

Separation of Ion Types in Tandem Mass Spectrometry Data Interpretation -- A Graph-Theoretic Approach (Extended Abstract)

Bo Yan^{1 2#}, Chongle Pan^{2 4#}, Victor N Olman^{1 2}, Robert L Hettich³, Ying Xu^{1 2*}

¹ Department of Biochemical and Molecular Biology, University of Georgia, GA, USA

² Computational Biology Institute, ³ Chemical Sciences Division, Oak Ridge National Laboratory, TN, USA

⁴ Genome Science and Technology Graduate School, University of Tennessee, TN, USA

[#]Both authors contributed equally to this work

^{*}Corresponding E-mail: xyn@bmb.uga.edu

Abstract

Mass spectrometry is one of the most popular analytical techniques for identification of individual proteins in a protein mixture, one of the basic problems in proteomics. It identifies a protein through identifying its unique mass spectral pattern. While the problem is theoretically solvable, it remains a challenging problem computationally. One of the key challenges comes from the difficulty in distinguishing the N- and C-terminus ions, mostly b- and y-ions respectively. In this paper, we present a graph algorithm for solving the problem of separating b- from y-ions in a set of mass spectra. We represent each spectral peak as a node and consider two types of edges: a type-1 edge connects two peaks possibly of the same ion types and a type-2 edge connects two peaks possibly of different ion types, predicted based on local information. The ion-separation problem is then formulated and solved as a graph partition problem, which is to partition the graph into three subgraphs, namely b-, y-ions and others respectively, so to maximize the total weight of type-1 edges while minimizing the total weight of type-2 edges within each subgraph. We have developed a dynamic programming algorithm for rigorously solving this graph partition problem and implemented it as a computer program PRIME. We have tested PRIME on 18 data sets of high accurate FT-ICR tandem mass spectra and found that it achieved ~90% accuracy for separation of b- and y-ions.

1. Introduction

Tandem mass spectrometry has become a dominant proteomics technique because of its ability to identify proteins in a high-throughput manner [1]. In a typical LC/MS/MS experiment, a protein mixture of interest is digested by protease into peptides and then separated by high performance liquid chromatography (HPLC). When eluted from HPLC column, peptides are transformed to gas-phase positively-charged ions by electrospray and then introduced into mass spectrometer in batches. After measuring the mass/charge ratio (m/z) of all ions, mass spectrometry can precisely isolate each peptide by its m/z and fragment the peptide through collisional-induced dissociation (CID) into two complementary sets of pieces, namely N- and C-terminus ions respectively. The resultant fragments from this peptide are then measured for their m/z ratios. This process involves two sequential measurements of m/z for a peptide, thus called tandem mass spectrometry (MS/MS) experiment. Modern mass spectrometers can acquire thousands of high-resolution MS/MS spectra per day. Interpretation of such high-throughput mass spectral data, in a reliable and efficient manner, represents a highly challenging computational problem.

There are two popular approaches to interpretation of tandem mass spectra for protein identification. The *database search* method compares experimental tandem spectra with theoretical tandem mass spectra of each peptide derived from protein sequence databases, and reports the best match or matches [2-6], assuming that the query peptides exist in the protein sequence database. This approach is highly effective and has been used successfully in several proteomics projects on organisms with well-studied genome [7-9]. However, it is not applicable to situations where a target sequence is not in the searching protein database.

This can happen for a number of reasons, including novel proteins, protein mutations, post-translational modifications, and DNA sequencing errors.

A second approach for peptide identification is called *de novo* sequencing, which attempts to derive a protein sequence directly from tandem mass spectra [10-13]. Theoretically, full-length peptide sequence could be derived from an ideal tandem mass spectrum through computing mass differences between adjacent fragment ions' masses of the same ion type, in a complete series of fragment ions. *De novo* methodology usually employs a graph theory in which tandem mass spectrum is typically represented as a graph. Each spectral peak is represented as a vertex and a pair of spectral peaks that differ by precisely one amino acid in mass, represented as an edge (we call it type-1 edge). The partial sequence of target protein is then predicted through finding one or a set of longest directed paths in the spectrum graph. In these methods, no special attempt was made to distinguish ion types and solely the information of type-1 edges was used. However as shown in Fig.1, type-1 edge may also arise from two different type ions by coincidence. Since a tandem mass spectrum generally contains various ions of different types, such kind of misconnection decreases the accuracy of *de novo* sequencing.

In this paper, we treat the problem of ion type identification separately from the problem of *de novo* sequencing. We present a novel graph-theoretic approach for identification of ion types in a set of high quality FT-ICR MS/MS spectra. Since the majority of MS/MS spectral peaks are either b- or y-ions [11, 14], our algorithm attempts to separate a set of MS/MS peaks into (a) b-ions and their variants (i.e., loss of water or ammonia), (b) y-ions and their variants, and (c) the other ion types. We use a spectrum graph to represent a given set of spectral peaks in a similar fashion to some of the previous works [10-13]. The main difference in our graph representation is that we consider two types of edges, one representing the connection between a pair of ions suspected to be of the same type (type-1 edge) and the other representing the connection between a pair of ions suspected to be of different types (type-2 edge), based on the observations: the mass difference between any two ions of the same type must be equal to the total mass of some amino acids; and if the mass difference is not equal to the mass of any composition of amino acids, it must be from different type ions. Edge weights are assigned based on the estimated probabilities of whether the edges truly connect ions of the same or different types of ions. We formulate the ion-type identification problem as a graph partition problem, which is to partition the graph into three subgraphs, namely B, Y and U respectively, so to maximize the

total weight of type-1 edges while minimizing the total weight of type-2 edges in the each subgraphs. This problem is rigorously solved efficiently using a dynamic programming algorithm. The algorithm runs in $O(\sum_{i=1}^L 3^{|S_{i-1}|+|S_i|})$ time in the worst case, where i is the distance from the root in a breadth-first tree (*BFT*) of the spectrum graph, L is the depth of the *BFT*, and $|S_i|$ is the number of vertices on the i -th level of the *BFT*. On a typical data set with 40 spectral peaks, our identification program PRIME (PaRtition of Ion types of tandem Mass speCtra) takes a few seconds to find the globally optimal partition on a Dell Workstation with Pentium 4 (2.1 GHz).

2. Methodology

2.1. Graph representation and problem formulation

Let $S = \{s_1, s_2, \dots, s_k\}$ be a set tandem mass spectral data with k peaks $s_j = \{M_j, I_j\}$, where M_j and I_j denote the neutral mass and intensity of peak s_j . We define zero mass peak as $s_0 = \{0, I_{k+1}\}$ and parent mass peak as $s_{k+1} = \{M, I_{k+1}\}$, where $I_{k+1} = \max\{I_j, 1 \leq j \leq k\}$ and M is the neutral mass of the parent peptide. Theoretically each ion should have a complementary ion in an ideal tandem mass spectrum. That is for any ion with a mass X , there should be an ion with a mass Y such that $X + Y = M$. In an experimental spectral data set, some ions may have their complementary ions missing due to various reasons. In such cases, we add the complementary ions back. That is for any j , $1 \leq j \leq k$, if there is no peak $\{M_i, I_i\}$ with $M_i + M_j = M$, we add a peak $\{M - M_j, I_j\}$ to the spectrum.

We now examine the mass differences between pairs of spectral peaks. First note that the mass difference between two ions of the same type (b- or y- or others) always equals to the total mass of some amino acids. However, it is not necessarily true vice versa. That is, two ions with a mass difference of the sum of some amino acid masses do not necessarily belong to the same ion type, i.e., may arise from ions of different types by coincidence. To estimate the probability of a given mass difference to be from two ions of the same type or of different types, we conducted a tandem mass spectrometry experiment in silica on tryptic digested peptides from proteins in bacteria *R. Palustris* genome [15]. We counted and

tabulated the mass differences between two ions of the same type and between two ions of different types (only b- and y-ions are considered). The conditional probability that a given mass difference δ arises from two ions of the same type is then estimated by the ratio of the counts of δ between two ions of the same type to the total occurrences of δ . The result is shown in Fig.1. Apparently, mass differences that do not correspond to the total mass of some amino acids must arise from ions of different types, while mass differences being the mass of one or two amino acids are highly probable to come from ions of the same type (see the information rich zone, 0~200 Da).

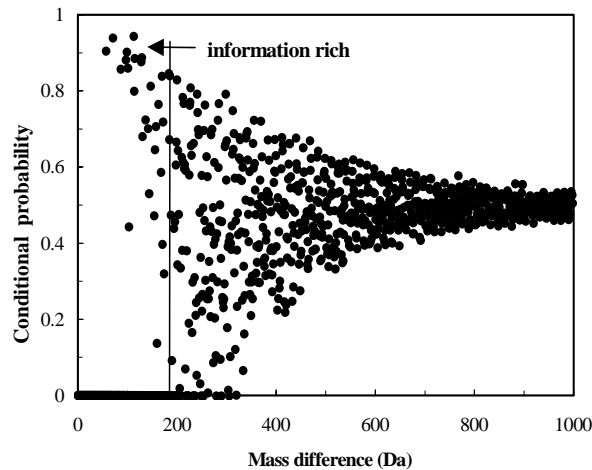


Figure 1. Conditional probability that two ions are of the same type at given mass difference. The distribution of probability is quite discrete in the information rich region (0 ~ 200 Da): either with high probabilities for some specific mass differences which correspond to the total mass of one or double amino acid(s) or with zero value for the other mass bin. For example, for a given mass difference of 57 Da (Gly), the probability that it arises from two ions of the same type is 0.905; while the probability that a mass difference of 50 Da comes from ions of the same type is zero, which means that the two ions definitely belong to different ion types. This feature forms the basis of our algorithm for distinguishing different ion types.

We use the following procedure to construct the spectrum graph. Each peak of tandem mass spectrum is represented as a vertex. A pair of peaks is connected by a type-1 edge if their mass difference is the same as the mass of a single amino acid, or is connected by a type-2 edge if their mass difference is less than or equal to 15 Da. We call this graph $G = (V, E)$ a spectrum graph, with V and E being the vertex and edge set, respectively.

Each edge is then assigned a weight, representing the confidence we have in considering two involved

peaks as the same or different ion types. The weight of a type-1 edge $E(V_m, V_n)$, is defined as follows,

$$W_1(E) = \ln(I_m + I_n) + \ln(\Pr(\delta_{mn})) + \ln(F_i) - \alpha \cdot |m(aa_i) - \delta_{mn}|, 1 \leq i \leq 20 \quad (1)$$

where I_m, I_n are intensities of ions m and n ; $\Pr(\delta_{mn})$ is the conditional probability that ions m and n have the same type, given the mass difference $\delta_{mn} = |M_n - M_m|$; aa_i is an amino acid type which has the smallest $|m(aa_i) - \delta_{mn}|$ value and satisfies the condition $|m(aa_i) - \delta_{mn}| \leq 0.05\text{Da}$, and $m(aa_i)$ is the mass of aa_i ; F_i is the frequency of amino acid aa_i occurring in the target genome. α is a scaling factor. This definition captures the intuition that a good type-1 edge usually has a small mass deviance and connects two peaks with high intensity.

The weight of a type-2 edge $E(V_m, V_n)$ is defined as

$$W_2(E) = \ln(I_m + I_n) \quad (2)$$

If we imagine type-1 edges carrying attractive force and type-2 edges repulsive force, the vertices of the same ion type in a spectrum graph should naturally cluster together, whereas vertices of different ion type repel away. Separation of b- and y-ions then can be done through cutting all the type-2 edges optimally. To cast this intuition into a graph partition problem, we formulate our objective function as to partition the vertices into 3 subsets, B, Y and U. We partition the graph in such a way to maximize the total weight of type-1 edges while minimizing the total weight of type-2 edges within B and Y subsets.

2.2. Mathematical model and Dynamic programming algorithm

Let Ω be the set of all possible tri-partitions of vertices set V of a spectrum graph G . Without loss of generality, we assume that G is a connected graph; otherwise G will be an individual connected component. We define the scoring function $Score(P, G)$ for any tri-partition $P = \{V_B, V_Y, V_U\} \in \Omega$ as

$$Score(P, G) = Q_1 \cdot (W_1(V_Y, V_Y) + W_1(V_B, V_B) - W_1(V_Y, V_B)) + Q_2 \cdot (W_2(V_Y, V_B) - W_2(V_Y, V_Y) - W_2(V_B, V_B)) \quad (3)$$

where $W_i(A, B)$ represents the total weight of type- i edges between vertices of subsets $A, B \subseteq V$, Q_i is a positive factor, $i=1,2$.

Our goal is to find the optimal partition $P^{opt} \in \Omega$ such that

$$Score(P^{opt}, G) = \max\{Score(P, G) \mid P \in \Omega\} \quad (4)$$

We now present a dynamic programming algorithm for solving the optimization problem defined above. For a carefully chosen vertex $v_0 \in V$ (see below), we construct a breadth-first tree [16], $BFT(v_0)$ of the spectrum graph G , with v_0 being the root. Let S_i be a set of vertices on the i -th level of $BFT(v_0)$, where $i = 0, 1, 2, \dots, L$ and L is the length of the longest path. We have $S_0 = \{v_0\}$ and $V = \bigcup_{i=0}^L S_i$. We define a set of subgraphs $G_i = (V_i, E_i)$ such that $V_i = \bigcup_{j=0}^i S_j$ and E_i consists of edges connecting vertices of V_i , $i = 0, 1, 2, \dots, L$. Note that in $BFT(v_0)$, there is no edges between V_i and S_j for any $j > i+1$. The definition is illustrated in Fig.2.

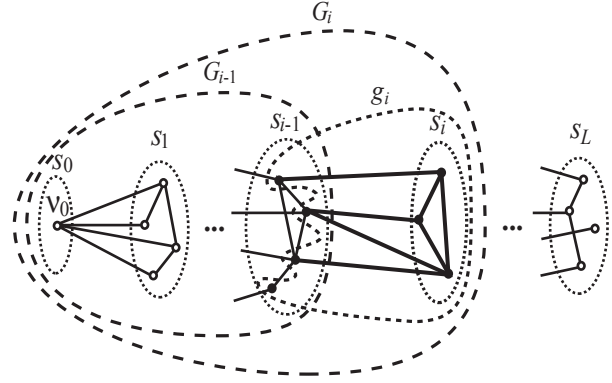


Figure 2. Scheme of graph partition using a dynamic programming algorithm.

For each partition P_i of vertices of S_i , let Ω/P_i be a subset of Ω satisfying P_i . We define the conditional optimal partition of vertices V_i of G_i , $P^{opt}(P_i) \in \Omega/P_i$, as

$$Score(P^{opt}(P_i), G_i) = \max\{Score(P, G_i), P \in \Omega/P_i\} \quad (5)$$

Theorem. For any subgraph G_i of G and any partition $P_i \in \Omega_i$ where Ω_i is a set of all possible partitions of S_i , the conditional optimal partition of vertices V_i of G_i can be decomposed into a conditional optimal partition of vertices V_{i-1} of G_{i-1} and a particular partition of S_i , $i = 1, 2, \dots, L$,

$$Score(P^{opt}(P_i), G_i) = \max\{Score(P^{opt}(P_{i-1}), G_{i-1}) + Score(P_{i-1} \otimes P_i, g_i), P_{i-1} \in \Omega_{i-1}\} \quad (6)$$

where g_i is a subgraph of G formed by vertex set $S_i \cup S_{i-1}$ connected by edges $E_i - E_{i-1}$, and $P_{i-1} \otimes P_i$ presents the virtual union of P_{i-1} and P_i .

Proof. Because the scoring function Eq.3 is additive in terms of the weights of edges, and subgraphs G_{i-1} and g_i do not have any common edges based on their definitions, we always have

$$Score(P, G_i) = Score(P, G_{i-1}) + Score(P, g_i) \quad (7)$$

Combining Eqs. 5 and 7 with the decomposition $\Omega/P_i = \bigcup\{\Omega/(P_{i-1} \otimes P_i), P_{i-1} \in \Omega_{i-1}\}$, we have

$$\begin{aligned} Score(P^{opt}(P_i), G_i) &= \max\{Score(P, G_i), P \in \Omega/P_i\} \\ &= \max\{Score(P, G_{i-1}) + Score(P, g_i), P \in \Omega/P_i\} \\ &= \max\{Score(P, G_{i-1}) + Score(P, g_i), P \in \bigcup\{\Omega/(P_{i-1} \otimes P_i), P_{i-1} \in \Omega_{i-1}\}\} \end{aligned} \quad (8)$$

However $Score(P, G_{i-1})$ is independent of P_i , and $Score(P, g_i)$ depends on P_{i-1} and P_i only. Thus we have

$$\begin{aligned} &\max\{Score(P, G_{i-1}) + Score(P_{i-1} \otimes P_i, g_i), P \in \Omega/P_{i-1}, P_{i-1} \in \Omega_{i-1}\} \\ &= \max\{Score(P^{opt}(P_{i-1}), G_{i-1}) + Score(P_{i-1} \otimes P_i, g_i), P_{i-1} \in \Omega_{i-1}\} \end{aligned} \quad (9)$$

This completes the proof of theorem.

2.3. Implementation

The following pseudo code describes the dynamic programming algorithm for solving the optimal partition problem defined by Eq. 6.

1. Select v_0 based on a specific rule (see below);
2. Build a $BFT(v_0)$ of G ;
3. **For** each partition $P_0 \in \Omega_0$ of S_0 **Do** $Score(P^{opt}(P_0), G_0) \leftarrow 0$;
4. **For** $i = 1$ To L **Step 1 Do**
5. **For** each partition $P_i \in \Omega_i$ of S_i **Do**
6. $max_score \leftarrow 0$;
7. **For** each partition $P_{i-1} \in \Omega_{i-1}$ of S_{i-1} **Do**
8. **If** $Score(P^{opt}(P_{i-1}), G_{i-1}) + Score(P_{i-1} \otimes P_i, g_i) > max_score$ **Do**
9. $max_score \leftarrow Score(P^{opt}(P_{i-1}), G_{i-1}) + Score(P_{i-1} \otimes P_i, g_i)$;
10. $P^{opt}(P_i) \leftarrow P^{opt}(P_{i-1}) \otimes P_i$;
11. Select the partition $P^{opt}(P_L)$ with the maximal $Score(P^{opt}(P_L), G_L)$ value as the final optimal partition of G ;
12. Determine the real ion types of the partitioned vertices as follows.

The dynamic programming algorithm classifies all spectral peaks into three classes, B, Y and U. In the algorithm, we did not specifically use properties that are directly associated with b-ions or y-ions. Therefore B may actually be y-ions and Y may be b-ions. We use two properties of tandem mass spectra to decide which group is the b-ion set/y-ion set. (i) The b-ion group should include a vertex with a mass of 0 and a vertex with a mass of peptide parent mass – 18 Da (i.e., minus H and OH), while the y-ion group should include a vertex with a mass of 18 Da (the complementary ion of the b-ion with full peptide length) and a vertex with parent mass (see Fig.3B); (ii) Statistical analysis of tandem mass spectra has shown that the average intensity of y-ions is typically more than twice of that of b-ions although their ion numbers are almost the same [11, 14]. Based on such information, PRIME reports the ion type of each ion as output.

2.4. Computational complexity

Let $C(G)$ be the computational complexity for calculating the optimal partition of a spectrum graph G defined by Eq. 6 and define $C(G)$ as the number of function $Score(P, G)$ calls. The computational complexity of our dynamic programming algorithm can be derived from the lines 4, 5 and 7 of the pseudo code directly,

$$C(G) \leq O\left(\sum_{i=1}^L 3^{|S_{i-1}|+|S_i|}\right) \quad (10)$$

where i is the distance from the root v_0 of the breadth-first tree $BFT(v_0)$, L is the length of the longest path of the $BFT(v_0)$, and $|S_i|$ is the number of vertices on the i -th level of $BFT(v_0)$. To make $C(G)$ as small as possible, we always exhaustively search through all vertices to find the root v_0 that gives the smallest O value before starting the partition procedure.

2.5. Generalized algorithm—considering chemical variants

Neutral losses from fragment ions are common in tandem mass spectra. A loss of water or ammonia will reduce fragment ion masses by 18 or 17 Da. In addition, protein post-translational modifications (PTMs), a common and important phenomenon in cell functioning, also change the pattern of spectra by shifting a portion of peaks with specified masses. We introduce a new type of pseudo amino acid with the specified mass for each suspected chemical variant, i.e., loss of water or ammonia, or PTMs, and treat the ions arising from neutral losses the same way as b- or y-ion group. For example, for loss of water, we add a pseudo amino acid namely “water” with a mass of -18.011 Da to the amino acid library. Since we don’t know the frequency of these pseudo amino acids in tandem mass spectra, we use a simplified Eq. 11 to calculate the weight of type-1 edge forming by these new residue types,

$$W_1(E) = \ln(I_m + I_n) - \alpha \cdot |m(aa_i) - \delta_{mn}| \quad (11)$$

By doing so, virtually no change is needed in our partition algorithm for dealing with situations with chemical variants.

Table 1. Test results on 18 experimental FT-ICR tandem mass spectra

No	Sequence ^a	b-ions ^b	y-ions ^b	CPU ^c
1	VEAD <u>IA</u> GHGQ <u>EV</u> LIR	18/18	15/15	5
2	HGT <u>VV</u> L <u>TAL</u> G <u>GIL</u> K	7/7	10/10	<1
3	HGT <u>VV</u> L <u>TAL</u> G <u>GIL</u> KK	10/10	6/6	<1
4	VEAD <u>IA</u> GHGQ <u>EV</u> LIR	5/5	13/13	3
5	KGH <u>HEA</u> EL <u>KPL</u> AQSHATK	6/6	2/4	<1
6	LFT <u>GH</u> P <u>ET</u> LEK	2/7	6/7	<1
7	Acetyl-LVFF <u>AE</u> D <u>VGS</u> SNK	6/6	6/6	1
8	GK <u>AK</u> V <u>TGR</u> WK	11/13	2/3	5
9	DA <u>FL</u> G <u>SFL</u> Y <u>EYS</u> R	17/17	11/11	1
10	LVNEL <u>TE</u> FAK	2/4	7/7	1
11	TVMEN <u>FVA</u> F <u>VD</u> K	7/7	3/6	1
12	LGE <u>YGF</u> Q <u>NAL</u> I <u>VR</u>	6/7	7/7	1
13	GL <u>VL</u> I <u>A</u> F <u>SQYL</u> Q <u>QCP</u> FE <u>HVK</u>	4/5	5/6	<1
14	HL <u>VDE</u> P <u>QNL</u> I <u>KQNC</u> D <u>QFE</u> K	2/3	5/6	<1
15	SL <u>HTL</u> F <u>GDE</u> L <u>CK</u>	5/5	6/7	<1
16	G <u>YSL</u> G <u>NW</u> V <u>CAA</u> K	7/7	7/7	1
17	K <u>IVSD</u> G <u>NGM</u> NA <u>WVA</u> WR	3/3	7/11	1
18	NLC <u>NI</u> PC <u>SALL</u> SS <u>DITAS</u> V <u>NCA</u> K	4/4	10/12	1

^a Peptides were either tryptic digested from horse myoglobin (no. 1-6), bovine serum albumin (no. 9-15), lysozyme (no. 16-18), or synthesized (no. 7-8). Peptide no.7 was acetylated. b- and y-ions those were correctly identified were labeled with underline (b-ions) and top-bar (y-ion); ^b The numerator indicates the number of the correctly identified b- or y-ions and their chemical variants; while the denominator are that of observed in experimental spectrum; ^c The unit of CPU used is second. All calculations were performed on a Dell Workstation with Pentium 4 (2.1 GHz).

3. Application and results

We have implemented our partition algorithm as a computer program PRIME (PaRtition of Ion types of tandem Mass spEctra) using C++ programming language and tested it on 18 sets of high accurate FT-ICR tandem mass spectra acquired with an IonSpec

(Irvine, CA) 7-Tesla HiRes electrospray Fourier transform ion cyclotron resonance mass spectrometer (ES-FTICR-MS) from various peptide sources, either tryptic digested from horse myoglobin, bovine serum albumin (BSA), Lysozyme proteins, or synthesized. Raw profile data of FT-ICR tandem mass spectrum served as input. The mass/charge ratio was first transformed automatically to neutral mass, where the isotopic peaks were automatically identified and removed but contributed their associated peak intensities to that of their monoisotopic mass. This preprocessing resulted in the number of ions in each spectrum varied from 22 to 50, which covering 25% to 90% of hypothetical b-ions and y-ions. A spectrum graph representation was then constructed and our dynamic programming algorithm was employed to differentiate the b- and y-ions. The prediction results are summarized in Table 1.

As shown in Table 1, for each of 18 test spectra, PRIME identified most of b- and y-ions and their chemical variants correctly. Among 18 data sets 7 were identified with 100% accuracy. The accuracy was defined as the percentage of the total number of b-, y-ions and their variants that were correctly identified over that of observed in experimental spectrum. The partition accuracy for each individual spectrum ranged from 57% to 100%, with an average accuracy of 88%. The experimental tandem mass spectra with good coverage of b- and y-ions always had high identification accuracy (e.g., the peptide with 100% accuracy).

As an example, Figure 3 shows the experimental tandem mass spectrum of test no.9 and its partition result. Several interesting results can be seen from this figure. (1) All the b- and y- ions and their variants (loss of water here, denoted by X) were identified correctly. (2) The full-length peptide sequence except the first two residues can be derived from either the series of b-ions or y-ions (no distinction between the amino acids L and I). (3) There exists a second continuous mass ladder (572.27, 719.34, 832.43, 995.50, 1124.55, 1287.63, 1374.67, and 1530.74) that apparently corresponds to the loss of water of b-ions series starting from position 6 of the peptide. This evidence strongly indicates that Ser6 lost water during fragmentation. This kind of secondary mass ladders information should be highly useful in the future applications for detecting the types and sites of protein post-translational modifications (PTMs), since X could be any other specified mass change of PTMs, for example, 98 Da of neutral mass loss for phosphopeptides (loss of H₃PO₄). (4) Three type-1 edges (labeled with dashed lines) were coincidentally formed between b-ions and y-ions. PRIME identified them all correctly based on the global optimization.

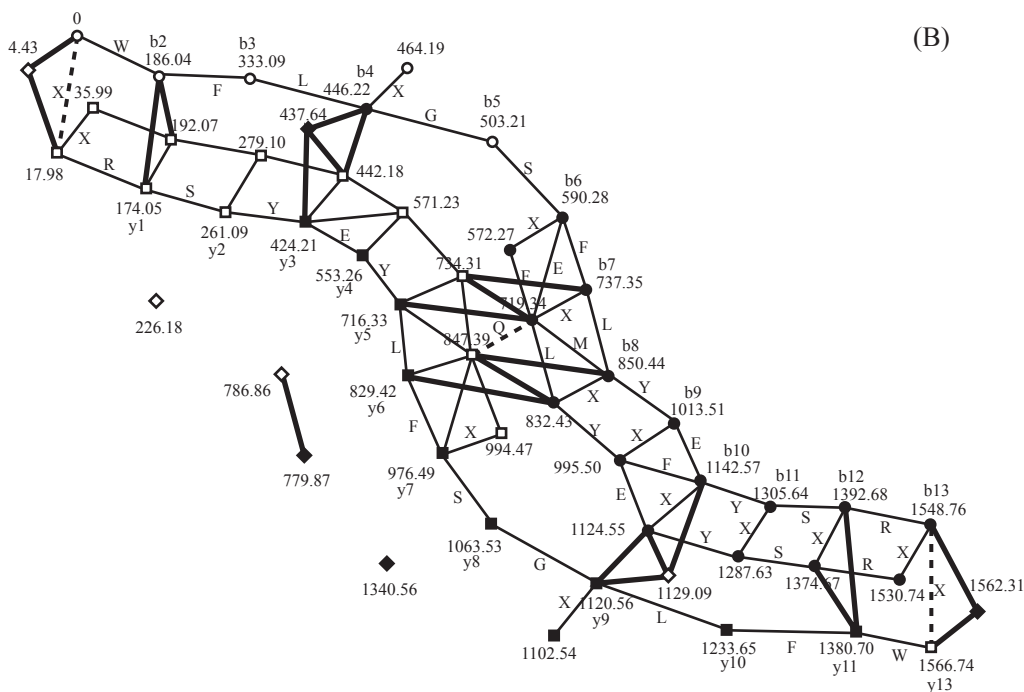
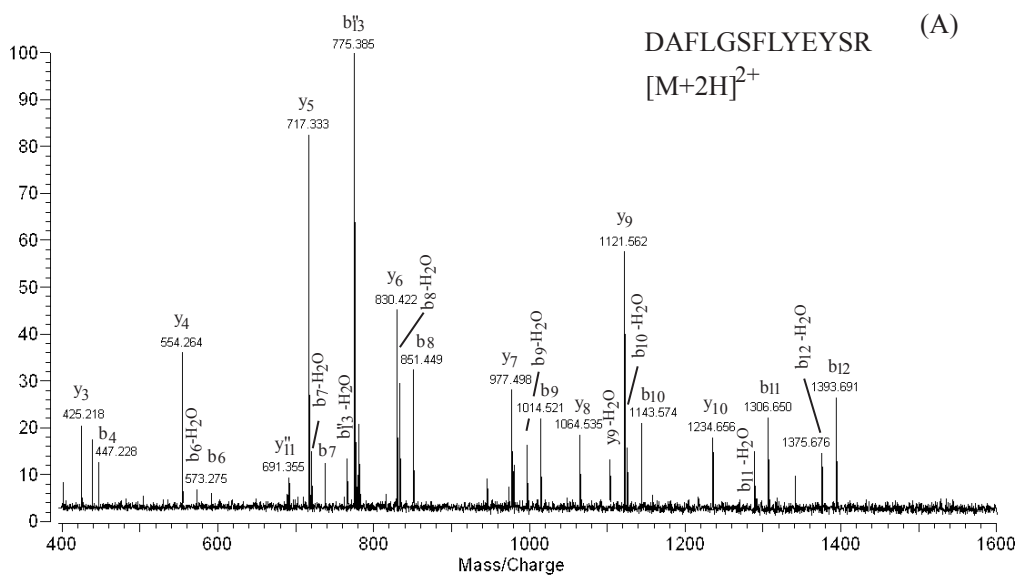


Figure 3. Experimental FT-ICR tandem mass spectrum of test no.9 (A) and its partition results (B). (B) was generated by program Pajek (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>) and polished by Adobe Illustrator 10.0. Neutral masses were used to label the vertices in (B). The b-ions (circles) and y-ions (squares) were partitioned in two subsets, where vertices in each subset were connected through type-1 edges (thin lines) and between the two subsets were connected through the type-2 edges (thick lines). Dashed lines represent the false type-1 edges that formed by ions of different ion types. X represents the loss of water with a mass of -18.011 Dalton. Noises were labeled with diamonds. The closed symbols represent the ions that existed in the experimental spectrum while the open symbols denote the adding back complementary ions.

4. Discussion

Methods for accurate identification of ion types provide the basis for many mass spectrometry data interpretation problems, including (a) *de novo* sequencing and (b) identification of post-translational modifications. Compared to previous *de novo* sequencing methods [10-13], the uniqueness of our approach is that we treat the problem of ion type identification separately from the problem of *de novo* sequencing. By decoupling the two problems, we build a conceptually clearer framework for solving the two problems separately rather than having the two problems tangled together.

We introduce a new type of edge, the type-2 edges which connect two peaks of possibly different ion types, in our spectrum graph representation; and assign each edge (both type-1 and type-2 edges) a weight which reflects the "probability" being of the same or different ion types. Separation of b- and y-ions is done through cutting all the type-2 edges optimally. Among the published papers on *de novo* sequencing methodologies and applications, solely the information of type-1 edges were utilized [10-13]. As shown in Fig. 1, type-1 edges may arise from different ion types, however the probability of having a false type-2 edge is intuitively much smaller than that of having a false type-1 edge. These facts imply that our algorithm does utilize more informational context of a given MS/MS data and probably describe the spectrum more completely compared to the existed *de novo* sequencing methods.

Identification of protein post-translational modifications (PTMs) represents a great interest of understanding the regulation and function of proteins [17]. The capacity of detecting the types and sites of PTMs efficiently should be a key feature of a good *de novo* sequencing algorithm. We have shown that our algorithm can be easily extended to consider chemical variants, i.e. loss of water or ammonia, or PTMs, by introducing a new type of pseudo amino acid with the specified mass for each suspected chemical variants to the amino acid library without increasing the computational complexity.

Several *de novo* sequencing algorithms and software packages, like Lutefisk [10] and PEAKS [13], were reported to perform *de novo* sequencing successfully and the latter also employed a dynamic programming algorithm. However as we pointed out early, the fundamental of these approaches are quite different from ours: no special attempt was made to distinguish ion types and solely the information of type-1 edges was used to construct the spectrum graph representation. Since our effort in this paper is to identify the ion type of each individual peak (a first

stage of *de novo* sequencing), no comparison was made to evaluate the performance of our algorithm with others. However the test of PRIME on the experimental FT-ICR data was encouraging: among the 18 data sets of high quality FT-ICR tandem mass spectra, PRIME achieved ~90% accuracy for identification of ion types. We expect that PRIME will prove to be highly useful for *de novo* sequencing and identification of protein post-translational modifications at the post-genomic era.

Acknowledgement

This research was supported in part by National Science Foundation (# NSF/DBI-0354771) and by the US Department of Energy's Genomes to Life program (www.deogenomestolife.org) under project, "Carbon Sequestration in *Synechococcus* sp.: From Molecular Machines to Hierarchical Modeling" (www.genomes-to-life.org).

References

- [1] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics", *Nature*, 2003. **422**(6928): pp. 198-207.
- [2] J.K. Eng, A.L. McCormack, and J.R. Yates, 3rd, "An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database", *J Am Soc Mass Spectrom*, 1994. **5**(11): pp. 976-989.
- [3] M. Mann and M. Wilm, "Error-tolerant identification of peptides in sequence databases by peptide sequence tags", *Anal Chem*, 1994. **66**(24): pp. 4390-9.
- [4] D. Fenyo, J. Qin, and B.T. Chait, "Protein identification using mass spectrometric information", *Electrophoresis*, 1998. **19**(6): pp. 998-1005.
- [5] D.N. Perkins, D.J. Pappin, D.M. Creasy, et al., "Probability-based protein identification by searching sequence databases using mass spectrometry data", *Electrophoresis*, 1999. **20**(18): pp. 3551-67.
- [6] K.R. Clauser, P. Baker, and A.L. Burlingame, "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching", *Anal Chem*, 1999. **71**(14): pp. 2871-82.
- [7] M.P. Washburn, D. Wolters, and J.R. Yates, 3rd, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology", *Nat Biotechnol*, 2001. **19**(3): pp. 242-7.

- [8] Y. Ho, A. Gruhler, A. Heilbut, et al., "Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry", *Nature*, 2002. **415**(6868): pp. 180-3.
- [9] E. LaSonder, Y. Ishihama, J.S. Andersen, et al., "Analysis of the *Plasmodium falciparum* proteome by high-accuracy mass spectrometry", *Nature*, 2002. **419**(6906): pp. 537-42.
- [10] J.A. Taylor and R.S. Johnson, "Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry", *Rapid Commun Mass Spectrom*, 1997. **11**(9): pp. 1067-75.
- [11] V. Dancik, T.A. Addona, K.R. Clauser, et al., "*De novo* peptide sequencing via tandem mass spectrometry", *J Comput Biol*, 1999. **6**(3-4): pp. 327-42.
- [12] T. Chen, M.Y. Kao, M. Tepel, et al., "A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry", *J Comput Biol*, 2001. **8**(3): pp. 325-37.
- [13] B. Ma, K. Zhang, C. Hendrie, et al., "PEAKS: powerful software for peptide *de novo* sequencing by tandem mass spectrometry", *Rapid Commun Mass Spectrom*, 2003. **17**(20): pp. 2337-42.
- [14] D.L. Tabb, L.L. Smith, L.A. Breci, et al., "Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides", *Anal Chem*, 2003. **75**(5): pp. 1155-63.
- [15] F.W. Larimer, P. Chain, L. Hauser, et al., "Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodospseudomonas palustris*", *Nat Biotechnol*, 2004. **22**(1): pp. 55-61.
- [16] T.H. Cormen, C.E. Leiserson, R.L. Rivest, et al., *Introduction to Algorithms*. 2nd ed. 2001, Cambridge, MA: The MIT Press.
- [17] M. Mann and O.N. Jensen, "Proteomic analysis of post-translational modifications", *Nat Biotechnol*, 2003. **21**(3): pp. 255-61.