Journal of Bioinformatics and Computational Biology


# PRIMA: PEPTIDE ROBUST IDENTIFICATION FROM MS/MS SPECTRA

JIAN LIU

*Department of Biomedical Engineering,*
*McGill University, Montreal, QC H3A 2B2 Canada.*
*Email: jian.liu4@mcgill.ca*

BIN MA

*Department of Computer Science, University of Western Ontario*
*London, ON N6A 5B7 Canada.*
*Email: bma@csd.uwo.ca*

MING LI *

*School of Computer Science, University of Waterloo*
*Waterloo, ON N2L 3G1, Canada.*
*Email: mli@uwaterloo.ca*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

In proteomics, tandem mass spectrometry is the key technology for peptide sequencing. However, partially due to the deficiency of peptide identification software, a large portion of the tandem mass spectra are discarded in almost all proteomics centers because they are not interpretable. The problem is more acute with the lower quality data from low end but more popular devices such as the ion trap instruments.

In order to deal with the noisy and low quality data, this paper develops a systematic machine learning approach to construct a robust linear scoring function, whose coefficients are determined by a linear programming. A prototype, PRIMA, was implemented. When tested with large benchmarks of varying qualities, PRIMA consistently has higher accuracy than commonly used software MASCOT, SEQUEST and X! Tandem.

*Keywords*: proteomics, peptide sequencing, tandem mass spectrometry, machine learning

## 1. Introduction

Proteomics aims at understanding proteins expressed in cells at different levels, during different time and in different forms. This task is critical as changes in protein expression levels are often associated with disease states or the variations of

*The author is also affiliated with City University of Hong Kong, Hong Kong SAR and Tsinghua University, Beijing, China.

metabolism. Mass spectrometers are currently the predominant tool to accomplish some of the primary goals of proteomics [1]: (1) identification of each protein in a cell; (2) determination of expression level of each protein (which does not always correlate with mRNA level); and (3) determination of post-translational modifications (PTMs), sites and types. However, due to the high-throughput capacity of mass spectrometers, software tools become a bottleneck to success. Today, in proteomics companies and academic consortiums worldwide, over half of the MS/MS data generated by mass spectrometers are rejected because they are not interpretable by currently available software (e.g. MASCOT or SEQUEST). The interpretable parts are further plagued by false positives. As pointed out in [2]: "... our ability to generate data now outstrips our ability to analyze it." Mass spectrometer accuracy and sensitivity varies greatly and this problem is particularly prominent with low-end but more popular ion trap devices.

This paper focuses on developing a robust and systematic method to deal with the lower quality data produced by the popular ion trap devices, as well as the high quality data consistently. There are two approaches for peptide identification from MS/MS data: *de novo* sequencing and database searching. In order to deal with the low quality data, we use the more popular database method. Using a linear programming formulation, we optimize a scoring function to score the experimental spectra against a protein sequence database. We have implemented the prototype PRIMA and demonstrated the improved performance over both MASCOT and SEQUEST on large spectrum benchmarks.

## 2. Background and related work

A protein sequence is a chain of amino acids connected by peptide bonds. Typically, peptide sequencing through tandem mass spectrometry follows a multistep procedure. First, proteins are digested by enzyme such as trysin, and the resultant peptides are separated by liquid chromatography. Then peptides undergo a fragmentation process to produce ions of different types (Fig. 1). Finally tandem mass spectra are generated by recording the mass/charge (m/z) ratio and intensity of each ion; and computer software are used to reconstruct the sequences from the spectra.

There are two classes of approaches for peptide sequencing via tandem mass spectrometry. *De novo* sequencing method determines the peptide sequence solely from the experimental spectra without using databases [3]. This method is useful when the protein is not in the database. The mainstream *de novo* sequencing software include program packages from mass spec vendors (MassLynx, BioAnalyst, denovoX, etc), the free program Lutefisk [4], and commercial programs PEAKS [5,6] and SpectrumMill. The basic *de novo* sequencing dynamic programming techniques were first introduced in [7,8].

Another approach is database search method, and it depends on the fact that the target protein sequence is in the database. Given an experimental spectrum $S$,

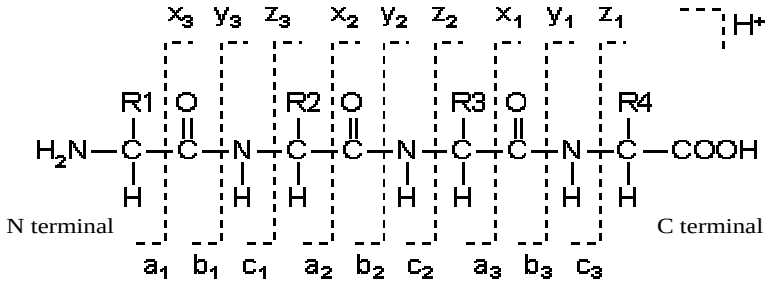


Fig. 1. Different ions produced by peptide fragmentation. $a/x$, $b/y$, $c/z$ are complementary ions, respectively. $b/y$ ions are the most common ones. (source: Matrix Science Inc.)

this method searches through a protein sequence database to find a peptide whose theoretical spectrum $S'$ matches $S$ the best. The mainstream software using the database method includes MASCOT [9] and SEQUEST [10,11]. SEQUEST compares the theoretical spectra against experimental spectrum using a correlation function to determine the score. MASCOT computes the score based on the probability that observed match of ions is a random event. Improvements to these programs are claimed with various criteria: fewer false positives [12], less time [13], validation [14], simultaneous analysis of multiple spectra [15], and new approaches [16,17]. Some other recent research have been presented in [18,19,20,21,22,23,24,25,26] and an empirical evaluation of various approaches was performed have in [27].

Since the coverage and quality of protein databases constantly improve, searching against databases is commonly used in practice. This paper focuses on a robust solution to the low quality spectra for database search.

Given a spectrum, we can find a set of candidate peptides from the protein database whose masses are within a predefined mass error tolerance to the precursor ion mass of the spectrum. For a large database (such as NCBInr), this list can be as large as 100,000 tryptic peptides, using $\pm 2$ dalton error tolerance. A powerful scoring function is then needed to single out the correct peptide from the entirety of candidates.

Constructing a good scoring function is tricky due to multifold reasons. First, the fragmentation of the peptides is determined by their physiochemical characteristics as well as many other factors, resulting many problems listed below.

- A peptide may be broken more than once, resulting in ions of internal fragmentations.
- Ions can be multiply charged (e.g. 2 or 3).
- Some ions may be missing in the experimental spectra, while noise peaks correspond to non-existing ions.
- Isotopic peaks may exist.
- Other ions ($a$, $c$, $x$, $z$) appear at different rates with various types of mass

spectrometers (See [28]).
- Ions (such as $b$ and $y$ ions) can lose an ammonium or water group.

To make the matter worse, each type of mass spectrometer has its own sensitivity and resolutions, the parameters of scoring function often need to be adjusted to achieve the best performance [29]. Further more, there are other problems such as post-translational modification (PTM) of the proteins. As a result, the spectra generated from mass spectrometers often have little resemblance of the corresponding theoretical spectra.

## 3. Constructing a linear scoring function

We are interested in designing a robust scoring function that is relatively insensitive to machine types, noise levels, and error tolerances. Our approach is to first find some features reflecting similarity between experimental and theoretical spectra from different perspective, then build a strong scoring function upon the ensemble of features.

### 3.1. *Selecting features*

Given the amino acid sequence of a peptide, its theoretical spectrum can be derived to include all ion types of interests including $a$, $b$, $c$, $x$, $y$, and $z$ ions and their variants (losing water and ammonium groups, isotopes, multiple charges). A simple algorithm is first applied to match each theoretical peak $p'$ with a closest experimental peak, with the preference to $b/y$ ions when there are multiple matches within the $m/z$ error tolerance.

Let $I$ denote the intensity of $p$ and $E$ denote the m/z error between $p$ and $p'$. Assuming the m/z error tolerance is $\Delta$, an experimental peak is a candidate to match if $\mid E \mid \leq \Delta$. Peak intensities in experimental spectra can vary drastically. We have observed that they can vary by a multiplicative factor of $10^6$. To minimize this problem, an empirical formula below is used to adjust the intensity for each candidate peak:

$$I^* = e^{-c\times(|E|/\Delta)^2} \times \sqrt{I} \tag{1}$$

where $c$ and $\Delta$ are empirically set to 3 and 0.5 dalton, respectively.

The following features are then extracted. These features are classified in to 4 groups.

(1) For each ion type, the sum of intensities (i.e. $I^*$ values) of all matched peaks of this type. The types we consider include $a$, $b$, $c$, $x$, $y$, $z$ ions, as well as all internal fragmentations, $b - NH_3$, $y - H_2O$.
(2) The weighted sum of intensities of all matched peaks. In other words, this is the weighted sum of all sums in Item 1. Each type of ions is assigned a weight. For CID (collision induced dissociation), most of the fragmentation produces

b-ions and y-ions. Therefore, higher weights (1.0) are given to $b$ and $y$ ions and lower weights (0.1) to other types of ions.

(3) The sums of products of the intensities for the complementary pairs of each type. These include: the sum of products of the intensities of all complementary $b/y$ ion pairs; the sum of products of the intensities of $y_i$ and $y_{i+1}$ pairs; the sum of products of the intensities $y_i$ and $y_i - H_2O$ pairs for all $i$; the sum of products of the intensities of $b_i$ and $b_i - NH_3$ pairs for all $i$, etc. For instance, the following formula is used to compute the the $b/y$ ion complementary pair intensities:

$$I_{by} = \sum_{i=1}^{n-1} I_b(i)^* \times I_y(n-i)^* \tag{2}$$

where $n$ is the peptide length.

(4) Average m/z error of the matched peaks for each ion type. The system error due to instrument calibration needs to be removed. Assume there are $n$ peaks in the ion series. Let $E_i$ be the error for each peak $p_i$ and $E_m$ be the mean of errors of the matched peaks, then the average error is adjusted as below:

$$E_{avg} = -\frac{\sum_{i=1}^{n} \mid E_i - E_m \mid}{n}. \tag{3}$$

Given an experimental spectrum and $n$ candidate peptides, a set of feature vectors $\{V_1, V_2, \cdots, V_n\}$ can be derived, each corresponding to one peptide. Let $V_i(j)$ be the value of $j$-th feature of $i$-th vector. Each feature value is normalized by

$$V_i(j)^* = \frac{V_i(j)}{\max\limits_{k=1,2,\cdots,n} \mid V_k(j) \mid} \tag{4}$$

According to the preceding formulation, each feature is a numerical value. It is expected that the correct peptide is more likely to have *high* feature values than incorrect ones. In practice some features are more distinguishing than others, due to the noises and missing ions. Thus it is necessary to find an appropriate weights for all the features to achieve the optimum discriminating capacity.

For each feature, given a training spectrum, the values for all candidate peptides are calculated, and then sorted in descending order. The percentile rank of the true peptide's value is recorded. Averaging over all training spectra, this feature's percentile ranking is obtained. Those features whose percentiles rank at top 5% most will be used to derive the final scoring function by a linear programming described in the next subsection.

### 3.2. *A linear programming formulation for the scoring function*

Given a spectrum and the peptide, the values of $l$ selected features form a vector $V = \langle v_1, v_2, \cdots, v_l \rangle$. In this work, the scoring function is formulated as a weighted sum of feature values. That is, we consider scoring functions of the form $S(V) = C \cdot V = \sum_{i=1}^{l} c_i \times v_i$, where $C = \langle c_1, c_2, \cdots, c_l \rangle$. Now the problem is to determine

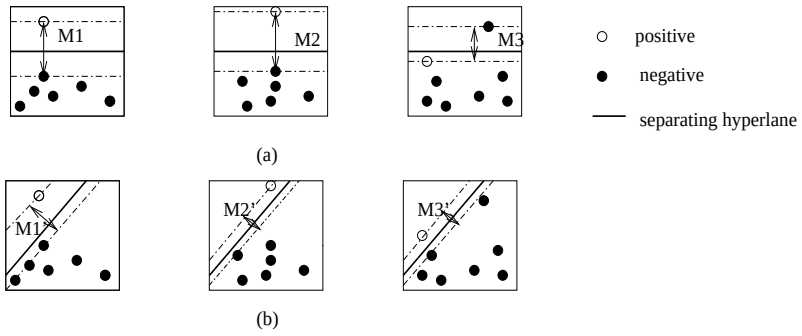


Fig. 2. An example of improving accuracy by bounding the functional margin. (a) Without the bounding, one sample is misidentified; (b) with bounded functional margin, all 3 samples are correctly identified.

values of $c_i$ to optimize the accuracy of identification. This is solved by a linear programming.

Assuming a sequence of experimental spectra $\langle s_1, s_2, \cdots, s_n \rangle$ is produced by peptides $\langle p_1, p_2, \cdots, p_n \rangle$, respectively. For each spectrum $s_i$, let $P_i$ be the feature vector for correct peptide $p_i$. The negative peptides are selected in a protein database by using the peptides with similar masses to $p_i$. Assume that the number of negative peptides for each spectrum is $K_1, K_2, \cdots, K_n$, respectively, and $N_{ij}$ is the $i$-th feature of the $j$-th negative peptide for $s_i$. The linear programming formulation is given below:

$$\begin{aligned} & \max \sum_{i=1}^{n} M_i \\ & \text{subject to} \\ c_i \geq 0 \quad & i = 1, 2, \cdots, l; \\ c_1 + c_2 + \cdots + c_l = 1; & \\ M_i \leq C \cdot (P_i - N_{ij}) \quad & j = 1, 2, \cdots, K_i, \ i = 1, \cdots, l; \\ M_i \leq \epsilon. & \end{aligned} \tag{5}$$

The geometrical interpretation of inner product of two vectors $X \cdot Y$ is the projection of $X$ onto $Y$ when $\| Y \| = 1$. In other words, it is the distance to a hyperplane $H$ which is perpendicular to $Y$. Thus the problem is equivalent to finding a good linear boundary separating hyperplane in the $\Re^l$ to identify positives and negatives. For $i$-th spectrum, the functional margin is $\min C \cdot (P_i - N_{ij})$. Intuitively, an ideal separating hyperplane leads to large margins for training samples. Nevertheless, maximizing sum of margins may damage the overall accuracy of identification. Fig. 2 (a) provides an example, where the third sample is not identified correctly if the

objective is to maximize the sum of functional margins.

To alleviate such problem, the forth constraint in Formula (5) is imposed to place a bound of functional margin distance. Fig. 2 (b) shows improved hyperplane for separation, where $M_i/M_i^{'}$, $i = 1, 2, 3$, are the functional margins for the individual samples, respectively. Noticeably Formula (5) works well for the $i$ such that the function margin, $\min(C \cdot (P_i - N_{ij}))$, is positive; but has no effect when the function margine is negative. However, when the training data are carefully selected and have acceptable quality, the function margin usually cannot have big negative values.

The coefficients are determined when the linear programming formulation is solved. As mentioned above, some samples cannot be recognized correctly, their functional margins are negative. As the objective goal is to maximize the sum of bounded functional margins, the overall identification accuracy might drop to offset some big negative margins. To further improve this situation, we used a heuristics to iteratively explore the proximity of the coefficients returned by LP solver. In each iteration, we adjusted one coefficient by a small step $\delta$ to improve 1) the identification accuracy or 2) the minimal functional margin of all samples without decreasing the accuracy.

Prototype PRIMA was implemented based on this formulation and the optimized coefficients.

## 4. Experimental results

We used three large third-party datasets to evaluate PRIMA. Dataset 1 contained 86 ion trap spectra from Richard Johnson.[a] These spectra served as training data. Dataset 2 contained 266 ion trap spectra obtained from a Finnigan LCQ Deca mass spectrometer [30], provided to us by Mark Cieliebak of ETH. Dataset 1 and 2 were searched against NCBInr database. Dataset 3 was a well-known dataset of 37,071 low quality ion trap spectra aimed at providing a standard test benchmark for researchers to compare their work with the SEQUEST program, given in [29]. This dataset was accompanied by a custom database. These spectra were produced by ion trap mass spectrometers of different resolutions and from different organizations, many not tryptic digested and many only tryptic digested at one end.

Since MASCOT and SEQUEST are the industrial standard and are most widely used, we compared PRIMA with them [b]. In our experiments, MASCOT online server at http://www.matrixscience.com/ was used for the experiments. Another open source software X! Tandem [18] is becoming popular recently, therefore its latest release (2005.06.01) was also downloaded from http://www.thegpm.org for testing and comparison. In particular, X! Tandem ran in two modes. In the first mode (default one), it only output the *valid* peptides whose scores were statistically sig-

[a]This dataset originally has about 144 spectra. Many of the spectra had large precursor mass discrepancies due to PTMs and these spectra are removed, with 86 left.

[b]For datasets 1 and 2, all the spectra were selected based on the criteria that their SEQUEST results are correct. Therefore, it is meaningless to compare with SEQUEST on these two datasets.

Table 1. Subset of selected features and their discriminating capacity. The second and third columns give the numbers of spectra where the corresponding feature ranks the correct peptides to top 5% and top 1 of all peptides with similar precursor mass, respectively.

| Feature | # of top 5% | # of No. 1 |
|---|---|---|
| sum of intensity for all ions | 85 | 73 |
| sum of intensity for y ions | 84 | 68 |
| sum of intensity product for complementary $b/y$ ions | 84 | 43 |
| sum of intensity for b ions | 74 | 10 |
| average m/z error for y ions | 56 | 5 |

Table 2. Training: Identification accuracy comparison between PRIMA, MASCOT and X! Tandem over dataset 1

| | ratio of No.1 | ratio of top 10 |
|---|---|---|
| PRIMA | 90.7% | 97.7% |
| MASCOT | 84.9% | 93.0% |
| X! Tandem (mode 1) | 55.6% | 58.1% |
| X! Tandem (mode 2) | 77.9% | 83.7% |

*Note*: In mode 1, X! Tandem did not return any peptides for 36 lower-quality spectra.

nificant, while it may not return any peptides for certain lower-quality spectra. However, in the second mode, X! Tandem returned all top ranked peptides for each spectrum.

In the training process, we identified the features used in the scoring functions. Table 1 displays the main features selected to form the scoring function, along with their discriminating capacity. With the selected features, the LP formulation in Section 3.2 is used to derive the linear scoring function. As observed by many prior researchers, for example in [6] and [24], $b/y$ ions are the most common and valid peaks for mass spec analysis for all types of instruments. Focusing on the features mainly related to $b/y$ ions makes the scoring function more instrument neutral.

After coefficients were determined, the scoring function was then applied to dataset 1 to assess its effectiveness. For each spectrum, the top ranked 10 peptides from PRIMA were output. As shown in Table 2, PRIMA outperformed both MASCOT and X! Tandem in identification accuracy.

PRIMA was then tested using datasets 2 and 3. Table 3 gives the performance of PRIMA, MASCOT and X! Tandem performance on dataset 2. It shows that PRIMA achieves better results than MASCOT and X! Tandem. For a closer look, Table 4 presents some peptides which were not correctly recognized either by PRIMA or MASCOT in the columns 2 and 3. However, for some of these peptides, correct sequence tags were identified and underlined in the table.

Dataset 3, from [29], provides a perfect benchmark for comparing PRIMA with

Table 3. Identification accuracy comparison between PRIMA and MASCOT, X! Tandem over dataset 2

| | ratio of No.1 | ratio of top 10 |
|---|---|---|
| PRIMA | 92.0% | 94.7% |
| MASCOT | 90.4% | 91.2% |
| X! Tandem (mode 1) | 67.4% | 72.4% |
| X! Tandem (mode 2) | 77.9% | 83.7% |

*Note*: In mode 1, X! Tandem did not return any peptides for 66 spectra.

Table 4. Peptides in dataset 2 incorrectly identified by either PRIMA or MASCOT.

| Correct peptides | PRIMA | MASCOT |
|---|---|---|
| KQTALVELLK | QEDGPDMHSK | (*) |
| DLGEQHFK | DLGEEHFK | (*) |
| KVPQVSTPTLVEVSR | KVPEVSTPTLVEVSR | (*) |
| FKDLGEEHFK | (*) | AGYVLELLDKK |
| KTGQAPGFTYTDANKNK | (*) | KLSNLIGLLWETDPNK |
| TGQAPGFTYTDANKNK | VQMDDAMVIHADTIR | (*) |
| HPYFYAPELLYYANK | CDLFKTEEYCLVGLTR | (*) |
| INPDKIKDVIGK | (*) | LFGHLTKIVAK |
| HPYFYAPELLYYANK | YPHMFINHNQQVSFK | (*) |
| DGISALQMDIK | DGISTGCSPARK | (*) |
| PSEGETLIAR | (*) | VSEGEFNHR |
| PGQDFFPLTVNYQER | (*) | IAQIIGPVLDVFFPPGK |
| PSEGETLIAR | (*) | AIEGSSGPKAR |
| DGISALQMDIK | KRSGKEEDNK | (*) |
| EIMQVALNQAK | (*) | TKTELAVEIIK |
| PSEGETLIAR | (*) | VSEGEFNHR |
| YSEIYYPTVPVK | LDNVEEGKENWK | NPETEWPPFLTK |
| PGQDFFPLTVNYQER | (*) | IAQIIGPVLDVFFPPGK |
| PGQDFFPLTVNYQER | (*) | VQLAGSHILEALRLHR |
| PSEGETLIAR | (*) | VSEGEFNHR |
| VISWYDNEWGYSNR | (*) | LVSWYDNEWGYSNR |

*Note*: An asteroid (*) indicates that the peptide was correctly identified.

SEQUEST. This dataset contains 37,071 spectra of low quality, measured from tryptic digestions of mixtures of 18 proteins. Using a specialized protein database (human plus the 18 proteins plus common contaminants), SEQUEST has correctly identified 2784 spectra. Among the 2784 spectra, which were corrected identified by SEQUEST, 2057 are fully tryptic, 646 are semi tryptic (one end of the peptide is cut at R/K), and 81 are non-tryptic. Because MASCOT online server and X! Tandem did not have an option to specialize on peptides that were only tryptic digested at one end, it was impossible to make a fair comparison between them and PRIMA. Therefore, we only used dataset 3 to compare SEQUEST with PRIMA. As summarized in Table 5, PRIMA correctly identified 3,090 spectra with highest scores, and 4,585 spectra with correct peptides ranked among top 10. Among the SEQUEST's

Table 5. Identification accuracy comparison between SEQUEST and PRIMA over dataset 3

| | Number of No. 1 | Number of top 10 |
|---|---|---|
| SEQUEST | 2,784 | Unknown |
| PRIMA | 3,090 | 4,585 |

2,784 correct spectra, PRIMA has correctly identified 2,295 peptides as No. 1 and 2,497 of them as top ten. PRIMA also correctly identified extra 795 spectra with the highest scores and 2,088 spectra with top 10 scores from the remaining 34,287 spectra that have failed by using SEQUEST.

We have further conducted limited test over higher quality data to evaluate the robustness of PRIMA. 17 spectra generated by Q-TOF were used as dataset 4 and they were provided by Bioinformatics Solutions Inc. These spectra are more accurate in m/z measurement and subject to less noise. Searching against NCBInr protein database, both PRIMA and MASCOT online server correctly identified 14 spectra, and all 17 positives were among top 10 ranked peptides. For this dataset, X! Tandem did not output any peptides for 9 spectra in first mode, and correctly recognized all the 8 positives for the rest. When it ran in second mode, X! Tandem successfully recognized 12 positives.

Besides the accuracy, a practical concern is the reliability of the scores. In general, the scores of positives in PRIMA were significantly high. For instance, given one spectrum in dataset 4, PRIMA ranked the positive peptide as No.1 with score 0.82, the distribution of scores for this spectrum is demonstrated in Fig. 3.

To further study the reliability of scores, we define confidence index for each spectrum as

$$CI = \frac{s_1 - s_2}{s_2 - s_3} \tag{6}$$

where $s_1, s_2, s_3$ are the highest, second and third highest scores of candidate peptides returned by PRIMA.

Fig. 4 (a) shows the scores and their confidence indices for all datasets 1, 2 and 4. Generally, for true positives, PRIMA returns high scores with high confidence; whereas either scores or confidence indices are low for false positives. Fig. 4 (b) depicts the ROC curve for the trade-off between sensitivity and specificity. The area under ROC curve was 0.96, therefore it provided low ratios for both false positive and false negative in classification.

The complete list of all results for all spectra can be found at http://monod.uwaterloo.ca/∼jianliu.

## 5. Discussions and future work

Our goal of this research was to design a robust scoring function and a prototype system to deal with the low quality data that flood the proteomics industry and

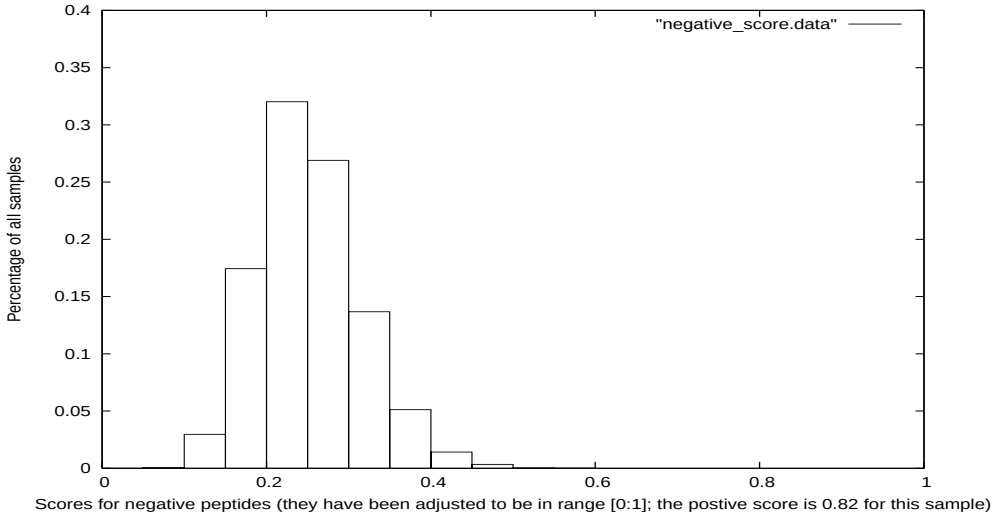


Fig. 3. Distribution of negative peptide scores for a sample spectrum

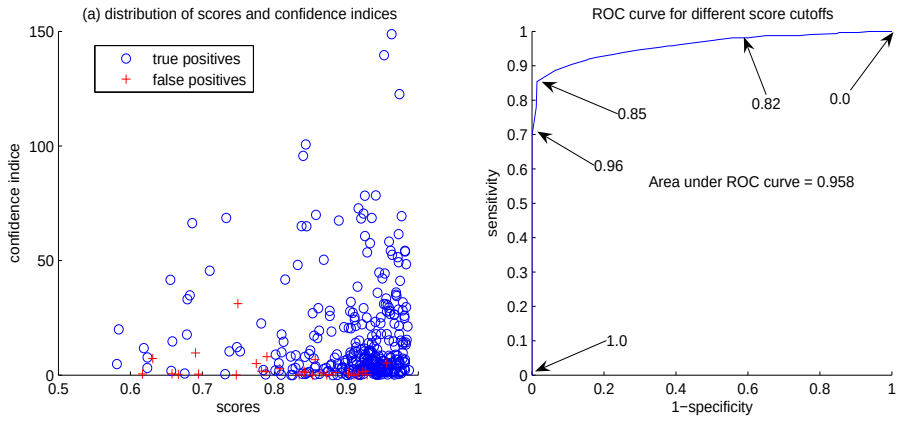


Fig. 4. Performance of PRIMA for datasets 1, 2, and 4. (a) Distribution of scores and confidence indices. (b) ROC curve: Sensitivity vs. specificity for different score cutoffs.

mass spectrometry research consortiums. We have presented a technique to construct a linear scoring function for MS/MS spectrum interpretation via a database. Some empirical values and formulas were used to normalize spectra and assigning weights to ions. Tests with over 30,000 spectra, produced from different centers, show that our prototype system PRIMA outperforms the mainstream software tools MASCOT, SEQUEST and X! Tandem on low quality ion trap data. This work also

provides a framework to effectively construct such a scoring function. For example, in contrast to collision induced dissociation, spectra generated from electron transfer dissociation have different patterns of ions. Therefore, it is necessary to reselect features and determine the coefficients during the training process. It was also noticed that a good selection of features can reduce the candidate peptide sets to manageable sizes, otherwise it will be computationally infeasible and fail conventional classification techniques like SVMs.

Further research is underway to deal with the post translational modifications, increase search speed, and effectively combine *de novo* sequencing with database search methods.

## 6. Acknowledgements

This work was partially supported by an NSERC grant OGP0046506 and CITO's Champion of Innovation Program, the Killam Fellowship, and the Canada Research Chair Program. The authors would like to thank Richard Johnson for providing dataset 1, Mark Cieliebak, Franz Roos and Sacha Baginsky for providing dataset 2, and L. DeSouza, Gilles Lajoie, and Michael K.W. Siu for their help on various aspects of mass spectrometry. The authors are also grateful to the reviewers for their constructive comments on the initial manuscript. The linear programs was solved by software package *lp_solve*, version 2.0, from [31].

## References

1. Kinter M and Sherman NE. *Protein sequencing and identification using tandem mass spectrometry*, Wiley-Interscience Publisher, 2000.
2. Paterson S, Data analysis – the Achilles heel of proteomics, *Nature Biotechnology* **21**:221-222, 2003.
3. Bartels C, Fast algorithms for peptide sequencing by mass septrometry, *Biomedical and Environmental Mass Spectrometry* **19**:363-368, 1990
4. Taylor JA, Johnson RS, Implementation and Uses of Automated de Novo Peptide Sequencing by Tandem Mass Spectrometry, *Analytical Chemistry* **73**:2594-2604, 2001.
5. Ma B, Zhang K, *et al.*, PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry. *Rapid Communication in Mass Spectrometry* **17(20)**:2337-2342, 2003.
6. Ma B, Zhang K, Liang C. An efficient algorithm for peptide *de novo* sequencing from MS/MS spectrum, *Proc. Conference on Combinatorial Pattern Matching* pp. 266-278, 2003.
7. Danck V, Addona T, Clauser K, Vath J, and Pevzner P, *De novo* protein sequencing via tandem mass-spectrometry. *Journal of Computational Biology* **6**:327-341, 1999.
8. Chen T, Kao MY, Tepel M,Rush J, Church G, A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry, *Journal of Computational Biology* **8(3)**:325-337, 2001.
9. Perkins DN, Pappin JC, Creasy DM, Cottrell JS, Probability-based protein identification by searching database using mass spectrometry data, *Electrophoresis* **20**:3551-3567, 1999.
10. Eng JK, McCormack AL, Yates, JR, An approach to correlate tandem mass spectral

data of peptides with amino acid sequences in a protein database, *Journal of American Society Mass Spectrometry* **5**:976-989, 1994.
11. Sadygov R, Liu H, Yates JR, Statistical models for protein validation using mass spectral data and protein amino acid sequence databases, *Analytical Chemistry* **76**:1664-1671, 2004.
12. Colinge J. *et al*., OLAV: Towards high throughput tandem mass spectrometry data identification. *Proteomics* **3**:1454–1463, 2003.
13. Craig R, Beavis R, A method for reducing the time required to match protein sequences with tandem mass spectra, *Rapid Communications in Mass Spectrometry* **17**:2310-2316, 2003.
14. Eddes JS, Kapp EA, *et al*., CHOMPER: A bioinformatics tool for rapid validation of tandem mass spectrometry search results associated with high-throughput proteomic strategies, *Proteomics* **2**:1097-1103, 2002.
15. Falkner J, Andrews P, Fast tandem mass spectra-based protein identification regardless of the number of spectra or potential modifications examined, *Bioinformatics*, **21**2177-2874, 2005.
16. Kapp A, *et al*., Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation, *Analytical Chemistry* **75**:6251-6264, 2003.
17. Hernandez P, Gras R, Frey J, and Appel RD, Popitam: towards new heuristic strategies to improve protein identification from tandem mass spectrometry data, *Proteomics* **3**:870-878, 2003.
18. Fenyo D and Beavis RC, A method for assessing the statistical significance of mass Spectrometry-based protein identifications using general scoring Schemes, *Analytical Chemistry* **75**:768-774, 2003.
19. Mann M, and Wilm M, Error-tolerant identification of peptides in sequence databases by peptide sequence tags, *Analytical Chemistry* **66**:4390-4399, 1994.
20. Fenyo D, Qin J, Chait B, Protein identication using mass spectrometric information, *Electrophoresis* **19**:998-1005, 1998.
21. Qin J, Fenyo D, Zhao Y, *et al*., A strategy for rapid, high-confidence protein identification, *Analytical Chemistry* **69**:3995-4001, 1997.
22. Pevzner PA, Dancik V, and Tang CL, Mutation-tolerant protein identification by mass-spectrometry, *Proc. of RECOMB*, pp. 231-236, 2002.
23. Yates JR, Eng JK, McCormack AL, Schieltz D, Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database, *Analytical Chemistry* **67**:1426-1436, 1995.
24. Bafna V, Edwards N, SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database, *Bioinformatics* **17**:S13-S21, 2001.
25. Wan Y and Chen T, A hidden Markov model based scoring function for tandem mass spectrometry. *Proc. of RECOMB*, pp.342-356, 2005.
26. Anderson DC, Li W, Payan DG and Noble WS, A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *Journal of Proteome Research* **2**:137-146, 2003.
27. Chamrad D, Evaluation of algorithms for protein identification form sequence databases using mass spectrometry, *Proteomics* **4**:619-628, 2004.
28. Snyder AP, *Interpreting protein mass spectra: a comprehensive resource*, Oxford University Press, 2002.
29. Keller A, Purvine S, *et al*., Experimental protein mixture for validating tandem mass spectral analysis, *OMICS: A Journal of Integrative Biology* **6(2)**:207-212, 2002.

30. Grossmann J, *Protein identification using mass spectrometry: development of an approach for automated de novo sequencing*, Master thesis, ETH Zurich, Department of Biology, 2003.
31. Linear Programming Solver, http://www.cs.sunysb.edu/∼algorithm/implement/lpsolve/implement.shtml, 2003.

**Jian Liu** received his PhD degree from Department of Computer Science, University of Minnesota, USA, in 2003. Before that, he also received his bachelor and master degrees, both in computer science, from Tsinghua University, Beijing, China in 1994 and 1996, respectively.

From 2003 to 2004, he was a postdoctoral research fellow with School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada. He is currently working on Genome Quebec project as a postdoctoral research fellow at Department of Biomedical Engineering, McGill University, Montreal, Quebec, Canada. His major research interests are machine learning in bioinformatics, peptide/protein identification through mass spectra analysis.

**Bin Ma** is a Canada Research Chair in Bioinformatics and an Associate Professor in the Department of Computer Science at the University of Western Ontario. He received his Ph.D. degree in Peking University in 1999. He is a recipient of Ontario Premier's Research Excellence award in 2003 for his research in bioinformatics. He is a coauthor of several well-known bioinformatics software programs, including PatternHunter and PEAKS.

**Ming Li** is a Canada Research Chair in Bioinformatics and professor of Computer Science at the University of Waterloo. He is a recipient of Canada's E.W.R. Steacie Fellowship Award in 1996, and the 2001 Killam Fellowship. Together with Paul Vitanyi they have pioneered the applications of Kolmogorov complexity and co-authored the book *An Introduction to Kolmogorov Complexity and Its Applications* (Springer-Verlag, 1993, 2nd Edition, 1997). He is a co-managing editor of Journal of Bioinformatics and Computational Biology. He is an associate editor-in-chief of Journal of Computer Science and Technology.