the following formula:

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

The score threshold is adjusted to reach FDR <= 1%. Scoring function is the heart of peptide identification algorithms. Hence in different search algorithms, scoring functions are mutative. e.g. To calculate FDR of sorting ion scores for Mascot, peptide length needs to be greater than 6; Whereas calculating FDR of sorting Xcorr scores for Sequest, peptide length ≥ 6 and $\Delta Cn \geq 0.1$; for PROPIN, the Sp scores are sorted to calculate FDR when peptide length ≥ 6.

Training Matching-Peak Intensity Identifiability

D39 data set: contains 97535 spectra and 3570 unique peptides with high quality and confidence, are consider as correct results. Simultaneously, corresponding reversed sequences were considered as incorrect results. These correct and incorrect peptides are applied to quantify matching-peak intensity identifiability in the search process of target-decoy database.

RESULT

Peak Selection in the Spectra

The quality of any peptide identification algorithm depends on the data it receives. Spectral quality is of great importance for any algorithm [4]. In order to establish a more reasonable algorithm, we have employed a simple effective method to select peak in the spectra. Specific operation as follows:

Step 1: Isotope removed. Peaks are considered as isotope peaks if closer than 1±0.25 Da, and then were filtered. The purpose of this is to get the minimum number of peaks can be defined for a spectrum to be considered for search, reduces random matches and enhance the accuracy.

Step 2: Peak selection: Open a window of 100 Da, selected the top six ion peaks in it since we consider the matching condition of six types of fragment ions, respectively named b, y, b-H2O, y-H2O, b-NH3, y-NH3.

Theoretical Spectra

The chemical properties of b/y-ions fragmentation decided the types of theoretical spectrum. 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 is not less than 2 and contained one of the R, K, H residues.

Quantifying Matching-Peak Intensity Identifiability

To quantify matching-peak intensity identifiability utilizing the intensity information, we divide the intensity interval [0, 1] between experimental and theoretical fragment ions into 12 intervals: [0, 0.05], [0.05, 0.1], [0.1, 0.2], [0.2, 0.3], [0.3, 0.4], [0.4, 0.5], [0.5, 0.6], [0.6, 0.7], [0.7, 0.8], [0.8, 0.9], [0.9, 0.95], [0.95, 1] respectively. In each interval, the correct and incorrect matching number of each type ions (b, b-H2O, b-NH3, y, y-H2O and y-NH3) in the training dataset can be statistically calculated in each interval. The matching-peak intensity identifiability Tij is calculated by the formula as follows:

$$T_{ij} = \frac{N(r_{ij})}{N(e_{ij})}$$

Where

j: The j-th interval of above-intervals (j=1, 2, ..., 12);

i: The i-th type ion in the j-th interval (i∈b/b-H2O/b-NH3/y/y-H2O/y-NH3);

N(rij) = the correct (correct peptides) matching peak number.

N(eij)= the incorrect (revered peptides) matching peak number.

The matching-peak intensity identifiability of each type ion is statistically obtained from the training dataset, as listed in Table 2.

TABLE 2. The quantified matching-peak intensity identifiability of each type ion (b, b-H2O, b-NH3, y, y-H2O and y-NH3).

Range	[0,0.05]	[0.05,0.1]	[0.1,0.2]	[0.2,0.3]	[0.3,0.4]	[0.4,0.5]	[0.5,0.6]	[0.6,0.7]	[0.7,0.8]	[0.8,0.9]	[0.9,0.95]	[0.95,1]
b	0.90	1.44	2.04	2.59	3.11	3.53	3.86	4.09	4.18	4.34	4.34	1.08
b-NH3	0.74	1.29	1.56	1.77	1.91	1.90	1.98	2.01	2.00	1.97	1.72	1.00
b-H2O	1.93	2.94	3.46	3.81	3.97	3.96	4.00	4.03	4.03	3.85	2.76	2.41
y	4.04	6.77	10.00	13.96	18.38	21.13	23.90	27.09	30.21	28.25	32.39	5.05
y-NH3	3.53	3.06	2.60	2.21	1.97	1.76	1.55	1.43	1.33	1.29	1.12	3.37
y-H2O	1.11	0.74	0.53	0.42	0.34	0.31	0.28	0.26	0.25	0.24	0.24	0.98

Scoring Function

A robust and accurate scoring function is the heart of any peptide identification algorithms. Due to the variability in the fragmentation, the extent of fragmentation and the intensities of the peaks, instruments and experimental method, made the task full of challenging. In order to establish PROPIN algorithm, we utilize the information of matching-peak intensity identifiability to evaluate the identification and matches. The scoring function is considered three aspects, respectively fragment ion matches, consecutive fragment ion matches and b/y fragment ion matches. The score model of all candidate peptides is described as follows:

(1) Scoring function for fragment ion matches. When matching an experimental peak to theoretical peak of fragment ion from a candidate peptide, we should consider the type of theoretical peak and which interval corresponding intensity identifiability belongs to. The fragment ion matching-peak intensity identifiability of the l-th matching is defined as Il=Tij the score of fragment ion matching is calculated as:

$$S_0 = \frac{k_0}{0.1811 n_0} \sum_l I_l$$

Where

k0 = number of experimental peaks matched.

n0 = number of peaks in the theoretical spectrum.

0.1811 = It represents incorrect matching probability of theoretical spectrum, which reflects the matched ability between experimental spectrum and decoy theoretical spectrum and is obtained by the following formula:

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

(2) Scoring function for consecutive ion matches. Multiple consecutive ion matches need to convert into a series of adjacent ion pairs matches. e.g. if b1, b2, b3 are consecutively matched, converted into two consecutive pairs: b1-b2 and b2-b3. the score of consecutive fragment ion matching is calculated as:

$$S_1 = \frac{k_1}{0.0828 n_1} \sum_l (I_m + I_p)$$

Where

Im = the intensity identifiability of the m-th matched peak.

Ip = the intensity identifiability of the p-th matched peak. Here, a consecutive ion match comprises of the l and m matches.

k1 = number of experimental consecutive matches.

n1 = number of theoretical consecutive matches.

0.1811 = It represents incorrect matching probability of theoretical spectrum, which reflects the matched ability between experimental spectrum and decoy theoretical spectrum and is obtained by the following formula:

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

(3) Scoring function for b/y-fragment ion. Due to intensity identifiability of b/y-ions (especially for y-ion) is mostly more than the other ion types. It can be inferred that b/y-ions matches are more efficient and important in the peptide identification algorithm. Hence, we take b/y-ion intensity identifiability separately and consider in the scoring function. To score the b/y-ion peaks, the b/y-ion identifiability is defined as:

$$Ib_l = T_{i=b,j}$$

$$Iy_l = T_{i=y,j}$$

And the score of b/y fragment ion matching is calculated as:

$$S_2 = \frac{k_2 (\sum_l Ib_l + \sum_l Iy_l)}{0.0604 n_2}$$

Where

k2 = number of the peaks matching to b-ions and y-ions

n2 = number of b-ions and y-ions in theoretical spectra

0.0604 = It represents b and y ions incorrect matching probability of theoretical spectrum, which is one third of the incorrect matching probability of theoretical spectrum (0.1811).

The final score Sp for the candidate peptide against the experimental spectrum is the sum of the above three scores:

$$Sp = 0.01 * (S_0 + S_1 + S_2) .$$

Comparison of Propin with Mascot and Sequest

All algorithms need to be compared after FDR calculation. PROPIN is compared with two widely-used MS identification algorithms Mascot and Sequest using S. pneumoniae D39 dataset, 18 standard proteins mixture and E. coli datasets.

The number of the identification results of S. pneumonia D39 dataset are more than 3000 peptides and 97500 spectra at FDR ≤ 0.01, including 2698 peptides and 81020 spectra identified by the above three-alhorithms (Figure 1A and 1B). We map the overlaps between Mascot and PROPIN for all identified peptides and spectra. It shows highly overlap ratios of 90.4% and 97.1%, respectively, and showing a good consistency of PROPIN with other algorithms. PROPIN identifies more peptides than Mascot and Sequest in the FDR range of 0.0%~4.5%.

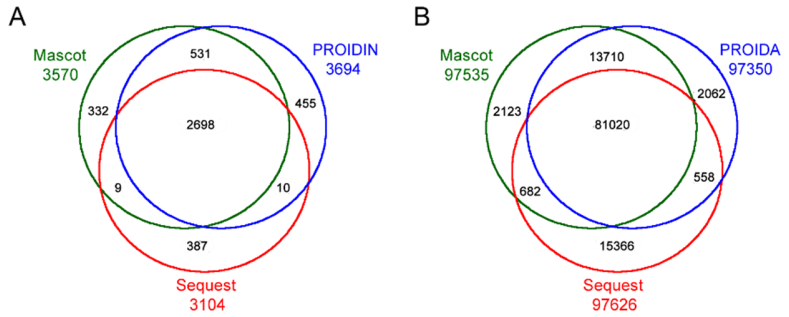

FIGURE 1. Comparison of Mascot, Sequest and PROIDIN using S. pneumoniae D39 dataset.

(A) Number of identified peptides. (B) Number of identified spectra

The 18 protein mixture dataset (FT, LTQ, LCQ, and QTOF) and E. coli dataset (LTQ-Orbitrap) show difference among PROPIN, Mascot, Sequest at all charge state. According to test PROPIN's adaptability and robustness at FDR ≤ 0.01 level (Figures 2 and 3), identify more peptides than Mascot and Sequest in all MS data, showing its robustness, stability and extensiveness.

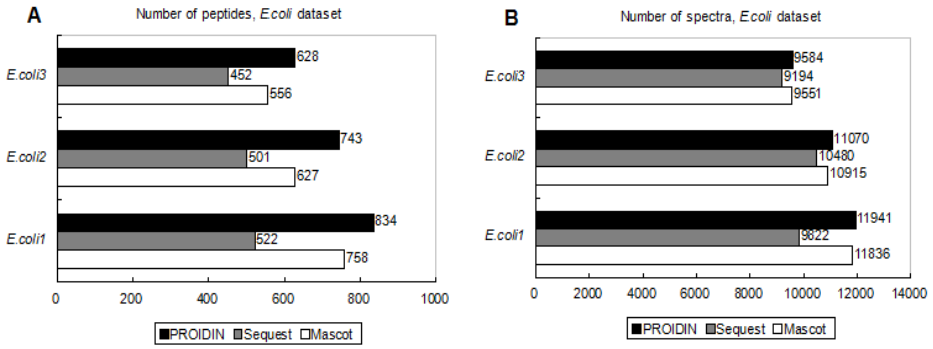

FIGURE 2. The number of identified peptides (A) and spectra (B) at FDR≤ 0.01 from E.coli dataset using PROIDIN, Mascot and Sequest

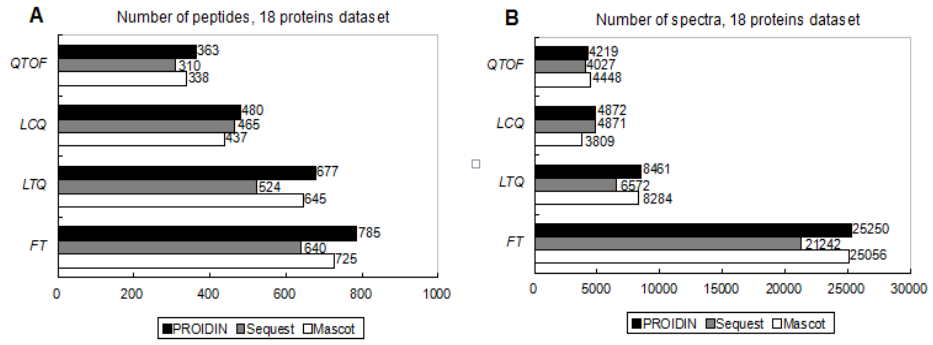

FIGURE 3. The number of identified peptides (A) and spectra (B) at FDR≤ 0.01 from standard 18 protein dataset using PROIDIN, Mascot and Sequest

The Number of High-Confidence Peptides Identified

Identification results is an reference standard of peptide identification algorithms, which embodies the algorithms are reasonable [4]. Clearly, different algorithms give different identification results [15]. Hence, a set of high-confidence peptides is required to compare the algorithms for their quality of matches [4,5]. We adopt a method of obtaining intersection among the multiple algorithms to enhance the confidence of the peptide identification.

'High-confidence' peptides (abbreviated as H), which signify peptides identified in at least two of the above three search algorithms and are calculated by the formula as follows:

$$H = (A \cap B) \cup (B \cap C) \cup (A \cap C)$$

Where

A: It represents the identified peptides from PROPIN.

B: It represents the identified peptides from Mascot.

C: It represents the identified peptides from Sequest.

The number of each two algorithm's intersection is shown in Figure 4. According to analyzing the number of High-confidence peptides in Figure 4, showing PROPIN exceed Mascot and Sequest in all case, simultaneously, proving it's robust and accuracy in diverse instruments and datasets.

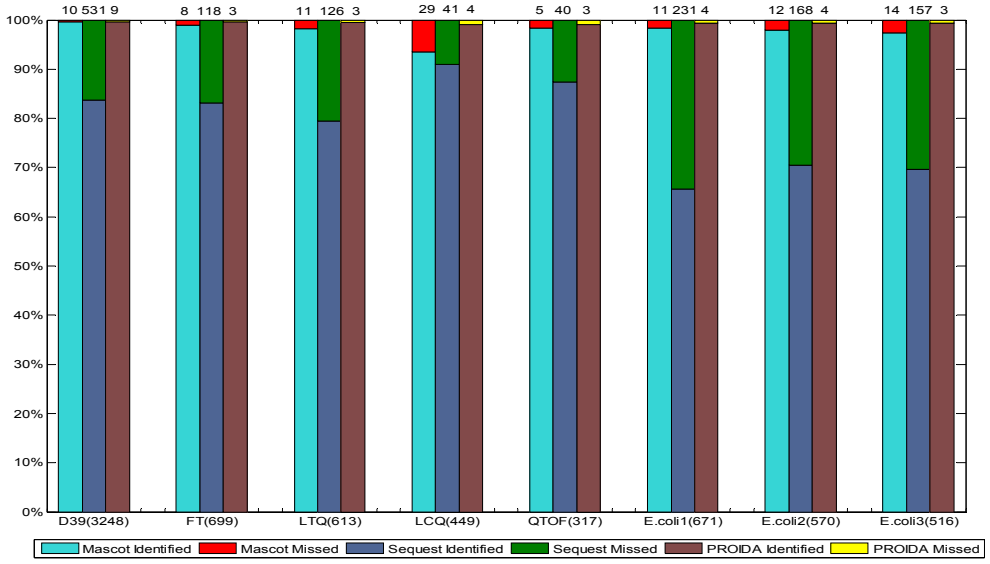

FIGURE 4. The high-confidence peptides of all three algorithms at FDR≤ 0.01

SUMMARY

PROPIN is a novel concept peptide identification algorithm based on matching-peak intensity identifiability, showing improved performance compared with two widely used algorithms Mascot and Sequest with three tested datasets. We believe that combining PROPIN with other algorithms will have potentially beneficial, such as increasing the number of confidence of identifications and integrated into new identification algorithms. As a new algorithm, PROPIN still requires further optimization to improve the overall performance, and developing more effective and robust algorithms for high-confidence peptide identification deeply.

REFERENCES

1. Li W, Ji L, Goya J, Tan G, Wysocki VH. SQID: an intensity-incorporated protein identification algorithm for tandem mass spectrometry. J Proteome Res 10: 1593-1602 (2011).
2. Karas M, Hillenkamp F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Analytical chemistry 60: 2299-2301 (1988).

3. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. Electrospray ionization for mass spectrometry of large biomolecules. Science 246: 64-71 (1989).
4. Yadav AK, Kumar D, and Dash D. MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. J Proteome Res 10: 2154-2160 (2011).
5. Xiao C-L, Chen X-Z, Du Y-L, Sun X, Zhang G, et al. Binomial Probability Distribution Model-Based Protein Identification Algorithm for Tandem Mass Spectrometry Utilizing Peak Intensity Information. Journal of Proteome Research 12: 328-335 (2012).
6. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature methods 4: 787-797 (2007).
7. Cottrell J, London U. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20: 3551-3567 (1999).
8. Eng JK, McCormack AL, Yates Iii JR. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry 5: 976-989 (1994).
9. Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20: 1466-1467 (2004).
10. An MR, Zou X, Wang QS, Zhao XY, Wu J, et al. High-Confidence de Novo Peptide Sequencing Using Positive Charge Derivatization and Tandem MS Spectra Merging. Analytical Chemistry 85: 4530-4537 (2013).
11. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. J Proteome Res 10: 1794-1805 (2011).
12. Dagda RK, Sultana T, Lyons-Weiler J. Evaluation of the consensus of four peptide identification algorithms for tandem mass spectrometry based proteomics. Journal of proteomics & bioinformatics 3: 39 (2010).
13. Kapp EA, Schütz F, Connolly LM, Chakel JA, Meza JE, et al. An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis. Proteomics 5: 3475-3490 (2005).
14. Geer LY, Markey SP, Kowalak JA, Wagner L, Xu M, et al. Open mass spectrometry search algorithm. Journal of proteome research 3: 958-964 (2004).
15. Xiao CL, Chen XZ, Du YL, Li ZF, and Wei L, et al. Dispec: A Novel Peptide Scoring Algorithm Based on Peptide Matching Discriminability. Plos One 8 (2013).
16. Klimek J, Eddes JS, Hohmann L, Jackson J, Peterson A, et al. The standard protein mix database: a diverse data set to assist in the production of improved peptide and protein identification software tools. The Journal of Proteome Research 7: 96-103 (2007).
17. Fu Y, Xiu L-Y, Jia W, Ye D, Sun R-X, et al. DeltAMT: a statistical algorithm for fast detection of protein modifications from LC-MS/MS data. Molecular & Cellular Proteomics 10 (2011).
18. Ramakrishnan SR, Vogel C, Prince JT, Wang R, Li Z, et al. Integrating shotgun proteomics and mRNA expression data to improve protein identification. Bioinformatics 25: 1397-1403 (2009).
19. Käll L, Storey JD, MacCoss MJ, Noble WS. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. Journal of proteome research 7: 29-34 (2007).
20. Elias JE, Haas W, Faherty BK, Gygi SP. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. Nature methods 2: 667-675 (2005).
21. Goeminne LJ, Gevaert K, Clement L. Experimental design and data-analysis in label-free quantitative LC/MS proteomics: A tutorial with MSqRob. Journal of Proteomics; 171:23-36 (2017).
22. Townsend C, Furukawa A, Schwochert J, Pye C, Edmondson Q, Lokey RS. CycLS: Accurate, whole-library sequencing of cyclic peptides using tandem mass spectrometry. Bioorganic & Medicinal Chemistry. 26:1232-1238 (2018).
23. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nature Methods. 14:513-520 (2017).