Predicting Molecular Formulas of Fragment Ions with Isotope Patterns in Tandem Mass Spectra

Jingfen Zhang, Wen Gao, Jinjin Cai, Simin He, Rong Zeng, and Runsheng Chen

Abstract—A number of different approaches have been proposed to predict elemental component formulas (or molecular formulas) of molecular ions in low and medium resolution mass spectra. Most of them rely on isotope patterns, enumerate all possible formulas for an ion, and exclude certain formulas violating chemical constraints. However, these methods cannot be well generalized to the component prediction of fragment ions in tandem mass spectra. In this paper, a new method, FFP (Fragment ion Formula Prediction), is presented to predict elemental component formulas of fragment ions. In the FFP method, the prediction of the best formulas is converted into the minimization of the distance between theoretical and observed isotope patterns. And, then, a novel local search model is proposed to generate a set of candidate formulas efficiently. After the search, FFP applies a new multiconstraint filtering to exclude as many invalid and improbable formulas as possible. FFP is experimentally compared with the previous enumeration methods, and shown to outperform them significantly. The results of this paper can help to improve the reliability of *de novo* in the identification of peptide sequences.

Index Terms—Isotope patterns, peptide sequencing, tandem mass spectra.

1 Introduction

Tandem mass spectrometry (MS/MS) is an essential and reliable tool for biologists to identify peptide and protein sequences. Many approaches have been designed for peptide sequencing [1], [2], [3], [4], [5], [6], [7]. One important class of these approaches is *de novo* sequencing, which directly derives a (partial) sequence from experimental spectrum [4], [5], [6], [7]. Due to the measurement errors in medium resolution spectrometry, a massive number of candidates will be produced, resulting in identification confusion.

The isotope patterns of ions obtained in tandem spectrum can be used toward resolving this confusion. In spectrum, a peak corresponds to an ion with an elemental component formula (i.e., molecular formula). As it will be shown later, the presented isotope patterns can help to accurately predict the elemental component formulas of ions which, in turn, can help to improve the reliability of *de novo* method. In this paper, we focus on predicting ions' elemental component formulas from isotope patterns.

Several methods and programs, e.g., [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], have been developed to predict the elemental component formulas of molecular ions using isotope patterns. In simple cases, the maximal and minimal numbers of each element are estimated by the intensities of the isotope peaks. Then, constraints of rings and double bonds are used to rule out certain improbable formulas [8], [9], [10], [11]. While, in more sophisticated cases [12], [13], [14], [15], [16], [17], the most possible formulas are predicted by comparing the theoretical isotope patterns with the experimental ones. More specifically, these methods deal with molecular ions containing elements C, H, N, O, S, Si, O, Cl, B, Br, and so on. They usually include three processes: 1) *candidate generation*, which exhaustively enumerates all possible elemental component formulas corresponding to a given mass and tolerant mass error, 2) *filtering*, which excludes formulas violating chemical constraints, and 3) *matching*, which compares the calculated theoretical isotope pattern of each remaining formula with the observed one. The best match is regarded as the most probable formulas.

These methods [12], [13], [14], [15], [16], [17] are capable of providing adequate solutions in some cases. However, they cannot be well generalized to the prediction of component formulas of peptide fragment ions for the following reasons. First, the fragment ions in MS/MS spectra are more complex than molecular ions (e.g., fragment ions have no general rings and double-bands constraints any more). Second, the number of possible formulas increases exponentially with the ions' masses. Due to the difficulty in differentiating the massive number of competing formulas, the reliability of identification and prediction will dramatically decrease. Furthermore, the computing time needed for the exhaustive enumeration and comparison of elemental component formulas will also become prohibitorily high. Therefore, these methods are not

- *J. Zhang, W. Gao, J. Cai, and S. He are with the Institute of Computing Technology, Chinese Academy of Sciences, JDL, Room 701, Power Creative A, No. 1, Shangdi East Road, Haidian District, Beijing, 100080, P.R. China. E-mail: {jfzhang, wgao, jjcai, smhe}@jdl.ac.cn.*
- *R. Zeng is with the Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Science, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai, 2,00031, P.R. China. E-mail: zr@sibs.ac.cn.*
- *R. Chen is with the Institute of Biophysics, Chinese Academy of Sciences, 15 Datun Road, Chaoyang District, Beijing, 100101, P.R. China. E-mail: crs@ict.ac.cn.*

Manuscript received 8 Sept. 2004; revised 4 Dec. 2004; accepted 13 Dec. 2004; published online 31 Aug. 2005.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-0139-0904.

suitable for a large mass range. Finally, some previous methods require that the intensity ratios of isotope peaks be very accurate to guarantee the reliability of the prediction [16], [17]. However, in real experiments, the peak intensities will be modified after the peak centroiding process, hence the intensity ratio may not be accurate enough.

A new method, called FFP (Fragment ion Formula Prediction), is proposed in this paper. FFP is specifically designed for predicting elemental component formulas of peptide fragment ions. Instead of the expensive and exhaustive enumeration of all possible formulas, FFP first converts the prediction of the best formulas to the minimization of "errors." Then, it generates a "seed" or a good starting point using an efficient quadratic programming (QP) technique, and searches the best candidates in the neighborhood of the starting point. We call this generate-and-search method "local search." With the local search, FFP is able to efficiently provide more accurate prediction results in medium and large mass ranges. However, due to the errors in the intensities of isotope peaks, the local search may still end up with many formulas invalid in the real world. Then, a novel multiconstraint filtering method is applied to eliminate those impossible candidates and adapt to the existing intensity accuracies. To do this, we have studied the influence of the elemental composition on the isotope intensities of fragment ions from the SWISS-PROT peptides, and obtained a set of theoretical mean mass-dependent and mass-independent isotope patterns. These mean isotope patterns are deployed to filter invalid and improbable formulas whose isotope patterns are far from the mean value. Combining the mean isotope patterns with the chemical constraint filtering, FFP can accurately predict the true formulas with a high reliability.

Experiments have been conducted to compare the performance of FFP with other two methods, namely, *MS_Enumerate* and *AC*, on a set of Q-TOF MS/MS data. *MS_Enumerate*, a baseline method, predicts possible formulas without using isotopic information. On the other hand, *AC*, proposed by Do Lago [17], is one of the most advanced methods using isotopic information. The performance of the predictions is evaluated with a single match score and a cumulative match score. The experimental results show that in the low mass range of (0~300u), FFP can achieve very high single and cumulative match scores, which are much better than those of the other two methods. In the medium and high mass ranges of (300~800u) and (800~2,000u), FFP still significantly outperforms the other two methods in the cumulative match score. In addition, the computing time that FFP needs is also much less than the previous methods need.

The remainder of this paper is organized as follows: In Sections 2 and 3, some background and materials of FFP are described, respectively. In Section 4, we formulate the problem, followed by a description of the new local search model and multiconstraint filtering. Section 5 provides experimental evaluation and comparison of FFP with previous approaches on several MS/MS spectra. Finally, discussions and conclusions occupy Sections 6 and 7, respectively.

TABLE 1
Ion Types and Formulas for Ion Mass Calculation

Ion type	Ion mass	Ion type	Ion mass
a	[N]+[M]-[CO]	x	[C]+[M]+CO
a^*	a-[NH_3]	y	[C]+[M]+[H_2]
a°	a-[H_2O]	y^*	y-[NH_3]
b	[N]+[M]	y°	y-[H_2O]
b^*	b-[NH3]	z	[C]+[M]-NH
b°	b-[H_2O]	Immonium	[N]+[A]-[CO]
c	[N]+[M]+NH_3		

[N] is the mass of N-terminal group, [N] = 1; [C] is the mass of C-terminal group, [C] = 17; [M] is the mass of the partial peptide, $[M] = \sum_{i \le t \le j} m(A_t)$*; [A] is the mass of the neutral amino acid residue mass, [A] = m(A); [CO] = 28,* $[H_2O]$ *= 18,* $[NH_3]$ *= 17.*

2 Background

In this section, we introduce some basic background of peptide fragment ions and isotope patterns in tandem spectra for the purpose of presenting our work.

2.1 Peptide Fragment Ions in MS/MS Spectra

By collision-induced dissociation (CID), peptides are fragmented and ionized, and the fragment ions are measured by a mass spectrometer for the mass/charge ratio (m/z). The fragment ion is classified as a, b, or c if the charge is retained on the N-terminal, and x, y, or z if the charge is retained on the C-terminal [18], [19]. An immonium ion is an internal fragment with only one single side chain, formed by a combination of the a-type and y-type cleavage [20], [21]. In low energy CID, the predominantly generated ions are a, b, and y types. In addition, ions with a lost ammonia (-17u, denoted as a^*, b^*, and y^*) and a lost water (-18u, denoted a°, b°, and y°) are observed in spectra [22].

For different types of ions, the ions' masses can be calculated from their primary amino acid sequences. Let A be an amino acid with molecular mass $m(A)$. A peptide $P = A_1, \ldots, A_n$, is a sequence of amino acids with $m(P) = \sum_{1 \le i \le n} m(A_i)$, and a partial peptide P' is a substring $A_i \ldots A_j$ of P with mass $m(P') = \sum_{i \le t \le j} m(A_t)$. Then, a fragment ion of a partial peptide P' can be characterized by a modification of P' with mass $m(P') + \delta$ [23]. For example, a y-ion of the partial peptide P' is $m(P') + 19$. Table 1 summarizes the ion types and formulas for the ion mass calculation that we use in this paper.

2.2 Isotope Patterns

Isotopes are elements that contain the same number of protons and electrons but differ in the number of neutrons in nucleus. As we know, the elements of H, C, N, O, and S have different stable isotope distributions (i.e., isotope patterns) in nature [24]. Most proteins are composed of the above five elements and, thereby, have relatively stable isotope patterns. Two ions have different isotope patterns if they have different elemental component formulas. Hence,

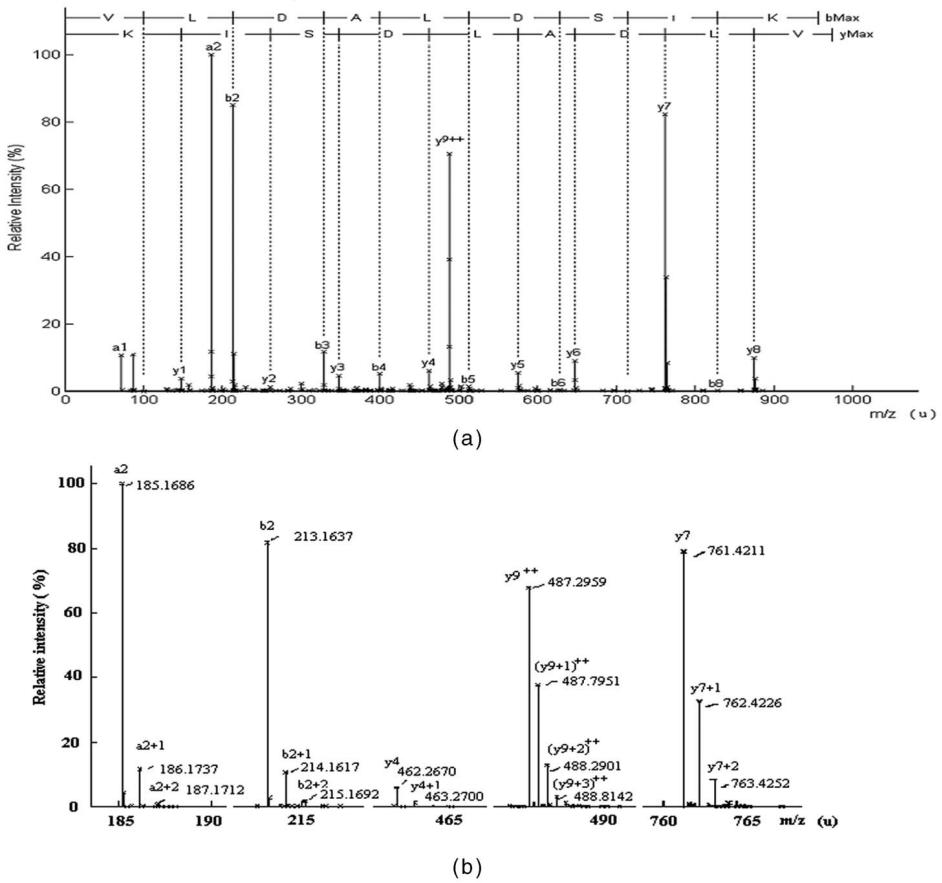

Fig. 1. MS/MS spectrum of the peptide VLDALDSIK (Precursor ion 973.557 u). (a) Major peaks corresponding to the y, b, or a-type ions have been labeled. (b) The close-up view of isotopic resolution of a2, b2, y4, y7, and $\mathrm{y9}^{++}$-ions and peaks.

the component formula of an ion can be derived from the presented isotope pattern in spectrum, and the ion can be identified reliably.

In mass spectrometry, if the selection window is set to transmit roughly 5u precursors in width, the entire group of isotopes of precursors will be fragmented. Therefore, the isotopes of fragment ions will be presented in tandem spectrum and, at the same time, the intensities of an ion's isotopes represent the isotope pattern of the ion. In this paper, it is shown that these isotope patterns can be used to predict the component formulas of ions accurately.

3 Materials

Our experimental data is from Dr. R. Johnson, including 60 Q-TOF spectra from tryptic digestion peptides, which includes 46 peptides without PTM (posttranslational modification) and 14 spectra of peptides with M-oxidation or C-carbamidomethylated.

A computational method is employed to investigate the isotopic information contained in the experimental spectra and it is observed that the relative peak heights of isotope peaks are approximately coincident with the expected isotope patterns of ions (see Section 4.1 for more details). As an example, one spectrum of the peptide VLDALDSIK is shown in Fig. 1. Specifically, Fig. 1a shows the spectrum in which the major peaks corresponding to the y, b, or a-type ions are labeled and Fig. 1b shows the close-up view of isotopic resolution of some ions and peaks. From the close-up view, we can observe that the different relative peak heights indicate different ions' isotope patterns which enable the prediction of component formulas of ions.

4 Method

In this section, a new method, FFP, is presented to predict elemental component formulas of ions based on the isotope patterns presented in experimental tandem spectra. This solution has three new contributions: a key concept of Isotope Pattern Vector (IPV), a "local search" model, and a multiconstraint filtering process.

We first introduce IPV, and define the match score between a formula and the experimental data in Section 4.1. The best formulas will have a minimal match score (or "errors"), which reformulates the problem into a minimization problem. Then, in Section 4.2, a "local search" model is described to find the match with the minimal score. By transforming the match score into a quadratic function of atom's numbers, FFP computes an optimal formula X_R in real domain with a quadratic programming (QP) routine. For example, for the ion with mass = 646.4050 in the spectrum of peptide VLDALDSIK (see Section 3), the computed X_R is $\mathrm{C}_{28.73}\mathrm{H}_{40.22}\mathrm{N}_{7.80}\mathrm{O}_{9.49}\mathrm{S}_{0.00}$ (while the true formula is $\mathrm{C}_{28}\mathrm{H}_{52}\mathrm{N}_{7}\mathrm{O}_{10}\mathrm{S}_{0}$). Since the formula is expressed in real numbers, it cannot be a valid component formula. We regard X_R as a "seed" and then search the integral

TABLE 2
Examples of Isotope Peaks in a Spectrum of the Peptide VLDALDSIK

Partial Peptide	Ion Type	Molecular Formula	Peak m/z	Peak Relative Intensity	Experimental *eIPV*	Theoretical *tIPV*	Match Score
VL	a_2	$C_{10}H_{21}N_2O_1$	185.1686	100%	185.1686	185.1654	0.0078
	a_2+1		186.1737	11.5942%	0.11594	0.12292	
	a_2+2		187.1712	1.0495%	0.01049	0.00895	
VL	b_2	$C_{11}H_{21}N_2O_2$	213.1637	85.1074%	213.1637	213.160285	0.0093
	b_2+1		214.1647	11.0945%	0.13036	0.13449	
	b_2+2		215.1692	1.6992%	0.01996	0.01242	
DALDSIK	y_7	$C_{32}H_{57}N_8O_{13}$	761.4211	82.2089%	761.4211	761.404415	0.0195
	y_7+1		762.4226	33.7331%	0.41033	0.40136	
	y_7+2		763.4252	8.2459%	0.10030	0.10500	
VLDALDSIK	y_9^{++}	$C_{43}H_{77}N_{10}O_{15}$	487.2959	70.4148%	973.58398 [a]	973.556875	0.0374
	$(y_9+1)^{++}$		487.7951	39.1804%	0.55642	0.53570	
	$(y_9+2)^{++}$		488.2901	13.1434%	0.18666	0.17132	

[a] *The observed mass value corresponding to the peak is calculated by* $mass = z \times m/z - (z-1) \times mass(H) = 487.2959 \times 2 - 1.00782 = 973.58398$.

candidates (such as $C_{29}H_{66}N_8O_9S_0$, $C_{28}H_{52}N_7O_{10}S_0$, and $C_{29}H_{10}N_8O_9S_1$) locally in the neighborhood of the seed. Finally, in Section 4.3, a multiconstraint filtering is applied to rule out invalid formulas (such as $C_{29}H_{10}N_8O_9S_1$, which violates chemical constraints) and improbable formulas (such as $C_{29}H_{66}N_8O_9S_0$, which do not match the experimental isotope pattern).

4.1 Isotope Pattern Vector (IPV)

Suppose that the mass of a monoisotopic (partial) peptide P is M, and its first and second isotopes (i.e., with one and two additional neutrons) are P_1 and P_2, respectively. We define the isotope pattern vector (denoted as *IPV*) of P as $\overline{T} = (M, T_1, T_2)$, where T_1 and T_2 are the relative abundance values of P_1 and P_2 with respect to P, respectively. Furthermore, we define *eIPV* as experimental (or observed) *IPV* if M, T_1, and T_2 are obtained from spectrum, and *tIPV* as theoretical *IPV* if M, T_1, and T_2 are calculated from a given elemental component formula.

An ion peak in mass spectrum is characterized in terms of (*m/z*, *intensity*), where *m/z* is the value of mass to charge ratio and *intensity* is the height of the peak. For convenience purpose, we normalize $z = 1$, and use the value of mass to refer to *m/z* in the rest of this paper.

To calculate the value of *eIPV* for isotope peaks of an ion, we consider a group of ion peaks (p_1, p_2, p_3) with (*m/z*, *intensity*) pairs of (M_{e1}, I_{e1}), (M_{e2}, I_{e2}), and (M_{e3}, I_{e3}) in a tandem spectrum. Here, M_{e2} and M_{e3} are approximately equal to $M_{e1} + 1$ and $M_{e1} + 2$, respectively. Then *eIPV* can be obtained by:

$$eIPV = (M_{e1}, I_1, I_2) = (M_{e1}, I_{e2}/I_{e1}, I_{e3}/I_{e1}). \quad (1)$$

To compute the *tIPV* for the elemental component formula of a (partial) peptide, we use the natural isotope distributions of elements C, H, N, O, and S [24] and assume that each isotope of an atom in the (partial) peptide appears independently. Considering a (partial) peptide P with monoisotopic mass M and component formula $C_{n1}H_{n2}N_{n3}O_{n4}S_{n5}$, M is denoted as $M = V \times X$, where $V = (12, 1, 14, 16, 32)$ is the mass vector of the five elements and $X = (n_1, n_2, n_3, n_4, n_5)^T$ is the number vector of the five elements in the formula. In particular, for carbon element, each carbon atom appears randomly as either ^{12}C with probability $q = 0.9889$ or ^{13}C with probability $p = 0.0111$ [24]. Thus, the monoisotopic P appears with probability q^{n1} and its first and second isotopes, P_1 and P_2, appear with probabilities $\binom{n_1}{1}pq^{n_1-1}$ and $\binom{n_1}{2}p^2q^{n_1-2}$, respectively. In other words, the relative abundance values of P_1 and P_2 with respect to P are $T_1 = n_1q_C$ and $T_2 = \frac{1}{2}T_1^2 - \frac{1}{2}n_1q_C^2$, where $q_C = p/q$. After considering all elements of C, H, N, O, and S, $tIPV = (M, T_1, T_2)$ can be obtained as follows:

$$M = V \times X, \quad (2)$$

$$T_1 = n_1q_C + n_2q_H + n_3q_N + n_4q_{O1} + n_5q_{S1}, \quad (3)$$

$$T_2 = n_4q_{O2} + n_5q_{S2} + \frac{1}{2}T_1^2 - \frac{1}{2}(n_1q_C^2 + n_2q_H^2 + n_3q_N^2 + n_4q_{O1}^2 + n_5q_{S1}^2), \quad (4)$$

where q_C, q_H, and q_N are the relative abundance values of ^{13}C to ^{12}C, D to H, and ^{15}N to ^{14}N, and $q_{O1}, q_{O2}(q_{S1}, q_{S2})$ are the ratio of ^{17}O to ^{16}O, ^{18}O to ^{16}O (^{33}S to ^{32}S, ^{34}S to ^{32}S), respectively.

Consider a component formula $C_{n1}H_{n2}N_{n3}O_{n4}S_{n5}$ with $tIPV = (M, T_1, T_2)$ (denoted as $\overline{T}$) and a group of observed ion peaks (p_1, p_2, p_3) with $eIPV = (M_{e1}, I_1, I_2)$ (denoted as $\overline{I}$). We define the match score between the formula and the observed peaks as the Euclidian distance E of $\overline{T}$ and $\overline{I}$ as

$$E = \sqrt{\delta_m^2 + \delta_1^2 + \delta_2^2} = \sqrt{(M - M_{e1})^2 + (T_1 - I_1)^2 + (T_2 - I_2)^2}. \quad (5)$$

As an example, Table 2 simply illustrates some isotope patterns of peptide VLDALDSIK (see Section 3), which includes (*m/z*, *intensity*) pairs of the major ion peaks, *eIPV*,

tIPV and the match score E between the isotope peaks and its component formulas. Observe that the match scores are very small indicating the consistency between the experimental isotope patterns and the expected ones. Hence, the component formulas of the ions can be predicted with a high reliability by comparing the experimental patterns with the theoretical patterns.

4.2 Local Search Model

For *candidate generation*, most previous work exhaustively enumerates all possible integral component formulas, resulting in a high computational complexity and low prediction reliability. To overcome these weaknesses, we introduce a local search model in the following sections. Specifically, we convert the prediction of the best formulas to the minimization of the match between the theoretical and observed isotope patterns. First, by expressing the match score with a quadratic function of elements' numbers, an optimal formula X_R can be computed with a QP technique in a continuous space. The only problem is that X_R is expressed in real numbers (such as $C_{28.73}H_{40.22}N_{7.80}O_{9.49}S_{0.0000}$). Then, the true integral formulas (such as $C_{28}H_{52}N_7O_{10}S_0$) is searched in a discrete space around X_R.

4.2.1 Predicting the Initial Formula with Quadratic Optimization

In this section, the prediction of ions component formulas is transformed into a quadratic optimization problem. By introducing formulas (2), (3), (4), and (5), there are

$$\delta_m = n_1 * 12 + n_2 * 1 + n_3 * 14 + n_4 * 16 + n_5 * 32 - M_{e1}, \quad (6)$$

$$\delta_1 = n_1 * q_C + n_2 * q_H + n_3 * q_N + n_4 * q_{O1} + n_5 * q_{S1} - I_1, \quad (7)$$

$$\begin{aligned}\delta_2 = {} & n_4 * q_{O2} + n_5 * q_{S2} \\ & - \frac{1}{2}(n_1 * q_C^2 + n_2 * q_H^2 + n_3 * q_N^2 + n_4 * q_{O1}^2 + n_5 * q_{S1}^2) \\ & + (n_1 * q_C + n_2 * q_H + n_3 * q_N + n_4 * q_{O1} + n_5 * q_{S1}) \\ & * I_1 - \frac{1}{2}I_1^2 - I_2 + \frac{1}{2}\delta_1^2. \end{aligned} \quad (8)$$

We characterize $[\delta_m \delta_1 \delta_2]^T = AX + B$ by omitting the residue $\frac{1}{2}\delta_1^2$ in the formula for δ_2; hence, E^2 can be transformed into a quadratic function of elements' numbers, given as follows:

$$Q(X) = E^2 = [\delta_m \delta_1 \delta_2] \begin{bmatrix} \delta_m \\ \delta_1 \\ \delta_2 \end{bmatrix} = X^T A^T A X + 2B^T A X + B^T B, \quad (9)$$

where $X = [n_1, n_2, n_3, n_4, n_5]^T$ is a vector representing the elements' numbers of a component formula, and A and B are constant matrixes which can be derived (not shown).

Moreover, to make elemental component formulas chemically meaningful, several constraints for elements are set as follows:

1. The mass calculated from the formula must be within the range of $[M_{e1} - \delta, M_{e1} + \delta]$ with respect to the maximal tolerant *m/z* error of δ, i.e., $|VX - M_{e1}| < \delta$.
2. The maximal number of a given element is the integer part of the *m/z* divided by the mass of the lowest weight isotope.
3. The number of C is less than the number of H, and the numbers of O and N are less than the number of C, etc. These constraints are implicitly derived from the molecular formulas of amino acids and the components of the main ion-types.
4. The sum of H and N in a one-charged ion is odd. The reason is that an unsaturated chemical chain exists if the ion is singly charged. In addition, H and N have odd valences, while C, O, and S have even valences.

Finally, by converting these constraints to linear inequalities denoted as $DX \leq G$, the prediction of component formula of a (partial) peptide ion can be transformed into a minimization of match score. The minimization problem can be solved by a standard quadratic programming, which is formulated as follows:

$$\begin{aligned} \text{minimize } E = & \sqrt{Q(X)} = \sqrt{X^T A^T A X + 2B^T A X + B^T B} \\ & \text{Subject to} \quad DX \leq G. \end{aligned} \quad (10)$$

The optimal X_R, which is associated to the minimal value of E, can be computed by solving (10) in the real domain. The true formula, on the other hand, must be expressed in the integral domain, and would be near X_R since it should have a small E. In the next section, we will discuss the method to find such a formula.

4.2.2 Searching the True Integral Formulas

To find the true formula, we treat X_R as a starting point, and then locally search the integral candidates in the neighborhood of X_R, which is a discrete space. More specifically, all integral formulas within a distance d from X_R are scored. The scale of d determines the number of candidate formulas. In experiment, the value of d is adapted by the mass range and the measured mass error of the ions. For example, in the medium mass range of (0~800u), the value of d can be set to 3. In the large mass range of (800~2,000u), d can be set to 3 if the measured mass error is less than 50ppm, or 5 if the mass error is larger than 50ppm. We use the default value of 5 in this paper. In this way, without enumerating all possible formulas, FFP can predict the elemental component formulas of ions within a large mass range while simultaneously ensuring a high reliability and efficiency.

As it will be mentioned in Section 5, the advantage of local search model becomes more evident as the ions' masses increase. This is because the exhaustive enumeration method inevitably generates a large number of noise formulas which are difficult to be filtered; while local search can ensure more accurate prediction by reducing the search space. Furthermore, the computing time of local search is constant with respect to the ion's mass while the time of the exhaustive enumeration increases exponentially.

4.3 Multiconstraint Filtering

After local search, a number of potential candidate formulas may still be generated. To improve the accuracy of the prediction, one needs to discard as many invalid and improbable formulas as possible. In FFP, we achieve this goal by multiconstraint filtering which uses the mean isotope patterns, the chemical constraints, and cross-validation. To

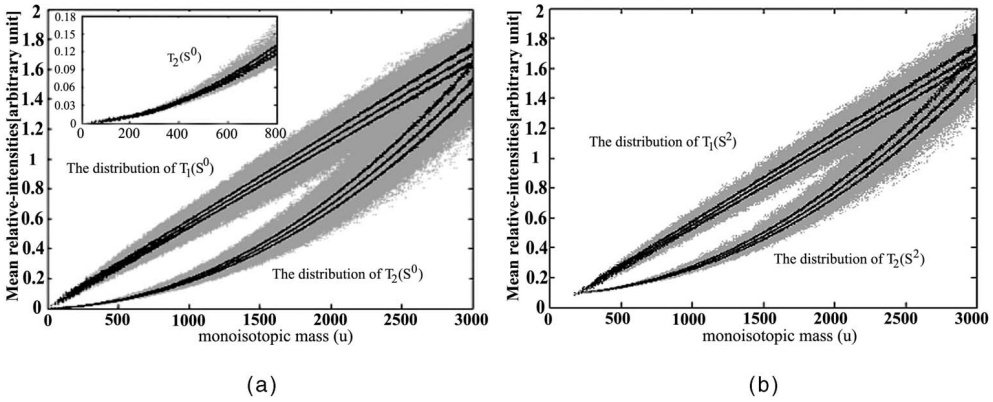

Fig. 2. Distributions of T_1, T_2 are shown here as a function of the monoisotopic ion's mass. Shaded areas denote the range between the minimal and maximal T_1, T_2 of all possible formulas. The curves represent the mean and standard deviations, respectively.

the best of our knowledge, this is the first time that the mean isotope patterns and cross-validation are used to exclude those inappropriate formulas in prediction.

To find the mean isotope patterns of fragmental ions, we calculate isotopic distributions and standard deviations of all peptides to reveal the relationship among the three components (M, T_1, T_2) of *tIPV*. Specifically, we first compute peptides through a theoretical enzymatic cleavage of 10,000 proteins contained in the release 44.2 of SWISS-PROT. Then, we select peptides in the mass range of (60~3000u), which corresponds to the standard domain of Q-TOF MS/MS experiments, resulting in a list of 1.68 million different formulas. It is noted that the element sulfur has an abundant isotope $^{+2}$S (frequency of 0.04210, 20 times more than the same isotope for the oxygen), but most peptides rarely contain more than five sulfurs. Hence, we classify the above 1.68 million formulas into six categories: S^0, S^1, S^2, S^3, S^4, and S^{5+} corresponding to the partial peptides containing 0, 1, 2, 3, 4, and 5 or more sulfurs, respectively. The statistics of T_1, T_2, and M according to the six categories are collected and shown in the following sections.

4.3.1 Mass-Dependent Mean Isotope Patterns

The mass-dependent mean isotope patterns of category S^0 within interval $\delta = 1$u is depicted in Fig. 2. Specifically, a wide range distribution within (60~3000u) is shown in Fig. 2a, which indicates that T_1 is linearly dependent on M, while T_2 quadratically increases with M. The narrow range distribution within (0~800u) is shown in the upper left small figure. It shows that the oscillation of T_2 is tiny (the amplitude is less than 0.03) and, hence, within the narrow range, the mean of T_2 can be used to discard random formulas whose T_2 is out of the oscillations. The mean of T_1 can be used in a similar way. For other five categories (i.e., S^1, S^2, S^3, S^4, and S^{5+}), T_1 and T_2 have similarly linear and quadratic distributions as in S^0. For simplicity, only the example for S^2 is illustrated in Fig. 2b.

The mean and standard deviations of T_1 within interval $\delta = 1$u for each category are calculated and fitted by polynomial curves. The detailed processes are depicted in Appendix A. From the polynomial fitting, it is observed that the mean and standard deviations of T_1 of the other five categories are similar to that of S^0, while in the polynomial expression, the power coefficients of the mean T_2 of different categories are very different. This is because the element S has a more abundant isotope $^{+2}$S. Hence, we can use the mean of T_2 to filter those improbable formulas in which the containment of sulfur does not match the experimental data well. We illustrate the mean curves of T_2 of different categories in Fig. 3.

To conclude, the distributions (i.e., the mean and standard deviations) of T_1 and T_2 can be applied to discard not only invalid formulas, but also improbable formulas with the same mass while from different categories.

4.3.2 Mass-Independent Mean Isotope Patterns

Besides the mass-dependent mean isotope patterns, the mass-independent relationship between T_1 and T_2 can also be statistically calculated, which is shown in Fig. 4. We observe that 1) T_2 increases quadratically with T_1 in all six categories, and 2) points (T_1, T_2) calculated from different categories are located in different distribution bands. For example, within the narrow range [(0, 0.5), (0, 0.4)], points (T_1, T_2) can be clearly distinguished, as illustrated in Fig. 4b. In other words, the formulas from different categories can be easily distinguished by the relationship of T_1 and T_2. The

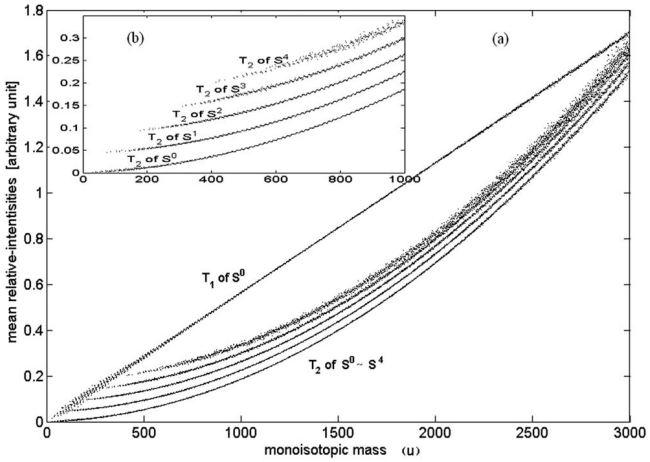

Fig. 3. The mass-dependent mean of T_2 of different categories. (a) and (b) depict the curves in wide and narrow ranges, respectively.

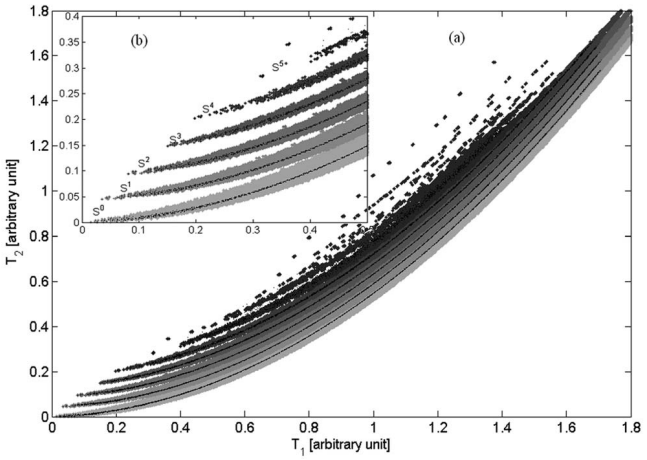

Fig. 4. Mass-independent mean isotope patterns are shown here. The shaded areas denote the range between minimum and maximum values of (T_1, T_2) and the curves represent the mean values. (a) and (b) depict distributions of different categories in the wide and narrow ranges, respectively.

derivation of the mean and standard deviations of T_2 with respect to T_1 is depicted in Appendix B.

While the mass-dependent mean isotope patterns discussed earlier can constrain T_1 and T_2 of a valid formula according to the mass, the mass-independent mean isotope patterns constrain the relationship of T_1 and T_2 of a valid formula. Furthermore, the mean isotope patterns of (T_1, T_2) can be used to filter those improbable candidates by distinguishing formulas from different categories, similar to the case of mass-dependent mean isotope patterns.

It should be mentioned that there have been a couple of previous works for calculating the mean isotope patterns [25], [26], [27], [28]. However, they focus on isotope patterns of peptides under certain criteria, which are completely different from the concept of *IPV* proposed in this paper. Hence, these results cannot be used for fragment ions.

4.3.3 Cross-Validation

Finally, cross-validation is applied in the multiconstraint filtering. In particular, the b-series ions of a partial peptide, including b, a, b^*, a^*, b°, a°-ions (with mass difference of 28, 17, 18, see Table 1), are homologous. They share the same primary sequences [23] and, consequently, share similar isotope patterns. The same is true for the y-series ions. FFP regards two *eIPVs* of $\overline{I_1} = (M_{e1}, I_{11}, I_{12})$ and $\overline{I_2} = (M_{e2}, I_{21}, I_{22})$ as homologous if the difference between M_{e1} and M_{e2} is 28, 17, or 18, and (I_{11}, I_{12}) is close to (I_{21}, I_{22}). Then, FFP uses the homology of *eIPVs* to cross-validate the predictions. For example, consider $\overline{I_1} = (M_{e1}, I_{11}, I_{12})$ and $\overline{I_2} = (M_{e1} - 28, I_{21}, I_{22})$. If there exits a candidate $C_{n1}H_{n2}N_{n3}O_{n4}S_{n5}$ predicted for $\overline{I_1}$ while formula $C_{n1-1}H_{n2}N_{n3}O_{n4-1}S_{n5}$ does not appear at the candidate list of $\overline{I_2}$, then FFP regards $C_{n1}H_{n2}N_{n3}O_{n4}S_{n5}$ as a random result and discards it.

5 Experimental Investigations

In this section, the performance of FFP is evaluated via a set of Q-TOF MS/MS data. We first introduce data preprocessing in Section 5.1, and performance metrics in Section 5.2. Then, the performance of FFP is demonstrated and compared with other two previously proposed methods, *MS_Enumerate*, and *AC* (we reimplement *AC* according to the algorithm presented in [17]). In Section 5.3, the experimental results show that FFP outperform these two methods significantly. Finally, in Section 5.4 the efficiency of the two key techniques, the local search model and multiconstraint filtering in FFP is demonstrated.

5.1 Data Preprocessing

For a given MS/MS spectrum, FFP first searches all potential isotope peaks. In order to predict the elemental component formulas from *eIPV*s, there must be at least one ion with three isotope peaks in the spectrum. We examine the potential isotope peaks in a spectrum and discard spectrum that contains no ions with three isotope peaks. After this initial sorting, the remaining data contain 50 spectra, which include 40 spectra of peptides without PTM and 10 spectra of peptides with M-oxidation or C-carbamidomethylated.

To search potential groups of isotope peaks and compute the *eIPVs* from a spectrum, we first set some thresholds and discard small peaks (e.g., noise) below the thresholds. For the high mass (> 500u), a threshold of 4 percent in relative height of the monoisotopic peak of an ion is chosen, while for the low mass (≤ 500u), a threshold is set to 2 percent. Then, FFP uses the mass-dependent mean isotope patterns (which are also used as multiconstraint filtering as in Section 4.3) to determine whether a group of peaks is a group of isotope peaks of one ion. More specifically, if an observed *eIPV* of $\overline{I} = (M_{e1}, I_1, I_2)$ corresponds to a group of peaks (p_1, p_2, p_3), and I_1 is larger than $(mean_1(k) + plus_1(k) + \delta)$ or less than $(mean_1(k)\text{-}minus_1(k)\text{-}\delta)$ (see Appendix A and Appendix B), for $k = 0 \sim 5$, FFP regards (p_1, p_2, p_3) as noise or overlapping signals, and discards it. Here, δ represents the tolerant error of relative intensities in spectrum, set by the users. The similar procedure is also applied in I_2.

After searching potential isotope peak groups, there are 906 groups (i.e., 906 ions) selected from the 50 spectra in total. In these 906 groups, there are 726 groups whose true formulas can be identified from the known peptides' sequences while 180 groups are unknown ions. The performance of FFP and other two algorithms, *MS_Enumarate* and *AC*, are evaluated on these 726 known ions.

5.2 Performance Metrics

For performance metrics, we define a single match score and a cumulative match score as follows: For each group of isotope peaks $\overline{I} = (M_e, I_1, I_2)$, FFP outputs a prediction including a list of candidate formulas. The true formula of each known ion, on the other hand, can be calculated according to the peptide sequence. Then, we can find the rank number of the true formula in the rank list (the smaller the better) of the candidate formulas. For the rank number k, the single match score (*m_score*) is defined as the percentage of true formulas appearing at rank k, and the cumulative match score (*cm_score*) is the percentage of true formulas appearing at or below k.

With the increase of the ions' masses, the possible combinations of the candidate formulas will increase exponentially. Thereby, the reliability of component formula prediction will decrease. The following section shows the

TABLE 3
The Predictive Accuracy of FFP, *MS_Enumerate*, and *AC*

Mass range	Number of ions	count / *cm_score* of FFP			count / *cm_score* of *MS_Enumerate*			count / *cm_score* of *AC*		
		Top 1	Top 5	Top 20	Top 1	Top 5	Top 20	Top 1	Top 5	Top 20
0~300u	237	196 / 0.83	231 / 0.97	235 / 0.99	164 / 0.69	222 / 0.94	236 / 0.996	52 / 0.22	150 / 0.63	229 / 0.97
300~500u	135	67 / 0.50	128 / 0.95	130 / 0.96	30 / 0.22	87 / 0.64	133 / 0.99	13 / 0.10	41 / 0.34	95 / 0.70
500~800u	155	24 / 0.15	100 / 0.65	140 / 0.90	5 / 0.03	29 / 0.19	82 / 0.53	6 / 0.04	27 / 0.17	76 / 0.49
800~2000u	199	17 / 0.09	66 / 0.33	143 / 0.72	0 / 0	9 / 0.05	26 / 0.13	2 / 0.01	11 / 0.06	38 / 0.19

performances of FFP compared with the other two methods in four mass ranges (0~300u), (300~500u), (500~800u), and (800~2,000u).

5.3 Comparing FFP with *MS_Enumerate* and *AC*

In this section, we compare the performance of FFP with that of two previous proposed methods. First, we consider a baseline *MS_Enumerate* to validate the efficiency of the isotope patterns. *MS_Enumerate* performs component prediction based on mass information but does not consider isotopic information. Second, isotope patterns are considered. Although many such methods (e.g., [12], [13], [14], [15], [16], [17]) have been proposed, it is concluded [17] that *AC* quantitatively achieves better results. Hence, we only compare the performance of FFP with *AC* in this paper. The performance of FFP is compared with *MS_Enumerate* and *AC* on the same data sets. The three methods share the same chemical constraints. We first investigate the predictive accuracy of FFP, *MS_Enumerate*, and *AC* on the 50 spectra. The experimental results are summarized in Table 3.

We can draw the following three conclusions from Table 3:

1. FFP predicts the best in (0~300u). For example, the *cm_score* of top-1 and top-5 are up to 0.83 and 0.97, respectively. The reason is that in this range, the mass errors in spectra, which play a key role in the formula prediction (see (5) in Section 4.1), are very small due to the short flight time of ions in spectrometry. Furthermore, FFP can achieve higher *m_scores* in top-1, roughly 14 percent and 60 percent (derived from the three *cm_scores* of Top 1 in Table 3, which are 0.83, 0.69, and 0.22, respectively), than the other two methods.
2. In (300~500u) and (500~800u), FFP achieves higher *cm_score* in top-5 and top-10 than the other two methods. With increase in ions' masses, the reliability of prediction decreases. However, the local search model and multiconstraint filtering play important roles in improving the predictive accuracy of FFP. In these ranges, it is difficult to distinguish the elemental components of ions only depending on the ions' masses and chemical constraints and, hence, the accuracy of *MS_Enumerate* decreases dramatically. *AC* is the worst because of its exhaustive enumeration and ignorance of the mass errors in experiments.
3. Even in the large mass range of (800~2,000u), which results in larger mass and intensity errors, FFP can still make reasonably adequate predictions. For example, it achieves *cm_score* of top-20 nearly up to 0.72. At the same time, the prediction of *MS_Enumerate* and *AC* are much lower (0.13 and 0.19, respectively).

For detailed comparison, Fig. 5 illustrates the performance curves of FFP, *MS_Enumerate*, and *AC* in terms of *cm_scores* at each rank number. From Fig. 5, it is obvious that FFP outperforms the other two methods at almost all ranks. However, in the range of (300~500u), *MS_Enumerate* produces a slightly higher *cm_scores* than FFP at a high rank of above 16 (see Fig. 5b). This is because either the local search model or the multiconstraint filtering may occasionally exclude the true formula, while *MS_Enumerate* includes all possible candidates.

It is also shown that the accuracy of prediction generated by all the three methods decreases with the increase of the ions' masses. However, the decrease of FFP is much slower than the other two methods, which means that the advantage of FFP is more evident as the ions' masses increases, making FFP more useful in a wide range of ion masses.

Next, we compare the performance of FFP with the other two methods on individual spectrum. From Table 3, we can derive that, for the 50 spectra data, FFP provides 525 predictions in which the true formula is ranked within the top 5. In other words, FFP can provide an average of 10.5 (525/50) predictions ranking the true formula within the top 5 for each spectrum. However, *MS_Enumerate* and *AC* produce only 7.0 (347/50) and 4.6 (229/50), respectively. As shown in Section 6, the more the formulas are predicted correctly, the more the confidence FFP will provide to the peptide sequences constructed by *de novo*.

Furthermore, FFP is evaluated in terms of the single match score (*m_score*) for each individual rank and the detailed performance is depicted in Fig. 6. From Fig. 6a, it is observed that in (0~300u), the *m_score* of rank 1 is above 0.83, indicating that the predicted formula in top 1 is highly credible. In (300~500u), the *cm_score* of the top 5 is still up to 0.948 (see Fig. 6b), indicating that the top five candidates are very credible. It will be discussed in Section 6 that the credible formulas can help *de novo* to refine peptide sequencing. It is our future work to improve FFP in the large mass range.

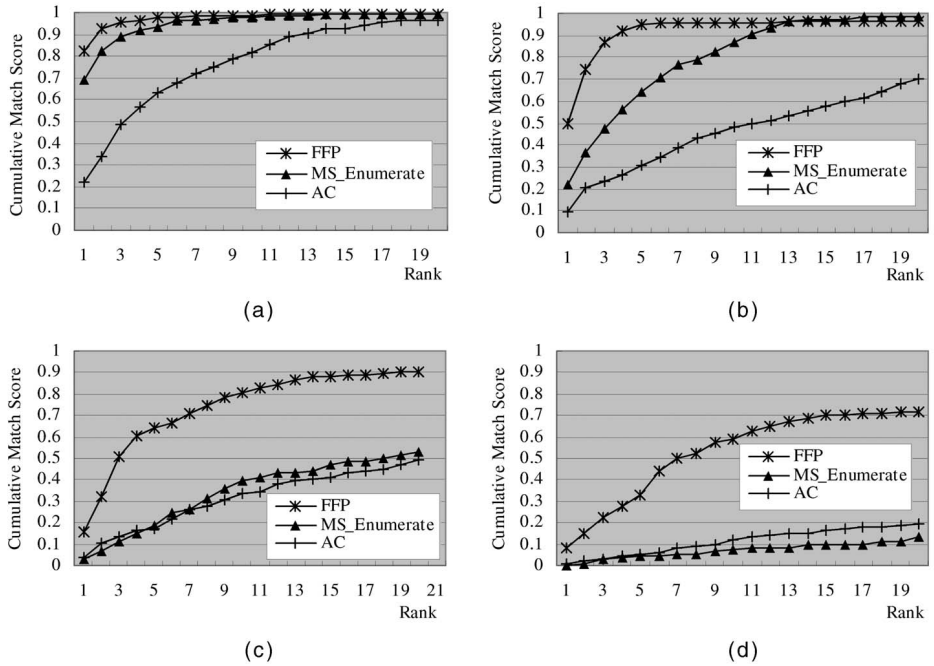

Fig. 5. Performance curves of the *cm_scores* of FFP, *MS_Enumerate*, and *AC*. (a), (b), (c), and (d) depict the performances of FFP, *MS_Enumerate*, and *AC* in the mass ranges of (0~300u), (300~500u), (500~800u), and (800~2,000u), respectively.

Finally, FFP is more computationally efficient than the other two methods. Because of the local search model, the computing time of FFP is roughly constant with respect to the ion's mass. However, the computing time of the exhaustive enumeration, which is used by *MS_Enumerate* and *AC*, increases exponentially. For example, when running on the same PC (CPU: Pentium IV, RAM: 256Mb, OS: Windows), FFP takes 8 seconds to finish all the computation on the 50 spectra data, while both *MS_Enumerate* and *AC* need 58 seconds. In the next section, the reasons for the good performance of FFP will be discussed.

5.4 The Source of Good Performance of FFP

As described in Section 4, the local search model and multiconstraint filtering are the two novel contributions in FFP. Here, we investigate the efficiency of these two techniques. We compare the "Local Search" (see Section 4.2) with the "Global Enumeration" strategy which enumerates all possible formulas, and evaluate "Multiconstraint Filtering" by comparing it with "Chemical-constraint Filtering." The comparison shows that Local Search is more effective when ion's mass is larger than 500u, while the advantage of Multiconstraint Filtering is prominent in the range of (0~800u). In addition, the relative contribution of these two techniques in FFP is also evaluated.

5.4.1 The Efficiency of Local Search

We compare "Local Search" with "Global Enumeration" in the *candidate generation* process and keep other processes (i.e., *filtering by* Multiconstraint Filtering and *matching* by (5) in Section 4.1) intact in the experiments. Fig. 7 depicts the

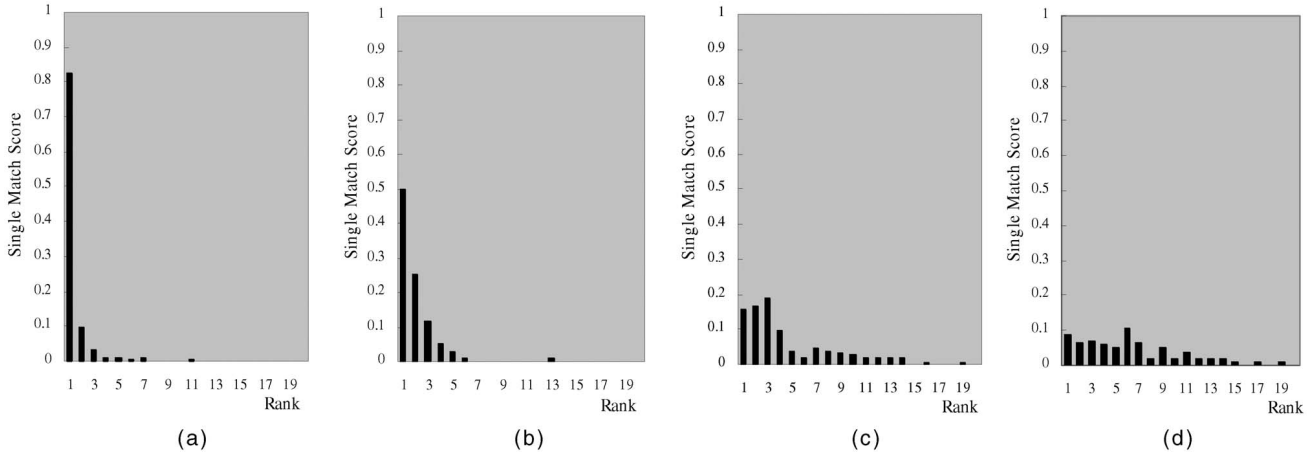

Fig. 6. Performance of FFP in terms of the single match score. (a), (b), (c), and (d) depict FFP's performance in the mass ranges of (0~300u), (300~500u), (500~800u), and (800~2,000u), respectively.

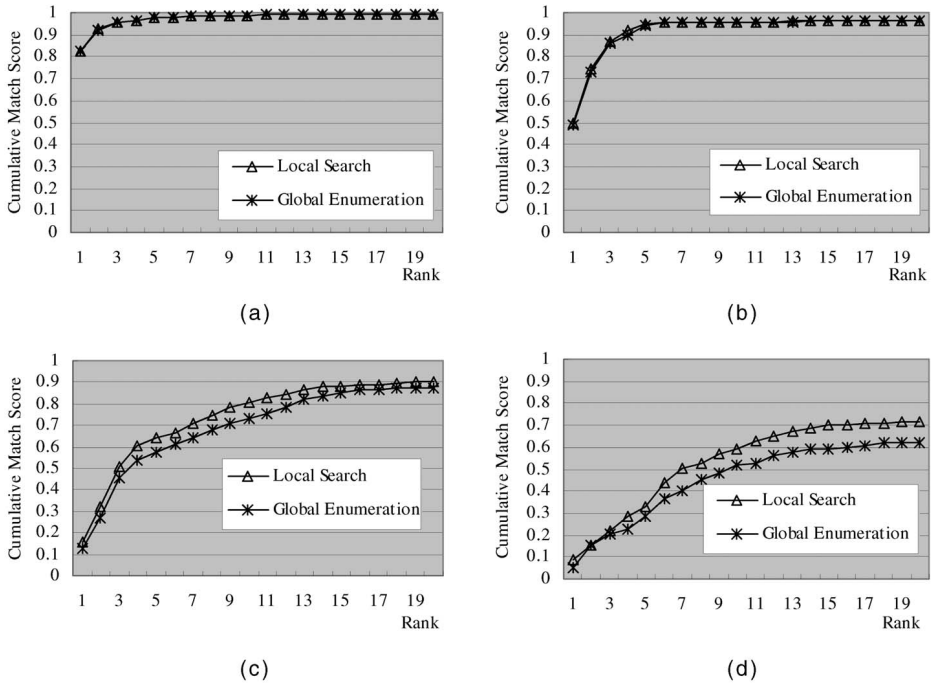

Fig. 7. Efficiency of Local Search and Global Enumeration. (a), (b), (c), and (d) depict the performances of Local Search and Global Enumeration in mass ranges of (0~300u), (300~500u), (500~800u), and (800~2,000u), respectively.

*cm_score*s of Local Search and Global Enumeration in four mass ranges.

In Figs. 7a and 7b, the curve of Local Search is virtually overlapping with that of Global Enumeration, indicating these two methods have almost equal results in (0~300u) and (300~500u). The advantage of Local Search is not prominent because the possible combinational formulas are relatively few in these ranges, and the Multiconstraint Filtering process discards most of impossible formulas. However, in Figs. 7c and 7d the distance between the two curves is significant, which indicates the efficiency of Local Search in (500~800u) and (800~2,000u). Local Search achieves a higher *cm_score* than Global Enumeration as much as 2~10 percent at each rank. This is because Global Enumeration inevitably generates a large number of noisy formulas difficult to be filtered and discarded, while Local Search can achieve more accurate prediction by reducing search space significantly.

5.4.2 The Efficiency of Multiconstraint Filtering

To evaluate the efficiency of "Multiconstraint Filtering", we compare it with "Chemical-constraint Filtering." The Multiconstraint Filtering includes the chemical constraints $DX \leq G$, the mean isotope patterns, and cross-validation (see Section 4.3). The other processes (i.e., *candidate generation* by Local Search and *matching* by (5)) are kept intact as discussed earlier. Fig. 8 depicts the comparative *cm_score* curves of Multiconstraint Filtering and Chemical-constraint Filtering.

From Figs. 8a, 8b, and 8c, it can be observed that Multiconstraint Filtering achieves significantly higher *cm_scores* of a low rank (below 5) than Chemical-constraint Filtering, indicating that in the small and medium mass range of (0~800u), Multiconstraint Filtering can filter impossible formulas more effectively. This can be ascribed to the small oscillation of T_1 and T_2 in the mean isotope patterns. However, for heavy ions (> 800u), the two *cm_score* curves are very close (Fig. 8d), which indicates that the mean isotope patterns are not evident to differentiate the noisy formulas.

5.4.3 Relative Contribution of Local Search and Multiconstraint Filtering

Here, the *Expected Rank No.* of the true formulas, which is calculated from the prediction results, is utilized as the performance metric to evaluate the relative contribution of Local Search and Multiconstraint Filtering. For comparison, the *Expected Rank No.* predicted by FFP and *MS_Enumerate* is also calculated. The results are summarized in Fig. 9.

Fig. 9 shows that in the range of (0~300u), the *Expected Rank No.* of the true formulas calculated by Local search and Multiconstraint Filtering are similar (1.759494 versus 1.50211), which implies that the contributions of these two techniques are almost same in FFP. In the range of (300~500u), Multiconstraint is more efficient than Local search (2.659259 versus 4.133333). However, Local Search makes more contribution in the ranges of (500~800u) and (800~2,000u). The figure also shows that due to these two techniques, FFP can significantly outperform the baseline method *MS_Enumerate* as the ion's mass increases.

6 Discussion

From the experimental results described in Section 5, we can see that FFP performs very well in predicting elemental component formulas of fragment ions in Q-TOF spectra. Furthermore, by predicting the formulas for a series of ions, FFP provides helpful information to refine the peptide sequencing by *de novo* in three ways. First of all, FFP can distinguish different ions with the same mass. As we know, for medium resolution MS/MS data, some acid amino residues are difficult to be differentiated by their mass, which causes confusion in *de novo*. For example, the

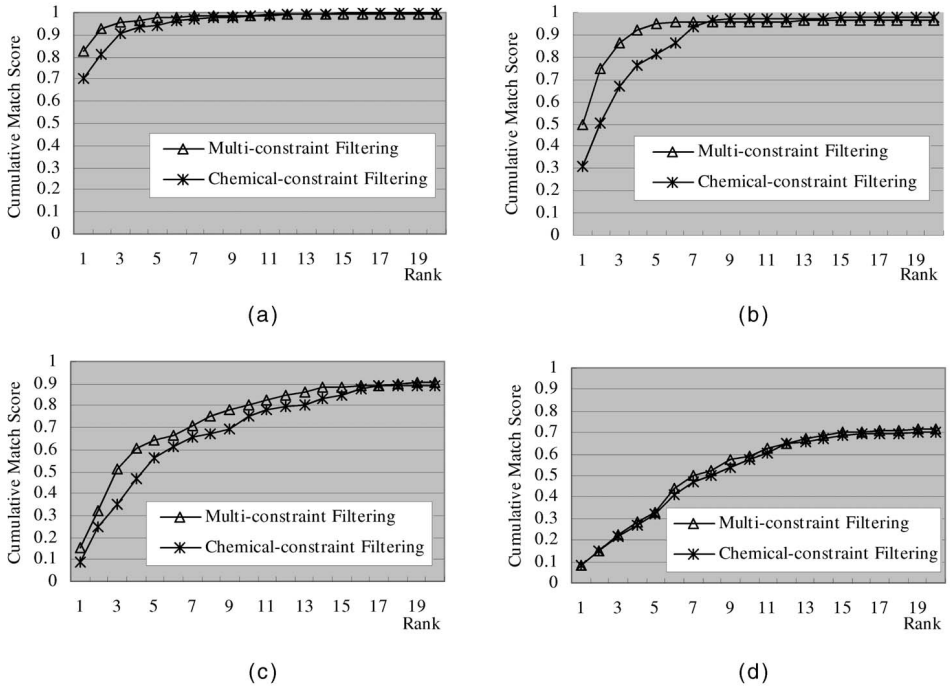

Fig. 8. Efficiency of Multiconstraint Filtering and Chemical-constraint Filtering. (a), (b), (c), and (d) depict the performances of Multiconstraint Filtering and Chemical-constraint Filtering in mass range of (0~300u), (300~500u), (500~800u), and (800~2,000u), respectively.

immonium ion of *oxidized Met* ($[C_4H_9NOS + H]^+$) and *Phe* ($[C_8H_9N + H]^+$) have almost the same mass, 120.0483 and 120.0813, respectively. However, *oxidized Met* and *Phe* have very different *tIPVs* of (120.0483, 0.05837, 0.04787) and (120.0831, 0.09480, 0.00398), respectively, and they can be differentiated by FFP from the *eIPVs* in spectrum. Therefore, the component prediction of FFP can refine the interpretation of individual ion peaks for *de novo*.

Second, the *de novo* approach directly derives a (partial) sequence from spectrum by computing the mass matches, i.e., the matches between the masses of amino acids and the distances of peaks in spectrum. Due to the measurement errors, there are a massive number of matches between two peaks and amino acid sequences. Since a peak corresponds to an ion with a component formula, when the difference of the two ions' formulas predicted by FFP matches the amino acid's formula, the constructed sequences by *de novo* will be more reliable.

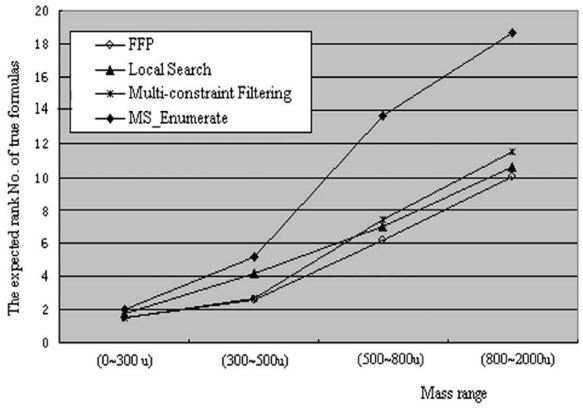

Fig. 9. Relative efficiency of four methods measured by *Expected Rank No.*

For example, assume that p_1, p_2, and p_3 are three peaks within a close range in a spectrum, and the mass distances between p_1 and p_3, and between p_2 and p_3 match amino acids A_1 and A_2, respectively. Let us further assume that after FFP predicts the formulas of p_1, p_2, and p_3, the difference between formulas of p_1 and p_3 matches the formula of A_1, while the difference between formulas of p_2 and p_3 does not match the formula of A_2. In this case, we can conclude that the subsequence conducted by p_1 and p_3 is more reliable than that constructed by p_2 and p_3.

Here is an example to further illustrate this point. Fig. 10 shows one spectrum of the peptide CCTESLVNR, in which both the two amino acids "C" are carbamidomethylated, and Table 4 depicts the predictions for isotope peaks of ions selected from the spectrum by FFP. From Table 4, we can observe that FFP predicts formulas for 16 ions with a high reliability. These predictions support a series of matches of amino acids. In this way, FFP can provide significant confidence to the true candidate sequence and, hence, will improve the reliability of *de novo* sequencing and reduce the computing effort of *de novo*. In the future work, we will implement FFP in a *de novo* algorithm, and show its efficiency in peptide sequencing.

Finally, FFP can predict certain formulas for "unknown" ions to improve the confidence of a candidate peptide sequence. For example, consider the fourteenth ion with a mass of 231.0874 in Table 4. If the main ion types introduced in Section 2 are employed, the 14th ion is unknown. However, FFP makes a prediction matching the internal ion of "TE," which identifies this ion and improve the confidence of the sequence "CCTESLVNR."

7 Conclusions

In this paper, a novel method, FFP (Fragment ion Formula Prediction), is proposed to predict component formulas of

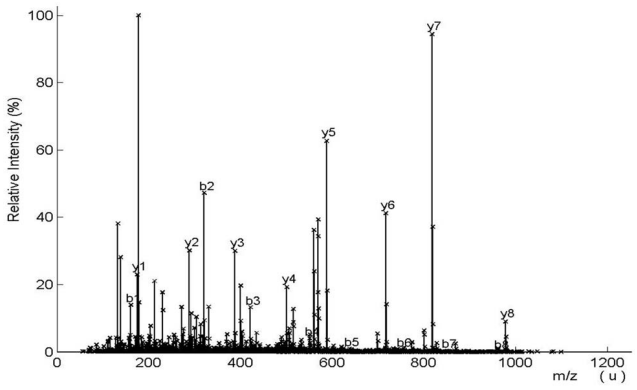

Fig. 10. MS/MS spectrum of the peptide CCTESLVNR with major peaks corresponding to y or b-type ions.

fragment ions based on the isotope patterns in the tandem spectrum. The main contributions of this paper are summarized as follows: First, FFP uses a "local search" model to generate and search candidate formulas. This ensures that FFP can predict elemental components of ions within a larger mass range with a low computation complexity. Second, FFP uses a multiconstraint filtering, which includes the mean isotope patterns, chemical constraint filtering and cross-validation, to discard as many invalid and improbable candidates as possible. In turn, these filtering can improve the prediction reliability. The experimental results show that FFP can not only predict formulas with a higher reliability, but also enjoy a lower computation complexity than the other methods. Third, FFP is also shown to help *de novo* to refine peptide sequencing in protein identification. In the future work, we will improve FFP's performance in the large mass ranges, and implement FFP in a *de novo* algorithm to improve its efficiency in peptide sequencing.

APPENDIX A

FITTING OF MASS-DEPENDENT MEAN ISOTOPE PATTERNS

As described in Section 4.3, the mean and standard deviations of T_1 within interval $\delta = 1\mathrm{u}$ are calculated quantitatively for each category, and fitted by polynomial curves. In category S^0, for example, the mean of T_1 can be extrapolated by the linear fit $mean_1$:

$$mean_1 = 0.0005669 \times M - 0.001243,$$

and mass-dependent standard deviations are given by

$$plus_1 = 0.000019 \times M + 0.006585,$$
$$minus_1 = 0.000017 \times M + 0.006191,$$

where $plus_1$ and $minus_1$ describing the oscillation of T_1, respectively.

Similarly, distribution of T_2 can be represented by $mean_2$, $plus_2$, and $minus_2$, which are given as follows:

$$mean_2 = 0.0000001663 \times M^2 + 0.00002253 \times M + 0.0008897,$$
$$minus_2 = 0.000000009509 \times M^2 + 0.000001197 \times M - 0.0001284,$$
$$plus_2 = 0.00000001132 \times M^2 + 0.0000002527 \times M + 0.0001947.$$

TABLE 4
The Prediction of Fragment Ions by FFP from One Spectrum of the Peptide CCTESLVNR

No	Ion type	Formula	Ion Mass	Rank (FFP)	Peak Relative Intensity %
1	b_9^{++}	$C_{43}H_{74}N_{15}O_{16}S_2$	1120.48918	>20	36.1643
2	y_8	$C_{38}H_{68}N_{13}O_{15}S_1$	978.4698	7	9.0032
3	y_7	$C_{33}H_{60}N_{11}O_{13}$	818.4371	5	94.3472
4	y_7-NH_3	$C_{33}H_{57}N_{10}O_{13}$	801.4130	5	5.1427
5	y_7-H_2O^{++}	$C_{33}H_{58}N_{11}O_{12}$	800.42257	1	19.647
6	y_6	$C_{29}H_{53}N_{10}O_{11}$	717.3854	1	41.1416
7	y_6-H_2O	$C_{29}H_{51}N_{10}O_{10}$	699.3696	7	5.4598
8	y_5	$C_{24}H_{46}N_9O_8$	588.3403	3	62.6775
9	y_4	$C_{21}H_{41}N_8O_6$	501.3098	2	19.1921
10	y_3	$C_{15}H_{30}N_7O_5$	388.2260	1	29.9462
11	y_3-NH_3	$C_{15}H_{27}N_6O_5$	371.1984	1	5.1841
12	b_2	$C_{10}H_{17}N_4O_4S_2$	321.0693	1	47.1805
13	y_2	$C_{10}H_{21}N_6O_4$	289.163	1	30.1255
14	'TE'-internal[a]	$C_9H_{15}N_2O_5$	231.0874	1	12.3397
15	$c1$	$C_5H_{12}N_3O_2S_1$	178.071	1	100
16	a_1	$C_4H_9N_2O_1S_1$	133.0515	1	38.1084
17	a_1-NH_3	$C_4H_6N_1O_1S_1$	116.0272	2	4.0949

[a] *Double backbone cleavage gives rise to internal fragments. Usually, these are formed by a combination of b type and y type cleavage.*

The distributions of the other five categories are also calculated. It is observed that the mean and standard deviations of T_1 are similar as that of S^0, while in polynomial expression, power coefficients of T_2 are different from that of S^0. The mean fit curves of T_2 are illustrated in Fig. 3 in Section 4.3.1. Specifically, we can use polynomial $mean(k) = c_2(k) \times M^2 + c_1(k) \times M + c_0(k)$ to describe the mean fit function of T_2 in S^k, for $k = 0 \sim 5$. These six categories have almost the same coefficients of $c_2(k)$ and $c_1(k)$, $k = 0 \sim 5$, but different coefficients of $c_0(k)$, which are 0.0008897, 0.04456, 0.09001, 0.1389, 0.1878, and 0.1932, respectively. In other words, with respect to $c_0(0)$, $c_0(k)$ increase as $c_0(k) \simeq c_0(0) + (0.0426 + 0.001 \times k) \times k, k = 1 \sim 5$, which corresponds to the fact that sulfur has an abundant isotope ^{+2}S of frequency of 0.04210.

APPENDIX B

FITTING OF MASS-INDEPENDENT MEAN ISOTOPE PATTERNS

Similarly, to reveal mass-independent relationship between T_1 and T_2, we calculate the mean and standard deviations fittings of T_2 with respect to T_1 in category S^0, which is given as follows:

$$mean(T_2) = 0.50003 \times T_1^2 + 0.04845 \times T_1 - 0.00072,$$
$$minus(T_2) = -0.000581 \times T_1^2 + 0.006891 \times T_1 + 0.000396$$
$$plus_2(T_2) = -0.001357 \times T_1^2 + 0.008658 \times T_1 - 0.000146.$$

For the other five categories, it also can be observed similar mean and standard deviations as that of S^0. In terms of polynomial $meanT_2(k) = c_2(k) \times T_1^2 + c_1(k) \times T_1 + c_0(k)$, for $k = 0 \sim 5$, all curves of mean T_2 have similar curvature (i.e., $c_2(k) \simeq 0.5$), while different coefficients of $c_0(k) = -0.00072, 0.04295, 0.08679, 0.1300, 0.1734, k = 0 \sim 4$, respectively.

ACKNOWLEDGMENTS

The authors thank Dr. R. Johnson for kindly providing Q-TOF data. The authors would also like to thank Chunjie Zhang from the University of Western Ontario, Wantao Ying from the Department of Genomics and Proteomics, Beijing Institute of Radiation Medicine, and Ruixiang Sun, Yan Fu, and Xiaobaio Wang from the Institute of Computing Technology, Chinese Academy of Sciences, for their insightful discussions. The authors would also like to thank Bin Pang and Bin Wu for reading the draft of the paper. This work was funded by the National Key Basic Research and Development Program (973) of China under Grant No. 2002CB713807.

REFERENCES

[1] J.R. Yates, III, P. Griffin, L. Hood, and J. Zhou, "Computer Aided Interpretation of Low Energy MS/MS Mass Spectra of Peptides," *Techniques in Protein Chemistry II,* pp. 477-485, 1991.

[2] J.K. Eng, A.L. McCormack, and J.R. Yates, III, "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database," *J. Am. Soc. Mass Spectrometry,* vol. 5, no. 11, pp. 976-989, Nov. 1994.

[3] M. Mann and M. Wilm, "Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags," *Analytical Chemistry,* vol. 66, no. 24, pp. 4390-4399, Dec. 1994.

[4] V. Dancik, T.A. Addona, K.R. Clauser, J.E. Vath, and P.A. Pevzner, "De Novo Peptide Sequencing via Tandem Mass Spectrometry," *J. Computational Biology,* vol. 6, nos. 3-4, pp. 327-342, Fall-Winter 1999.

[5] J.A. Taylor and R.S. Johnson, "Implementation and Uses of Automated De Novo Peptide Sequencing by Tandem Mass Spectrometry," *Analytical Chemistry,* vol. 73, no. 11, pp. 2594-2604, June 2001.

[6] T. Chen, M.Y. Kao, M. Tepel, J. Rush, and G.M. Church, "A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry," *J. Computational Biology,* vol. 8, no. 3, pp. 325-337, June 2001.

[7] B. Ma, K.Z. Zhang, C. Hendrie, C.Z. Liang, M. Li, A. Doherty-Kirby, and G. Lajoie, "PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS," *Rapid Comm. in Mass Spectrometry,* vol. 17, no. 20, pp. 2337-2342, Oct. 2003.

[8] J.I. Brauman, "Least Squares Analysis and Simplification of Multi-Isotope Mass Spectra," *Analytical Chemistry,* vol. 38, no. 4, pp. 607-610, Apr. 1966.

[9] R.W. Rozett, "FIMS: Least-Squares Fitted Fractional Abundance of Isotopes," Quantum Chemistry Program Exchange (QCPE) 11, 271, 1975.

[10] C.J. Robinson and G.L. Cook, "High Ionizing Voltage, Low Resolution Mass Spectrometric Analysis of Gas Oil Aromatic Fractions," Quantum Chemistry Program Exchange (QCPE) 11, 266, 1974.

[11] J.E. Evans and N.B. Jurinski, "Program ELAL: An Interactive Minicomputer Based Elemental Analysis of Low and Medium Resolution Mass Spectra," *Analytical Chemistry,* vol. 47, no. 6, pp. 961-963, May 1975.

[12] B.D. Dombek, J. Lowther, and E. Carberry, "A Computer Program for the Prediction of Mass Spectrum Isotope Peaks," *J. Chemical Education,* vol. 48, no. 11, p. 729, Nov. 1971.

[13] B.D. Dombek, J. Lowther, and E. Carberry, "IPPKS: The Prediction of Mass Spectrum Isotope Peaks," Quantum Chemistry Program Exchange (QCPE) 11, 294, 1976.

[14] H.M. Bell, "Computer Analysis of Isotope Clusters in Mass Spectrometry," *J. Chemical Education,* vol. 51, no. 8, p. 548, Aug. 1974.

[15] P.E. Kavanagh, "Program for Elemental Analysis Using Low or Medium Resolution Mass Spectra," *Org. Mass Spectrom.,* vol. 15, pp. 334-335, 1980.

[16] A. Tenhosaari, "Computer-Assisted Composition Analysis of Unknown Compounds by Simultaneous Analysis of the Intensity Ratio of Isotope Patterns of the Molecular Ion and Daughter Ions in Low-Resolution Mass Spectra," *Org. Mass Spectrom.,* vol. 23, pp. 236-239, 1988.

[17] C.L. Do Lago and C. Kascheres, "New Method of Isotope Pattern Analysis," *J. Computational Chemistry,* vol. 15, pp. 149-155, 1991.

[18] P. Roepstorff and J. Fohlman, "Proposal for a Common Nomenclature for Sequence Ions in Mass Spectra of Peptides," *Biomedical Mass Spectrometry,* vol. 11, no. 11, p. 601, Nov. 1984.

[19] R.S. Johnson, S.A. Martin, K. Biemann, J.T. Stults, and J.T. Watson, "Novel Fragmentation Process of Peptides by Collision-Induced Decomposition in a Tandem Mass Spectrometer: Differentiation of Leucine and Isoleucine," *Analytical Chemistry,* vol. 59, no. 21, pp. 2621-2625, Nov. 1987.

[20] A.M. Falick, W.M. Hines, K.F. Medzihradszky, M.A. Baldwin, and B.W. Gibson, "Low-Mass Ions Produced from Peptides by High-Energy Collision-Induced Dissociation in Tandem Mass Spectrometry," *J. Am. Soc. Mass Spectrometry,* vol. 4, no. 11, pp. 882-893, 1993.

[21] I.A. Papayannopoulos, "The Interpretation of Collision-Induced Dissociation Tandem Mass Spectra of Peptides," *Mass Spectrometry Rev.,* vol. 14, no. 1, pp. 49-73, Jan. 1995.

[22] J.C. Rouse, W. Yu, and S.A. Martin, "A Comparison of the Peptide Fragmentation Obtained from a Reflector Matrix-Assisted Laser Desorption-Ionization Time-of-Flight and a Tandem Four Sector Mass Spectrometer," *J. Am. Soc. Mass Spectrometry,* vol. 6, no. 9, pp. 822-835, Sept. 1995.

[23] P. Pevzner, *Computational Molecular Biology,* MIT Press, pp. 229-249, 2000.

[24] Z.D. Sharp, "Introduction to Stable Isotope Geochemistry," http://epswww.unm.edu/facstaff/zsharp/bio1.htm, 2005.

[25] S. Gay, P.-A. Binz, D.F. Hochstrasser, and R.D. Appel, "Modeling Peptide Mass Fingerprinting Data Using the Atomic Composition of Peptides," *Electrophoresis,* vol. 20, pp. 3527-3534, 1999.
[26] S. Gay, P.-A. Binz, D.F. Hochstrasser, and R.D. Appel, "Peptide Mass Fingerprinting Peak Intensity Prediction: Extracting Knowledge from Spectra," *Protemics,* vol. 2, pp. 1374-1391, 2002.
[27] R. Gras, M. Müller, E. Gasteiger, S. Gay, P.-A. Binz, W. Bienvenut, C. Hoogland, J.-C. Sanchez, A. Bairoch, D.F. Hochstrasser, and R.D. Appel, "Improving Protein Identification Form Peptide Mass Fingerprinting through a Parameterized Multi-Level Scoring Algorithm and an Optimized Peak Detection," *Electrophoresis,* vol. 20, pp. 3535-3550, 1999.
[28] M. Wehofasky and R. Hoffman, "Isotopic Deconvolution of Matrix-Assisted Laser Desorption/Ionization Mass Spectra for Substance-Class Specific Analysis of Complex Samples," *European J. Mass Spectrometry,* vol. 7, pp. 39-46, 2001.

Jingfen Zhang received the BS degree in mathematics from Central China Normal University in 1993 and the ME degree in computer science from HuaZhong University of Science and Technology in 1998, respectively. Since 2001, she has been working toward her PhD degree at the Institute of Computing Technology, Chinese Academy of Sciences, P.R. China. Her current interests are in the areas of DNA fragments assembly, protein interaction network, and protein identification.

Wen Gao received a PhD degree in computer science from Harbin Institute of Technology in 1988, and another PhD degree in electronics engineering from the University of Tokyo, Japan in 1991. He is a professor at the Institute of Computing Technology, Chinese Academy of Sciences, the chief editor of the *Chinese Journal of Computers*, honorary professor of computer science, City University of Hong Kong, etc. His research interests include multimodal user interface, multimedia data compression, computer vision, and artificial intelligence. He has published more than 150 papers and books. He is a member of the IEEE.

Jinjin Cai graduated from Beijing Jiaotong University in 2002, with a bachelor's degree in computer science and technology. Since 2003, she has been working toward her master's degree at the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, P.R. China. Her current interests are in the area of *De novo* sequencing, posttranslational modifications, and machine learning algorithms.

Simin He received the BE and PhD degrees from Tsinghua University, China. He is currently with the Institute of Computing Technologies, Chinese Academy of Sciences, China, as a research associate professor. His research centers on the design and analysis of high performance algorithms in bioinformatics and networking. He is a member of the IEEE.

Rong Zeng received the PhD degree from the Shanghai Institute of Biochemistry in 2000. She is a professor and deputy director of the Research Center for Proteome Analysis at the Institute of Biochemistry and Cell Biology of Shanghai Institutes for Biological Sciences of Chinese Academy of Sciences. Her research interests are proteomics related to human diseases, biological mass spectrometry methodology, and development of high-throughput platform for proteomics.

Runsheng Chen is a professor in the Graduate School at the University of Sciences and Technology of China, the Vice Editor-in-Chief of *ACTA Biophysics Sinica*, the standing editor of *Progress of Biochemistry and Biophysics* (in Chinese). His current research interests include structure, function and evolution of "junk" DNA, human genome informatics, cryptography of DNA and new method of DNA sequence analysis, comparative genomics, system's biology and gene network, molecular modeling and design of protein, and RNA and drugs. In 1996, he received the "Kotani Prize" during the 15th International CODATA Conference, Tsukuba, Japan.