PRIMA: PEPTIDE ROBUST IDENTIFICATION FROM MS/MS SPECTRA

JIAN LIU *and BIN MA †and MING LI ‡

In proteomics, tandem mass spectrometry is the key technology for protein identification from the cells. However, partially due to the deficiency of peptide identification software, over half 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 end data from low quality but cheaper devices such as the ion trap instruments. In order to deal with the noisy and low quality data, this paper develops a systematic approach to construct a robust linear scoring function, whose coefficients are determined by a linear program. A prototype, PRIMA, is implemented. When exhaustively tested with large benchmarks of varying qualities, PRIMA consistently outperforms the commonly used software MASCOT and SEQUEST with higher accuracy.

1. Introduction

Proteomics aims at understanding proteins expressed in cells at different levels, during different times, and in different forms. These questions are critical steps connecting the genomes to drug discovery and modern medical advances. Mass spectrometers are currently the predominant tool to accomplish some of the primary goals of proteomics: (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. 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. There are two approaches to do 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 powerful database method. Using a

*School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada. This work is partially supported by an NSERC grant OGP0046506 and CITO's Champion of Innovation Program. Email: jianliu@monod.uwaterloo.ca

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

‡School of Computer Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada. This work is partially supported by an NSERC grant OGP0046506, CITO's Champion of Innovation Program, the Killam Fellowship, and the Canada Research Chair Program. Email: mli@uwaterloo.ca

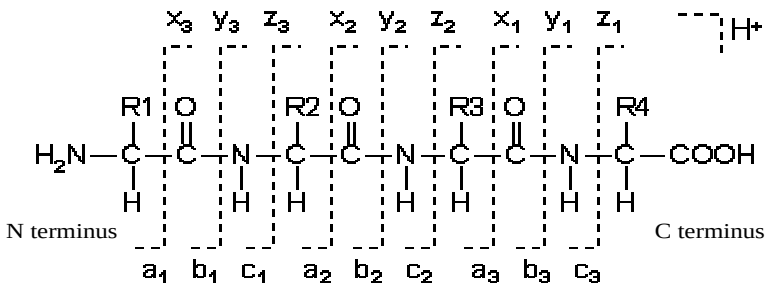

Figure 1. Different ions produced by peptide fragmentation. a/x, b/y, c/z are complementary ions, respectively. b/y ions are the most common ones.

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 supremacy of PRIMA over both MASCOT and SEQUEST on large spectrum benchmarks.

2. Background and related work

Tandem mass spectrometry (MS/MS) is currently the method of choice for high throughput identification of proteins due to its speed and high sensitivity. In such an approach, a protein is digested or chemically cleaved into many peptides. These peptides are fragmented and ionized to carry one or more units of charge. Peptides typically break at the peptide bonds, forming b-ions and y-ions, as shown in Fig. 1. The ions are then separated according their mass/charge ratios in the mass analyzer. Finally ions are collected by ion detector to produce mass spectra. Each spectrum includes a sequence of peaks indicating the mass/charge ratios and abundance of ions.

Software tools, database search method or *de novo* sequencing, are finally applied to interpret each MS/MS spectrum to infer the peptide sequence, and then the protein which contains the peptide.

De novo sequencing method determines the peptide sequence solely from the experimental spectra without using databases.[2] This method is useful when the protein is not in the database. The mainstream *de novo* sequencing software include program packages from mass spectrometry vendors (MassLynx, BioAnalyst, denovoX, etc), the free program Lutefisk[16] and commercial programs PEAKS[13] and SpectrumMill. The basic *de novo* sequencing dynamic programming techniques were first introduced by Dancik $et\ al.$[7] and Chen $et\ al.$[4]

The database search method is more powerful, but it depends on the fact that the target protein sequence is in the database. Given an experimental spectrum S, 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 are MASCOT [14] and SEQUEST.[9,15] SEQUEST compares the theoretical spectra against real 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. Recent research have been reviewed recently by Chamrad.[3] Improvements to these programs are claimed with various criteria: fewer false positives[5] less time[6], validation[8], and new approaches.[10]

This paper focuses on the database method in order to obtain a robust solution to the low quality spectra. We aim at developing a theoretically sound and practically feasible approach, avoiding currently infeasible problems such as computing the probability of each spectrum given a peptide.[1]

At the heart of all search methods is a scoring mechanism to rank the candidate peptides. Constructing a good scoring function is tricky. The fragmentation of the peptides is determined by their physiochemical characteristics as well as many other factors, resulting many problems listed below.

- Internal fragmentations. A peptide may be broken more than once.
- Some ions may be missing in the experimental spectra. The intensity of same ion may vary greatly for different runs.
- Isotopes. For example C^{13} adds one dalton. Furthermore if the ion has charge 2, then the distance is only 0.5 on the m/z axis.
- Other ions: a-ions, c-ions, x-ions, z-ions. They appear at different rates with different types of mass spectrometers.
- Each N-terminal ion (a-, b-, c-ions) can lose an ammonium group (NH_3, -17 daltons); each C-terminal ion (x-, y-, z-ions) can lose a water (H_2O, -18 daltons).
- Multiply charged ions. Noise peaks that correspond to nothing.

As the result, the spectra generated from mass spectrometers often have little resemblance of the corresponding theoretical spectra. For example, Fig. 2 illustrates an experimental and a theoretical b/y ion peak spectrum for peptide LVTDLTK. To make the matter worse, each type of mass spectrometer has its own sensitivity and resolutions, the scoring function often needs to be adjusted to achieve the best performance.[11]

Given a spectrum, we can find a list of candidate peptides from the protein database whose masses are within a predefined mass error tolerance to the precursor ion mass (i.e. the peptide mass measured by the mass spectrometer) of the spectrum. For a large database, this list can be as large as 100,000 tryptic peptides, using ± 2 dalton error tolerance. A scoring function is then needed to find the correct peptide.

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.

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 and their variants (losing water and ammonium groups, isotopes, multiple charges). A simple algorithm is

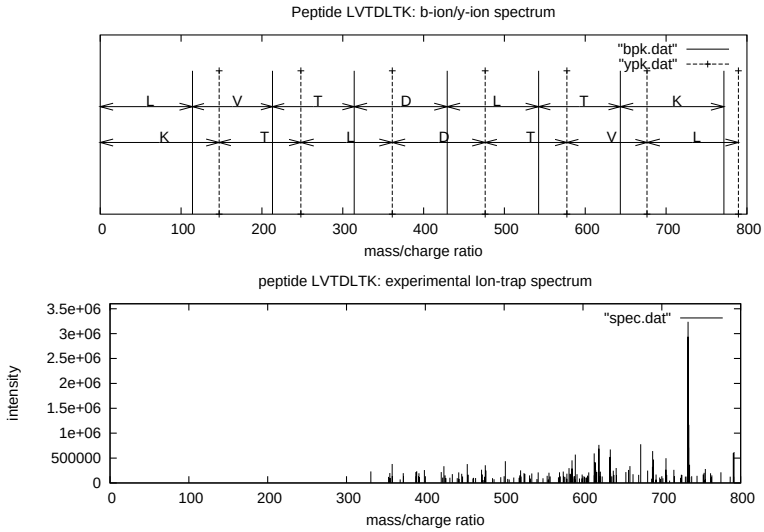

Figure 2. The theoretical b/y ion spectrum, top, and an experimental spectrum from [11], bottom, for peptide LVTDLTK.

first applied to match each theoretical peak p' with a closeby experimental peak, with the preference to b-/y- ions when there are multiple matches within the mass error threshold.

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 Δ, 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:

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

where c and Δ are empirically set to 3 and 0.5 dalton, respectively.

The following features are then extracted to indicate the similarity between theoretical and experimental spectrum. These features are classified into 4 groups:

(1) For each ion type, the sum of intensities, 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, I^* values, of all matched peaks. I.e., this is the weighted sum of all sums in Item 1. Each type of ions is assigned a weight. 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:

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

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:

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

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 are used to derive the final scoring function by a linear program described in the next section.

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 $S(V) = C \cdot V = \sum_{i=1}^{l} c_i * v_i$, where $C = \langle c_1, c_2, \cdots, c_l \rangle$. Now the problem is to determine 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 feature vector of the j-th negative peptide for s_i. The linear programming formulation is given below:

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

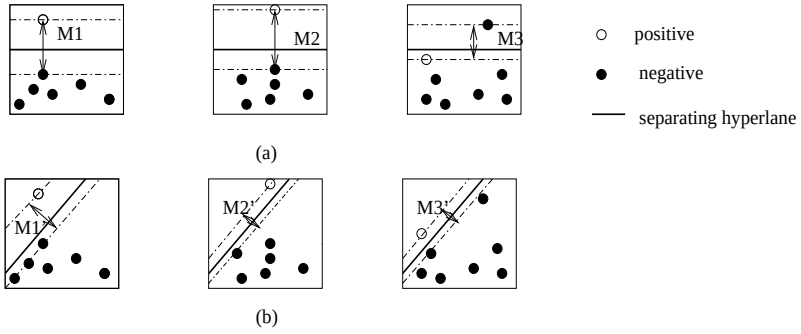

Figure 3. An example of improving accuracy by bounding the functional margin. (a) Without the bounding, one sample is misidentified; (b) with bounded functional margin, all 3 samples are correctly identified.

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 $\max C \cdot (P_i - N_{ij})$. 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. 3 (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. 3 (b) shows improved hyperplane for separation, where $M_i/M_i^{'}$, $i = 1, 2, 3$, are the functional margins for the individual samples, respectively.

The coefficients are determined when the linear programming formulation is solved. 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 iden-

tification accuracy might drop to offset the big negative margins. To further improve the situation, we use a heuristics to iteratively explore the proximity of the coefficients returned by LP solver. In each iteration, we adjust one coefficient by a small step δ to improve 1) the identification accuracy or 2) the minimal functional margin of all samples without decreasing the accuracy.

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

4. Experimental results

We used three large third-party datasets to evaluate PRIMA. Dataset 1 contains 86 ion trap spectra from Richard Johnson.[a] Dataset 2 contains 266 ion trap spectra obtained from a Finnigan LCQ Deca mass spectrometer[12], provided to us by Mark Cieliebak of ETH. Dataset 3 is 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 by Keller $et\ al.$[11] 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, are recognized as the leading database search programs and are most widely used, we compare PRIMA with these two programs. In our experiments, MASCOT online server at http://www.matrixscience.com/ is used for dataset 1 and 2. For dataset 3, as MASCOT online server does not accept external databases and does not have an option to specialize on peptides that are only tryptic digested at one end, it was impossible to make a fair comparison with PRIMA. We were only able to use dataset 3 to compare SEQUEST with PRIMA. On the other hand, for datasets 1 and 2, although we know partial SEQUEST results, it was impossible to make a fair comparison. Thus, dataset 1 is used for training. Dataset 2 is used to compare MASCOT with PRIMA, and dataset 3 is used to compare SEQUEST with PRIMA. In our experiment, ϵ of in Formula 5 was set to 1.0001 empirically; precursor error tolerance was set to $\pm 1.0/2.0$ daltons for training and testing, respectively.

In the training process, we identified the features used in the scoring functions. As observed by many prior researchers, 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. Table 1 displays the features selected to form the scoring function, along with their discriminating capacity. For each feature, the second column and the third column provide numbers of spectra for which the positive peptide feature value is ranked among top 5% and as No.1 among all candidate peptide feature values. With the selected features, the LP formulation in Section 3.2 is then used to derive the linear scoring function.

After coefficients are determined, the scoring function was then applied to dataset 1 to

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

Table 1. Training with dataset 1 (86 spectra, NCBI NR protein database): Subset of selected features and discriminating capacity of each feature

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 and MASCOT, both using NCBI non-redundant protein database.

	ratio of No.1	ratio of top 10
PRIMA	90.7%	97.7%
MASCOT	84.9%	93.0%

assess its effectiveness. For each spectrum, the top ranked 10 peptides from PRIMA were output. Table 2 provides a comparison between PRIMA and MASCOT.

PRIMA was then tested using datasets 2 and 3. Table 3 gives the PRIMA and MASCOT performance on dataset 2. It shows that PRIMA achieves better results than MASCOT. For a close look, Table 4 presents some peptides which are not correctly recognized either by PRIMA or MASCOT. In the columns 2 and 3, an asteroid (*) indicates that the peptide is correctly identified.

Table 3. Identification accuracy comparison between PRIMA and MASCOT, dataset 2, 266 spectra, both using NCBI non-redundant protein database.

	ratio of No.1	ratio of top 10
PRIMA	92.0%	94.7%
MASCOT	90.4%	91.2%

Dataset 3, provides a perfect benchmark for comparing PRIMA with SEQUEST. This dataset contains 37,071 spectra. According to Keller $et\ al.$[11] 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. 125 of them are charge 1, 1649 of them are charge 2 and 1010 of them are charge 3. Among the rest of 34287 spectra, 379 of them are charge 1, 16856 of them are charge 2 and 17052 of them are charge 3. Among the charge 2 and 3 spectra, there are 15435 duplicates, That is, these spectra have been saved in both charge 2 and 3 status.

After removing duplicates, PRIMA correctly identifies 3,090 spectra, with highest scores, and 4,585 spectra with correct peptides ranked among top 10. These are summarized in Table 5. Among the SEQUEST's 2,784 correct spectra, PRIMA has correctly identified 2,295 of them with the highest scores and 2,497 of them as top ten. Note that among the 2,784 spectra, 72 spectra have precursor mass error beyond PRIMA's precursor

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

error tolerance ± 2.0 daltons and 81 are non-tryptic, hence these 153 spectra are automatically not identifiable by PRIMA, *a priori*. PRIMA correctly identified extra 795 spectra with the highest scores and 2,088 spectra with top 10 scores, duplicates removed, from the remaining 34,287 spectra that have failed SEQUEST.

Table 5. Identification accuracy comparison between SEQUEST and PRIMA, dataset 3, 37,071 spectra both using the database given in 11.

	Total number of spectra	Number of correct	Number of top 10
SEQUEST	37,071	2,784	Unknown
PRIMA	37,071	3,090	4,585

A complete result list for all spectra can be found at http://monod.uwaterloo.ca/∼jianliu.

5. Conclusions and future work

Our goal of this research is to design a robust scoring function and a prototype system to deal with the low quality data that flood the proteomics industry and mass spectrometry research consortiums. We have presented a technique to construct a linear scoring function for MS/MS spectrum interpretation via a database. Tests with over 30,000 spectra, produced from three different centers, show that our prototype system PRIMA consistently outperforms the mainstream software tools MASCOT and SEQUEST on low quality ion trap data. This work also provides a framework to effectively construct such a scoring

function.

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

6. Acknowledgements

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 linear programs was solved by software package lp_solve, version 2.0, downloaded from ftp://ftp.es.ele.tue.nl/pub/lp_solve.

References

1. V. Bafna and N. Edwards. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics*, 17(2001), S13-S21.
2. C. Bartels. Fast algorithms for peptide sequencing by mass septrometry. *Biomedical and Environmental Mass Spectrometry*, 19(1990), 363-368.
3. D. Chamrad. Evaluation of algorithms for protein identification form sequence databases using mass spectrometry. *Proteomics* 4(2004), 619-628.
4. T. Chen, M-Y. Kao, $et\ al$. A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology* 8:3(2001), 325-337.
5. J. Colinge, A. Masselot, $et\ al$. Olav: Towards high throughput tandem mass spectrometry data identification. *Proteomics*, 3(2003), 1454-1463.
6. R. Craig and R. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Communications in Mass Spectrometry* 17(2003), 2310-2316.
7. V. Danck, T. Addona, K. Clauser, J. Vath, and P. Pevzner. *De novo* protein sequencing via tandem mass-spectrometry. *Journal of Computational Biology* 6(1999), 327-341.
8. J.S. Eddes, E.A. Kapp, S.F. Frecklington, $et\ al$. CHOMPER: A bioinformatic tool for rapid validation of tandem mass spectrometry search resutls associated with high-throughput proteomic strategies. *Proteomics* 2(2002), 1097-1103.
9. J.K. Eng, A.L. McCormack, and J.R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of Ammerican Society Mass Spectrometry*, 5(1994), 976-989.
10. E. A. Kapp, F. Schutz, $et\ al$. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Analytical Chemistry* 75(2003), 6251-6264.
11. A. Keller, S. Purvine, $et\ al$. Experimental protein mixture for validating tandem mass spectral analysis. *OMICS: A Journal of Integrative Biology*, 6:2(2002), 207-212.
12. J. Grossmann. *Protein identification using mass Spectrometry: development of an approach for automated de novo sequencing*, Master thesis, ETH Zurich, Department of Biology, 2003.
13. B. Ma, K. Zhang, C. Liang. An efficient algorithm for peptide *de novo* sequencing from MS/MS spectrum, *Proc. Conference on Combinatorial Pattern Matching* 2003, 266-278.
14. D.N. Perkins, J.C. Pappin, D.M. Creasy, and J.S. Cottrell. Probability-based protein identification by searching database using mass spectrometry data. *Electrophoresis* 20(1999), 3551-3567.
15. R. Sadygov, H. Liu, J.R. Yates. Statistical models for protein validation using mass spectral data and protein amino acid sequence databases. *Analytical Chemistry*, (76)2004, 1664-1671.
16. J.A. Tayor, R.S. Johnson. Sequence database searches via *de novo* peptide sequencing by mass spectrometry. *Rapid Communications Mass Spectrometry*, 11(1997), 1067-1075.