MSPosDM: A Novel Peptide Identification Algorithm Optimized for Tandem Mass Spectra

Simin Zhu, Chuanjie Yang, Wenya Wu

China Fire and Rescue Institute, Beijing 102202, China

Abstract. Tandem mass spectrometry (MS/MS) plays an extremely important role in proteomics research. Thousands of spectra can be generated in modern experiments, how to interpret the LC-MS/MS is a challenging problem in tandem mass spectra analysis. Our peptide identification algorithm, MSPosDM, is integrated the intensity information which produced by target-decoy statistics, although intensity information often undervalued. Furthermore, in order to combine the intensity information for better, we propose a novel concept scoring model which based on Poisson distribution. Compared with commonly used commercial software Mascot and Sequest at 1% FDR, the results show MSPosDM is robust and versatile for various datasets which obtained from different instruments. We expect our algorithm MSPosDM will be broadly applied in the proteomics studies.

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

1. Introduction

In biological sample analysis, mass spectrometry (MS)-based proteomics has evolved into an indispensable approach [7, 11]. In proteomics experiment, proteins can be cleaved into peptides by enzyme-selected, then separated and enter MS for subsequently analyzing [7, 12]. Thousands of fragmentation spectra generated in modern proteomics experiment, how to infer peptide sequence is a challenging and propose peptide identification algorithms are necessity [1, 12].

Algorithms model greatly affect the efficient and accuracy of searching spectra [10]. Scoring function is the core of protein identification algorithms, integrated the current algorithms types, they could be divided into four categories as follows [14, 16, 19].

(1) Correlation matching model: by mathematical simulation of protein digestion and mass spectrometry detection process, the theory of enzyme peptides transform into the corresponding predicted spectrum, then the degree of correlation between predicted spectrum and experimental spectrum needed to be evaluated by mathematical approach, obtained the appropriate search results. Representative algorithms such as Sequest [6] and pFind [12].

(2) Probabilistic matching model: utilizing the statistical probability which obtained by census the frequency of a certain m/z value in a certain error range in protein database to indicate the credibility of matching [15, 17], then constructed a reasonable algorithm model, searching the correct peptides. Representative algorithms such as Mascot, OMSSA [2], X!Tandem [4], Andromeda and ProVerB [18].

(3) Random matching model: using the information of proteins mass distribution and theory of enzyme peptides mass distribution, the corresponding spectrum was divided into several intervals, by calculating random probability of each section matching to the selected peptides, then built the identification algorithms model. Representative algorithms such as SCOPE [3] and Probity [8].

(4) Empirical weight matching model: by assigned different empirical weights to key ions, consecutive occurrence, intensities, pair-wise amino acid patterns and ect. Representative algorithms such as MassWiz [19] and SQID [13].

Integrated the characterizes information to scoring algorithms was inevitable, which could improve the confidence of searching results and efficiency. m/z value often as the main characterize information to be assigned the mainstream search engines, contained Mascot, Seuqest, X!Tandem and OMSSA. Peak intensity characterizer often undervalued because of its unreliability, in order to integrate the intensity information, favourable peaks selecting manner was imperative, SQID and ProVerB utilized diverse manner to select efficiency peaks, and studied the fragmentation intensity patterns, built the protein identification algorithms, achieved excellent results. Dispec proposed a novel concept characterize information based on peptide
matching discriminability (PMD) [1], which abundant reflected the properties of experimental spectrum. Hence, appended more luxuriant characterize information would improve the efficiency and reliable of identification results [6].

In this paper, we proposed a novel peptide identification algorithm, named MSPoisDM, which integrated a brand-new concept characterize information Peak Intensity Identification-ability (PII). PII measured the degree of real matching, meantime, we built a novel scoring model for adding PII information. To validate the reliability and accuracy of MSPoisDM, we utilized diverse datasets from various mass spectrometer platforms to test, compared with Mascot and Sequest at 1% FDR level [9], MSPoisDM showed more robust and higher identification.

2. Methods

2.1 MSPoisDM Identification Algorithm
MSPoisDM, which in virtue of Poisson distribution to construct a novel scoring model and consider to add the PII information. We adopted Matlab (version: 8.1.0.604. (R2013a)) as the programming language. How to via the training experimental spectra to obtain PII characterize information was crucial for our identification algorithm. Here, we through the following aspects for introducing the algorithm designing process.

(1) Isotopes discarding: plentiful isotopes exist in nature, so experimental spectra did. The existence of isotope peaks led to more random matching instead of real matching, the key of discarding isotope was that correctly judged isotope peaks or not. The specific rules as following: if the two peaks closer than $1 \pm 0.25 \text{Da}$ were considered as isotope peaks and the lower intensity peak needed to be discarded. This treatment could reduce random matches and enhance the accuracy.

(2) Peaks selecting: different peptide identification algorithms had diverse manner to select efficient peaks. Sequest and SQID selected the strongest 200 and 80 peaks in each experimental spectra respectively; OMSSA divided the spectra into several bins and then selected the top 5 peaks in each bin. MS Amanda selected the $m$ most intense peaks in each $100 \text{Da}$. In this article, we adopted dynamic approach which had been reported by ProVerB to select peaks, MSPoisDM selected the top 6 peaks in $100 \text{Da}$ window.

(3) Extracting PII characterize information: integrated abundant characterizes information to enhance the accuracy of scoring algorithms were necessary. PPI was a measure of real matching or not, and the specific extraction process included three aspects:

(a) Training dataset: utilized the rational datasets for training was extremely important. The training spectra of MSPoisDM was extracted from the identified spectra which Mascot, Sequest and ProVerB all identified and be controlled by 1% FDR level. Hence, we considered the training spectra were collect identified.

(b) Statistical method: different statistical method generated various results, in order to obtain PII characterize information, statistical process comprised three aspects: firstly, confirm key ion type, here we only defined $b$, $b-NH_{2}$, $b-H_{2}O$, $y$, $y-NH_{3}$, $y-H_{2}O$ as key ion type; secondly, divided the peaks intensities which had been normalized into 12 intervals, the details showed in table 1. Meantime, the method of normalization just like the following formula:

$$I_{\text{Normalization}} = \frac{I_{\text{Original}}}{I_{\text{Max}}}$$ (1)

Where $I_{\text{Max}}$ was the mean of the top three peaks, enhance the reliability of the statistical method. Third, searching the training data set based on forward and reverse reference sequence respectively, recorded the matching results.

(c) Quantitative mathematical: quantified the statistical results involved on the above was crucial for MSPoisDM, we adopted the following formula to quantify, which not only retained variation, but also made the quantitative results get better smoothness. Specific process as follows:

$$PPI(i,j) = \frac{N(F,i,j)}{\sum_{i=1}^{n} \sum_{j=1}^{m} N(F,i,j)}$$ (2)

Where $i$ denoted the sum of key ion types; $j$ denoted the sum of the number of intensity interval; $F$ denoted the forward reference sequence; $R$ denoted the reverse reference sequence; $N(F,i,j)$ was the number of fragment ion matches which ion type was $i$ and intensity value located at $j$ interval based on forward sequence; $N(R,i,j)$ was the number of fragment ion matches which ion type was $i$ and intensity value located at $j$ interval based on reverse sequence; $PPI(i,j)$ reflected the degree of real matching which ion type was $i$ and intensity value located at $j$ interval. Table 1 was the calculated PPI value.

(4) False Discovery Rate (FDR) calculated: no search engine could ensure all identified results were correct, the peptide spectrum matches (PSMs) were exported to calculate the FDR threshold. We used our in-house Matlab code to extract Mascot and MSPoisDM output result files which peptide length $\geq 6$, Sequest results were extracted from output files which PSMs with the highest rank and $\Delta Cn \geq 0.1$, meantime, the peptide length $\geq 6$. And the FDR was calculated by Kall’s method, respectively. The specific formula as follows:

$$\text{FDR} = \frac{\text{no. of decoy PSMs above threshold}}{\text{no. of target PSMs above threshold}}$$ (3)

(5) Scoring algorithm: scoring function is the heart of the peptide identification algorithms. In this article, firstly, we considered three-dimensional characterizes into MSPoisDM, contained fragment matches, consecutive fragment matches and $b/y$-ion matches [1]; secondly, constructed the scoring function based on Poisson distribution, respectively; finally, integrated PPI
characterize information into scoring model. Specific details as follows:

(a) Fragment matches: proposed a universal scoring function for various strategies is hard. We solved the problem by utilizing Poisson distribution to build appropriate function, the formula of Poisson distribution as below:

$$P_0 \{ X = K_0 \} = \lambda_0^{K_0} \cdot e^{-\lambda_0} / K_0 !$$

Where $K_0$ reflected the number of fragment matches; $P_0$ reflected the probability of $K_0$ matches, which embodied the confidence of fragment matches; $\lambda_0$ reflected the theoretical mean of fragment matches, and the value of $\lambda_0$ could be calculated from the following formula:

$$\lambda_0 = 0.06 N_0$$

Where $0.06$ reflected the random match probability, because of we selected the top 6 peaks from each 100 Da in the experimental spectra; $N_0$ reflected the number of theoretical matches. Meantime, when $\varepsilon_0 = [\lambda_0]$, the probability of Poisson distribution arrived maximum.

$$P_{0 \text{ max}} \{ X = \varepsilon_0 \} = \lambda_0^{\varepsilon_0} \cdot e^{-\lambda_0} / \varepsilon_0 !$$

The preliminary score of fragment matches calculated from the following formula.

$$S_{0 \text{ pre}} = \text{sgn}(K_0 - \varepsilon_0) \cdot \lg \frac{P_{0 \text{ max}}}{P_0}$$

(b) Consecutive fragment matches: consecutive fragment matches characterize information was hard to integrate into scoring algorithms, we via to take fragment matches and consecutive fragment matches as two independent information and scored separately, which could improve the efficiency of MSPoisDM. If termed two fragment matches as consecutive matches, they must satisfy two conditions: belonged to the same ion-type and the differ just equal the mass of a residue. And the scoring process as follows:

$$P_1 \{ X = K_1 \} = \lambda_1^{K_1} \cdot \frac{e^{-\lambda_1}}{K_1 !}$$

Where $K_1$ denoted the number of consecutive fragment matches; $P_1$ reflected the probability of $K_1$ consecutive fragment matches; $\lambda_1$ reflected the theoretical mean of consecutive fragment matches and calculated from the following method:

$$\lambda_1 = 0.09083 \cdot K_0 / N_0$$

Where $0.09083 \cdot K_0 / N_0$ reflected the random consecutive match probability, which reported by ProVerB; $N_1$ reflected the number of theoretical consecutive matches. When $\varepsilon_1 = [\lambda_1]$, the probability arrived at maximum.

$$P_{1 \text{ max}} \{ X = \varepsilon_1 \} = \lambda_1^{\varepsilon_1} \cdot \frac{e^{-\lambda_1}}{\varepsilon_1 !}$$

Like the scoring strategy of fragment matches, here, we also needed to calculate the preliminary of the consecutive matches. The specific process as follows:

$$S_{1 \text{ pre}} = \text{sgn}(K_1 - \varepsilon_1) \cdot \lg \frac{P_{1 \text{ max}}}{P_1}$$

Where $S_{1 \text{ pre}}$ is the preliminary score of the consecutive fragment matches. Then the final scoring strategy which integrated PPI information as following formula:

$$S_{1 \text{ Final}} = \begin{cases} \sum_{j=1}^{\# PPI} H_j \cdot S_{1 \text{ pre}} \geq 0, \\ \sum_{j=1}^{\# PPI} H_j \cdot S_{1 \text{ pre}} < 0 \end{cases}$$

If two matches were a couple of consecutive match, and corresponding PPI value were $PPI_s$ and $PPI_t$, respectively, and the value of $H_j$ could be calculated by following formula:

$$H_j = PPI_s + PPI_t$$
(c) b/y-ion matches: b/y-ion were the mainstream ion type under CID environment. Evaluated the efficiency of b/y-ion matches could improve the robust and accuracy of peptide identification algorithm. Hence, we considered the b/y-ion matches as the independent information into scoring model.

\[ P_2 \{ X = K_2 \} = \lambda_2^{-K_2} \cdot e^{-\lambda_2} / K_2! \]  
(15)

Where \( K_2 \) reflected the number of b/y-ion matches; \( P_2 \) reflected the probability of \( K_2 \) b/y-ion matches; \( \lambda_2 \) reflected the theoretical mean of b/y-ion matches and calculated from the following method:

\[ \lambda_2 = 0.02 \cdot N_2 \]  
(16)

Where 0.02 was the b/y-ion random probability, \( N_2 \) was the number of the theoretical b/y-ion matches. When \( \varepsilon_2 = [\lambda_2] \), the probability obtained the maximum value.

\[ P_{2,\text{max}} \{ X = e_2 \} = \lambda_2^{-\varepsilon_2} \cdot e^{-\lambda_2} / e_2! \]  
(17)

Similarly, the preliminary score of b/y-ion matches could be obtained by the following formula:

\[ S_{2, \text{pre}} = \text{sgn}(K_2 - e_2) \cdot \log \frac{P_{2,\text{max}}}{P_2} \]  
(18)

Meantime, the final score of b/y-ion matches as follows:

\[ S_{2, \text{final}} = \begin{cases} S_{2, \text{pre}} \cdot e^{\sum_{i=1}^{S_{2, \text{pre}}} \text{PPL}_{i\cdot b/y}}, & S_{2, \text{pre}} \geq 0 \\ S_{2, \text{pre}} \cdot e^{-\sum_{i=1}^{S_{2, \text{pre}}} \text{PPL}_{i\cdot b/y}}, & S_{2, \text{pre}} < 0 \end{cases} \]  
(19)

Where \( S_{2, \text{final}} \) was the score of b/y-ion mates.

(d) The score of candidate peptide: the score of candidate peptide could be calculated by the following formula:

\[ S = S_{0, \text{final}} + S_{1, \text{final}} + S_{2, \text{final}} \]  
(20)

Where \( S \) was the score of candidate peptide, it measured the similarly degree between experimental spectra and theoretical spectra. The highest score of candidate peptides was treated as the final searching result.

### 2.2 MS/MS Datasets and Search Engine

We utilized various data sets which based on different instrument platforms. Standard mixtures of 18 proteins obtained from four types of MS instruments, included Thermo Finnigan LTQ-FT, Thermo Finnigan LCQ DECA, Thermo Finnigan LTQ and Micromass/Waters QTOF Ultima. In order to narrate convenience, we abbreviated the names of mentioned above as FT, LCQ, LTQ and QTOF, and public download website is https://regis-web.systemsbiology.net/PublicData sets/. The public data sets of E.coli downloaded from http://macrotelab.org/MSdata/Data_03/, which contained three sub-data sets, named E.coli1, E.coli2 and E.coli3. S.pneumoniae D39 dataset based on LTQ-Orbitrap was obtained from http://bioinformatics.jnu.edu.cn/software/proverb/, not only served as the training data set for extracting PPI characterize information, but also utilized as test data set; the data set of yeast was obtained from ophthalmology central of Sun Yat-Sen university, the data set was generated by HCD.

Mascot and Sequest were widely used in proteomics research, which were adopted to compare with MSPoisDM. The version of Mascot and Sequest were 2.3 and 28.13 respectively. When used Mascot engine to search, the dta format files transferred into Mascot generic format (Mgf) files by merge.pl which was download from Mascot official web. Dta format files as input for Sequest and MSPoisDM. In addition, searching criteria were applied for Mascot, Sequest and MSPoisDM, specific contained cysteine (+57.021446 Da, Carbamidomethylation) defined as fixed modification, methionine (+15.994915 Da, oxidation) defined as variable modification and full tryptic specificity. Other parameters were set in table 2.

<table>
<thead>
<tr>
<th>Instruments</th>
<th>Mascot &amp; MSPoisDM</th>
<th>Sequest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PIT</td>
<td>FIT</td>
</tr>
<tr>
<td>LCQ_DecA</td>
<td>3.0 Da</td>
<td>0.5 Da</td>
</tr>
<tr>
<td>LTQ</td>
<td>3.0 Da</td>
<td>0.5 Da</td>
</tr>
<tr>
<td>LTQ-FT</td>
<td>10 ppm</td>
<td>0.5 Da</td>
</tr>
<tr>
<td>QTOF</td>
<td>0.2 Da</td>
<td>0.5 Da</td>
</tr>
<tr>
<td>LTQ-Orbitrap</td>
<td>10 ppm</td>
<td>0.5 Da</td>
</tr>
</tbody>
</table>

### 3. Results

MSPoisDM was compared with Mascot and Sequest after FDR calculation, the data set of S.pneumoniae D39 showed MSPoisDM & Mascot had higher overlap than MSPoisDM & Sequest. The overlap of MSPoisDM & Mascot was 88.3%, but MSPoisDM & Sequest only was 74.0%. Figure 1 showed the overlap between the two from Mascot, Sequest and MSPoisDM.
The data sets of standard mixtures of 18 proteins, which instruments contained FT, QTOF, LCQ and LTQ. MSPoisDM identified peptides was the most of the algorithms which mentioned above from the instruments except LTQ, showed its robust and steady. Meantime, MSPoisDM identified more spectra than Sequest from any MS instrument. Figure 2 and Figure 3 were the identified peptides and spectra from the four instruments mentioned above respectively.

The data sets of E.coli, which contained three subsets E.coli1, E.coli2 and E.coli3, the identified peptides and spectra from E.coli data sets were the most of all the three search engines, showed MSPoisDM had high identified and superiority. Specify details revealed in Figure 4 and Figure 5.

For verifying the accuracy of MSPoisDM, we should calculate the overlap between the two search engines of all. The peptides which identified at least by two search engines were defined as high confidence peptides. According to the Figure 6 and Table 3 showed MSPoisDM had more high confidence peptides than others.
Table 3. High confidence peptides of Mascot, Sequest and MSPoisDM from varies data sets

<table>
<thead>
<tr>
<th></th>
<th>M &amp; S</th>
<th>S &amp; MP</th>
<th>M &amp; MP</th>
<th>M &amp; S &amp; MP</th>
<th>H_P</th>
</tr>
</thead>
<tbody>
<tr>
<td>D39</td>
<td>2707</td>
<td>2712</td>
<td>3237</td>
<td>2700</td>
<td>3256</td>
</tr>
<tr>
<td>FT</td>
<td>573</td>
<td>577</td>
<td>686</td>
<td>569</td>
<td>698</td>
</tr>
<tr>
<td>QTOF</td>
<td>272</td>
<td>271</td>
<td>310</td>
<td>269</td>
<td>315</td>
</tr>
<tr>
<td>E.coli1</td>
<td>429</td>
<td>434</td>
<td>646</td>
<td>425</td>
<td>659</td>
</tr>
<tr>
<td>E.coli2</td>
<td>390</td>
<td>395</td>
<td>548</td>
<td>388</td>
<td>557</td>
</tr>
<tr>
<td>E.coli3</td>
<td>345</td>
<td>352</td>
<td>491</td>
<td>342</td>
<td>504</td>
</tr>
<tr>
<td>LCQ</td>
<td>379</td>
<td>421</td>
<td>413</td>
<td>376</td>
<td>461</td>
</tr>
<tr>
<td>LTQ</td>
<td>476</td>
<td>476</td>
<td>579</td>
<td>466</td>
<td>599</td>
</tr>
</tbody>
</table>

Table 4. High confidence spectra of Mascot, Sequest and MSPoisDM from varies data sets

<table>
<thead>
<tr>
<th></th>
<th>M &amp; S</th>
<th>S &amp; MP</th>
<th>M &amp; MP</th>
<th>M &amp; S &amp; MP</th>
<th>H_S</th>
</tr>
</thead>
<tbody>
<tr>
<td>D39</td>
<td>81702</td>
<td>81896</td>
<td>94975</td>
<td>81224</td>
<td>96125</td>
</tr>
<tr>
<td>FT</td>
<td>18080</td>
<td>18296</td>
<td>23810</td>
<td>17986</td>
<td>24214</td>
</tr>
<tr>
<td>QTOF</td>
<td>3212</td>
<td>3162</td>
<td>3906</td>
<td>3093</td>
<td>4094</td>
</tr>
<tr>
<td>E.coli1</td>
<td>7499</td>
<td>7418</td>
<td>10854</td>
<td>7345</td>
<td>11081</td>
</tr>
<tr>
<td>E.coli2</td>
<td>7769</td>
<td>7688</td>
<td>10103</td>
<td>7600</td>
<td>10360</td>
</tr>
<tr>
<td>E.coli3</td>
<td>6845</td>
<td>6776</td>
<td>8865</td>
<td>6681</td>
<td>9124</td>
</tr>
<tr>
<td>LCQ</td>
<td>3274</td>
<td>3324</td>
<td>3556</td>
<td>3149</td>
<td>4856</td>
</tr>
<tr>
<td>LTQ</td>
<td>5805</td>
<td>5325</td>
<td>6626</td>
<td>4933</td>
<td>7890</td>
</tr>
</tbody>
</table>

Figure 7 showed FDR from the range of 0 ~ 5%, MSPoisDM identified peptides was the most in 0.005 ~ 0.45 of all the mentioned peptide identification algorithms, showing its reliable and significance.

4. Discussion

MSPoisDM proposed a novel peptide identification algorithm optimized for tandem mass spectra, and integrated the PPI characterize information, according to the diversity data sets from different MS instruments, showed MSPoisDM robust, accuracy and steady. Meantime, for verifying the generality of MSPoisDM, we adopted the yeast data set from HCD to test [5], and MSPoisDM identified more peptides than Mascot. Hence,
MSPoisDM was a universal peptide identification algorithm for tandem mass spectra.

Acknowledgements

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

References


