CHARACTERIZATION OF MULTI-CHARGE MASS SPECTRA FOR PEPTIDE SEQUENCING

KET FAH CHONG, KANG NING,[†] HON WAI LEONG
Department of Computer Science, National University of Singapore,
3 Science Drive 2, Singapore 117543

PAVEL PEVZNER
Department of Computer Science & Engineering, University of California, San Diego,
La Jolla, CA 92093-0114

Sequencing of peptide sequences using tandem mass spectrometry data is an important and challenging problem in proteomics. In this paper, we address the problem of peptide sequencing for multi-charge spectra. Most peptide sequencing algorithms currently handle spectra of charge 1 or 2 and have not been designed to handle higher-charge spectra. We give a characterization of multi-charge spectra by generalizing existing models. Using these new models, we have analyzed spectra with charges 1-5 from the GPM [8] datasets. Our analysis shows that higher charge peaks are present and they contribute significantly to prediction of the complete peptide. They also help to explain why existing algorithms do not perform well on multi-charge spectra. We also propose a new *de novo* algorithm for dealing with multi-charge spectra based on the new models. Experimental results show that it performs well on all spectra, especially so for multi-charge spectra.

1 Introduction

Proteomics is the large-scale study of proteins, particularly their sequences, structures and functions. In proteomics, the identification of the protein sequences is very important, and peptide sequencing is essential to the identification of the proteins. Currently, peptide sequencing is largely done by tandem mass spectrometry. The analysis of the spectrum data is a non-trivial problem. This is in part because the spectrum obtained from MS/MS usually contains lots of noise, which do not belong to the peptide, but introduced because of the impurity of the peptide, and the inaccuracy of the machines. The problem becomes more difficult since for one peptide sequence, not all of its subsequences have the corresponding ions in the spectrum.

Deducing peptide sequences from raw MS/MS data is slow and tedious when done manually. Instead, the most popular approach is to do a *database search* of known peptide sequences with the un-interpreted experimental MS/MS data. A number of such database search algorithms have been described, the most popular being Mascot [1] and Sequest [2]. These methods are effective but often give false positives or incorrect identifications. Searching databases with masses and partial sequences (sequence tags) derived from MS/MS data give more reliable results [3]. For unknown peptides, *de novo* sequencing [4-7] is used in order to predict sequences or partial sequences. However, the

[†] Contact: ningkang@comp.nus.edu.sg.

prediction of peptide sequences from MS/MS spectra is dependent on the quality of the data, and this result in good predicted sequences only for very high quality data.

This paper focuses on the important issue of the *amount of charge* on the ions in the spectra, particularly multi-charge spectra (charges 3 to 5). In the case of an ESI/MALDI source, the parent ion and many fragments may have multiple charge units assigned to them. Multi-charged spectra (with charges up to 5) are available from the GPM [8] web-site. Current *de novo* methods work well on good quality spectra of charges 1 and 2. However, they do not work well on spectra with charges 3 to 5 since they do not explicitly handle multi-charge ions (one notable exception is PEAKS [6] which does conversion of multi-charge peaks to their singly-charged equivalent before sequencing). Lutefisk [7] works with singly-charged ion only, while Sherenga [4] and PepNovo [5] works with singly- and doubly-charged ions. Therefore, it is not surprising that some of the higher charged peaks are mis-annotated by these methods leading to lower accuracy.

In this paper, we propose *a generalized model* that better describes multi-charge spectra (multi-charge to mean charge $\geq$ 3) and *quality measures* for multi-charge spectra based on the new model. Our evaluation of multi-charged spectra from GPM with the new model shows that the theoretically attainable accuracy increases as we consider higher charge ions meaning that multi-charge ions are significant. In addition, we show that *any* algorithm that considers *only* charge 1 or 2 ions will suffer from low prediction accuracy. Our experiments show that the accuracy[1] of these methods on multi-charge spectra is very low (less that 35%), and this accuracy decrease as the charge of the spectra increases (for charge 4 spectra, the accuracy of Lutefisk is less than 7%).

We also proposed a *simple de novo* sequencing algorithm called *GBST* (greedy best strong tag) that considers higher charge ions based on our new model. Experimental results on GPM spectra show that *GBST* outperforms many of the other *de novo* algorithms on spectrum data with charge of 3 or more.

2 Modeling of Multi-Charge Spectra

Consider an experimental mass spectrum $S = \{p_1, p_2, \ldots p_n\}$ of maximum charge α that is produced by an MS/MS experiment on a peptide $\rho = (a_1 a_2 \ldots a_l)$, where a_j is the j^{th} amino acid in the sequence. The parent mass of the peptide ρ is given by $M = m(\rho) = \sum_{j=1}^{l} m(a_j)$. Consider a peptide prefix fragment $\rho_k = (a_1 a_2 \ldots a_k)$, for $k \leq n$, that has mass $m(\rho_k) = \sum_{j=1}^{k} m(a_j)$. Suffix masses are defined similarly. Then, the set of all possible prefixes and suffixes of a peptide forms the "full ladder" of the peptide. Let $TS_0(\rho) = \{m(\rho_1), m(\rho_2), \ldots, m(\rho_n)\}$ to be the set of all possible (*uncharged*) prefix fragment masses of the peptide ρ. A peak in the experimental spectrum S then corresponds to the detection of some charged prefix or suffix peptide fragment that results from peptide fragmentation in the mass spectrometer. Each peak p_i in the experimental spectrum S is described by its *intensity*(p_i) and *mass-to-charge ratio* $mz(p_i)$.

[1] The accuracy measure we use is defined in Section 3.3.

However fragmentation is usually not very clean and other types of fragments occur. Noise and contaminants can also cause a peak in the experimental spectrum. In peptide sequencing, we are given an experimental spectrum with true peaks and noise and the problem is to try to determine the original peptide ρ that produced the spectrum.

The Theoretical Spectrum for a Known Peptide: To theoretically characterize a multi-charge spectrum of a known peptide ρ, we consider the set of all *possible* true peaks that correspond to prefix fragments (N-terminal ions) and suffix fragments (C-terminal ions). Each peak p can be characterized by the ion-type, that is specified by $(z, t, h) \in (\Delta_z \times \Delta_t \times \Delta_h)$, where z is the charge of the ion, t is the basic ion-type, and h is the neutral loss incurred by the ion. In this paper, we restrict our attention to the set of ion-types $\Delta = (\Delta_z \times \Delta_t \times \Delta_h)$, where $\Delta_z = \{1, 2, \ldots, \alpha\}$, $\Delta_t = \{a\text{-ion}, b\text{-ion}, y\text{-ion}\}$ and $\Delta_h = \{\varnothing, -H_2O, -NH_3\}$.[2] The (z, t, h)-ion of the peptide fragment q (prefix or suffix fragment) will produced an observed peak p_i in the experimental spectrum S that has a mass-to-charge ratio of $mz(p)$, that can be computed using a shifting function, *Shift*, defined as follows:

$$m(q) = Shift(p_i, (z, t, h)) = mz(p_i) \cdot z + (\delta(t) + \delta(h)) - (z - 1) \qquad (1)$$

where $\delta(t)$ and $\delta(h)$ are the mass differences associated with the ion-type t and the neutral-loss h, respectively. We say that peak p_i is a *support peak* for the fragment q and *has ion-type* (z, t, h) and we say that the fragment q is *explained* by the peak p_i.

We define the *theoretical spectrum* $TS_\alpha^\alpha(\rho)$ *for* ρ *for maximum charge* α to be the set of all *possible* observed peaks that may be present in an experimental spectrum for the peptide ρ with maximum charge α. More precisely, $TS_\alpha^\alpha(\rho)$ = { p : p is an observed peak for the (z, t, h)-ion of peptide prefix fragment ρ_k, for all $(z, t, h) \in \Delta$ and $k=1,\ldots,n$}.

Extended Spectrum: Conversely, the *real* peaks in an experimental spectrum $S = \{p_1, p_2, \ldots p_n\}$ of maximum charge α, may have come from different ion-type of different fragments (may be prefix or suffix fragment, depending on the ion-type). We do not know, a priori, the ion-type $(z, t, h) \in \Delta$ of each peak p_i. Therefore, we "extend" each peak p_i by generating a set of $|\Delta|$ pseudo-peaks (or guesses), one for each of the different ion-types $(z, t, h) \in \Delta$. More precisely, in the extended spectrum S_α^α, for each peak $p_i \in S$ and ion-type $(z, t, h) \in \Delta$, we generate a pseudo-peak, denoted by $(p_i, (z, t, h))$, with an "assumed" (uncharged) fragment mass computed using the *Shift* function (1). At most one of these pseudo-peaks is a real peak, while the others are "introduced" noise.

We always express a fragment mass in experimental spectrum using its *PRM* (prefix residue mass) representation, which is the mass of the prefix fragment. For suffix fragments (y-ions), we use its corresponding prefix fragment. Mathematically, for a fragment q with mass $m(q)$, we define $PRM(q) = m(q)$ if q is a prefix fragment ({b-ion}); and we define $PRM(q) = M - m(q)$ if q is a suffix fragment ({y-ion}). By calculating the *PRM* for all fragments, we can treat all fragments masses uniformly.

[2] The definitions and results in this paper also apply to any set of ion-types considered.

We illustrate the extended spectrum with an example shown in Figure 1. For simplicity, we only consider ion-types Δ_t = {*b*-ions, *y*-ions} and Δ_h={Ø}. Given a peptide ρ = GAPWN, with parent mass $M = m(\rho) = 525.2$, and an experimental spectrum S = {113.6, 412.2, 487.2} with maximum charge 2. The first peak "113.6" is a (2, *b*-ion, Ø)-ion of the prefix fragment GAP; the peak 412.2 is a (1, *b*-ion, Ø))-ion of the prefix fragment GAPW; and "487.2" is a (1, *y*-ion, Ø)-ion for the fragment G. In Figure 1(a), only charge 1 is considered and S_1^2 = {112, 430, 411, 132, 486, 57}. The entries in the table are the *PRM* values. For example, the possible fragment masses of 112 and 430 correspond to the extension of the first peak for ion-types (1, *b*-ion, Ø) and (1, *y*-ion, Ø), respectively. However, if charge 2 is also considered, then S_2^2 = {112, 430, 225, 31, 411, 132, 486, 57} as shown in Figure 1(b).

Duality between Extended Spectrum and Theoretical Spectrum: We now describe a duality relationship between the extended spectrum S_β^α and the theoretical spectrum $TS_\beta^\alpha(\rho)$. Given an experimental spectrum S of a known peptide ρ, the set $RP_\alpha^\alpha(S,\rho)$ of *real peaks* in the spectrum S is given by:

$$RP_\alpha^\alpha(S,\rho) = TS_\alpha^\alpha(\rho) \cap S \tag{2}$$

The set $EF_\alpha^\alpha(S,\rho)$ of *explained fragments* in the peptide ρ, namely fragments that can be "explained" by the presence of support peak or pseudo-peak in S_α^α, is given by:

$$EF_\alpha^\alpha(S,\rho) = TS_0(\rho) \cap PRM(S_\alpha^\alpha)\,. \tag{3}$$

In the set $RP_\alpha^\alpha(S,\rho)$, there may be several real peaks that are support peaks for the same fragment . Similarly, in the set $EF_\alpha^\alpha(S,\rho)$, there may be multiple pseudo-peaks in S, that helps to "explain" the same fragment. Indeed, we have the following duality theorem:

Duality Theorem: Given an experimental spectrum S of a known peptide ρ, we have

$$EF_\alpha^\alpha(S,\rho) = PRM(Shift(RP_\alpha^\alpha(S,\rho))) \tag{4}$$

Modelling Current Algorithms: To take into account the fact that some algorithms consider only ion-types of charge up to β (usually β = 2), we extend the definition to $TS_\beta^\alpha(\rho)$ which is defined to be the subset of $TS_\alpha^\alpha(\rho)$ for which the charge $z \in \{1,2,\ldots,\beta\}$. The case β=1 reflects the assumption that all peaks are of charge 1, and makes use of the extended spectrum S_1^α. Algorithms such as PepNovo and Lutefisk works with a subset of the extended spectrum S_2^α, even for spectra with charge $\alpha > 2$. In general, $TS_\beta^\alpha(\rho)$ does not account for peaks that correspond to ion-types with higher charges $z=\beta+1, \ldots, \alpha$. Of course, the more charge we take into account, the more accurate will be the accuracy that can be attained since $TS_1^\alpha(\rho) \subseteq TS_2^\alpha(\rho) \ldots \subseteq TS_\alpha^\alpha(\rho)$.

The Extended Spectrum Graph: We also introduce an extended spectrum graph, denoted by $G_d(S_\beta^\alpha)$, where d is the "connectivity". Each vertex v in this graph represents a pseudo-peak (p_i, (z, t, h)) in the extended spectrum S_β^α, namely, the (z, t, h)-

ions for the peak p_i. Thus $v = (p_i, (z, t, h))$. Each vertex represents a possible peptide fragment mass given by $PRM(Shift(p_j, (z, t, h)))$. Two special vertices are added – the start vertex v_0 corresponding to the empty fragment with mass 0 and the end vertex v_M corresponding to the parent mass M.

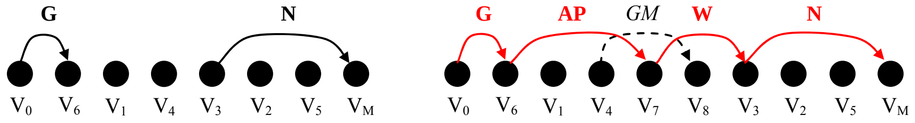
In the "standard" spectrum graph, we have a directed edge (u, v) from vertex u to vertex v if $PRM(v)$ is larger than $PRM(u)$ by the mass of a single amino acid. In the extended spectrum graph of connectivity d, $G_d(S_\beta^\alpha)$, we extend the edge definition to mean "*a directed path of no more than d amino acids*". Thus, we connect vertex u and vertex v by a directed edge (u, v) if the $PRM(v)$ is larger than $PRM(u)$ by the total mass of d' amino acids, where $d' \leq d$. In this case, we say that the edge (u, v) is connected by a path of length up to d amino acids. Note that the number of possible paths to be searched is 20^d and increased exponentially with d. We use d=2, unless otherwise stated.

z	***mz(p₁)*= 113.6**		***mz(p₂)*= 412.2**		***mz(p₃)*=487.2**	
	B	***Y***	***B***	***Y***	***B***	***Y***
1	V_1 112.6	V_2 430.6	V_3 411.2	V_4 132	V_5 486.2	V_6 57

(a) The spectrum S_1^2 (only B and Y ions considered)

z	***mz(p₁)*= 113.6**		***mz(p₂)*= 412.2**		***mz(p₃)*=487.2**	
	B	***Y***	***B***	***Y***	***B***	***Y***
2	V_7 225.2	V_8 318	- -	- -	- -	- -

(b) Extending the peaks for charge 2 ions.



(c) The spectrum graph $G_2(S_1^2)$

(d) The extended spectrum graph $G_2(S_2^2)$

Figure 1. Example of extended spectrum graph for mass spectrum regenerated from peptide GAPWN.

Two extended spectrum graphs (with connectivity d=2) are shown in Figure 1. The spectrum graph $G_2(S_1^2)$ is shown in Figure 1(c). We can see that only the edges (v_0, v_6) for amino acid G and (v_3, v_M) for amino acid N can be obtained. The subsequence APW is longer than 2 amino acids long and so $G_2(S_1^2)$ is unable to elucidate this information. By considering S_2^2 (in (a) and (b)), we obtain the graph $G_2(S_2^2)$ shown in (d). New edges can be obtained, edge (v_6, v_7) for path AP of length 2 amino acids and (v_7, v_3) for amino acid W. This gives a full path from v_0 to v_M and the full peptide can now be elucidated. However we also note that in $G_2(S_2^2)$, fictitous edges may also be introduced due to the introduction of more noise. One example is shown in (d) using dashed line for the fictitious edge (v_4, v_8). Many such fictituous edges can result in fictituous paths from v_b to v_e, thus giving a higher rate of false positives.

2.1. *Quality Measures for Evaluating Mass Spectra*

We have extensively analyzed many multi-charge spectra using our new characterization. In this exercise, we are only analyzing the quality of the spectra, and we are not doing sequencing or prediction. We define two *quality measures* of a multi-charge spectra

- $\mathit{Specificity}(\alpha, \beta) \quad = |TS^{\alpha}_{\beta}(\rho) \cap S| \;/\; |S| \quad = |RP^{\alpha}_{\beta}(S, \rho)| \;/\; |S|$
- $\mathit{Completeness}(\alpha, \beta) \quad = |TS_0(\rho) \cap PRM(S^{\alpha}_{\alpha})| \;/\; |\rho| \quad = |EF^{\alpha}_{\beta}(S, \rho)| \;/\; |\rho|$

Specificity measures the proportion of true peaks in the experimental spectrum *S*, and it can be also be consider the signal-to-noise ratio of *S*. However, for a given PRM, there may be multiple support peaks in $RP^{\alpha}_{\beta}(S, \rho)$, which lead to “double counting”. The completeness measure avoids this by computing the proportion of the fragment masses that are explained by support peaks. Multiple support peaks for the same fragments are not double-counted.

2.2. *Experimental Data and Analysis*

The data being used for analysis and experimentation is the Amethyst data set from GPM (Global Proteome Machine) [8] (obtainable from ftp://ftp.thegpm.org/quartz). The GPM system is an open-source system for analyzing, storing, and validating proteomics information derived from tandem mass spectrometry. The database was designed to store the minimum amount of information necessary to search and retrieve data obtained from the publicly available data analysis servers. One feature of the Amethyst dataset is that there are lots of multi-charge spectra (up to charge 5). These data are MS/MS spectra obtained from QSTAR mass spectrometers. Both MALDI and ESI sources were included.

Using the $G_d(S^{\alpha}_{\beta})$ extended spectrum graph model (with *d*=2), we have measured the average *Specificity*(*α*,*β*) and *Completeness*(*α*,*β*) on the enture Amethyst datasets from GPM using our extended spectra S^{α}_{β} for $1 \le \alpha \le 5$, and $1 \le \beta \le \alpha$. A mass tolerance of 0.5 Da is used for matching peak mass-to-charge ratios. All the data in the Amethyst dataset (12558 datasets in total, with 4000, 4561, 2483, 1175, 339 for charge 1, 2, 3, 4, 5, respectively) has been used for this purpose.

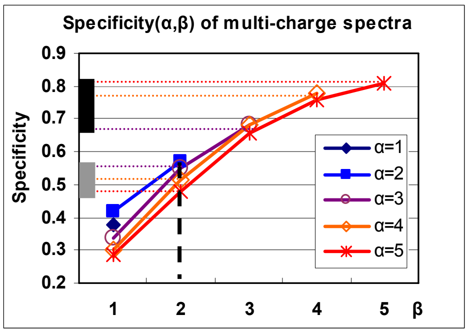


Figure 2. *Specificity(α,β)* of multi-charge spectra. Specificity increases as *β* increases. Most algorithms consider up to S^{α}_{2} (dashed black line). But considering S^{α}_{α} for spectra with $\alpha \ge 3$ improves the specificity (black line vs grey line).

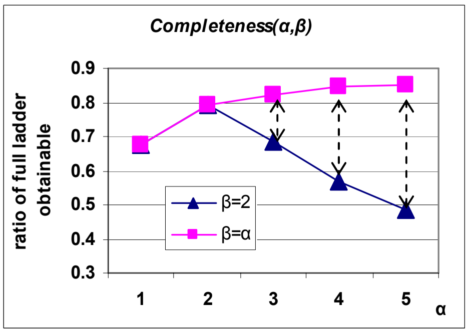


Figure 3. *Completeness*(*α*,*β*) of multi-charge spectra. We see that considering only S^{α}_{2} gives < 70% of the full ladder, which drops drastically as α gets bigger. On the other hand, considering S^{α}_{α} gives > 80% of full ladder.

The *Specificity*(α,β) results are showin in Figure 2. The results show that the GPM spectra contain an abundance of higher charged peaks in higher-charged spectra. For a fixed α, as β increases, the specificity increases – meaning that more true peaks are discovered. Furthermore, the increase is significant. For α=5, the specificity increases from 0.49 with β=2, to 0.81 when β=5. Algorithms that uses β =2 considering only charge 1 and 2 (like LuteFisk and PepNovo) are limited to specificity values of between 0.48 to 0.56, as indicated by the dashed vertical line at β=2.

The *Completeness*(α,β) results are showin in Figure 3. In this graph, we compare the *Completeness*(α,β) results for (a) using the full extended spectrum S_α^α versus (b) using only β=2, namely, S_2^α . Again, the results clearly show that significant improvement can be obtained by considering higher charge peaks. The disparity increases with α as seen from the widening gap indicated by the vertical arrows.

3 A Simple de Novo Algorithm for Multi-Charge Spectra

We now present a simple de novo peptide sequencing algorithm that takes into account multi-charged ion-types in the spectrum. Our main aim is to show that even with a simple algorithm, we can get improved results by considering multi-charged ions.

3.1 Strong Tags in the Multi-Charge Spectra

Tandem spectrum data analysis shows that peaks in many mass spectra can be grouped into closely-related sets, especially when the peptide is multi-charge. Within each set, the peaks can be interpreted as the same ion type (*b*-ions or *y*-ions), and the mass differences between "successive" peaks are such that they can form ladders (contiguous sequences). An example is shown in Figure 4, where we have computed the theoretical spectrum (the table) and the peaks from an experimental spectrum *S* are shown in bold. Several peaks are grouped together into contiguous sequences of *y*-ions and *b*-ions of charge 1.

This motivates us to call these contiguous sequences of strong ion-types (*b*-ions and *y*-ions of charge 1) "*strong tags*". More formally, they are defined as follows: Consider the extended spectrum graph, $G_1(S_1^\alpha)$, namely, only charge 1 ion-types. We define a *strong tag T of ion-type* (1, *t*, Ø) to be a maximal path ($v_1, v_2, ..., v_r$) in $G_1(S_1^\alpha)$ where each vertex $v_i \in T$ has the same ion-type (1, *t*, Ø) and each (v_i, v_{i+1}) is an edge in the graph , namely, their mass difference is the mass of one amino acid. (For our current algorithm, we consider only *b*-ions and *y*-ions, namely, *t* = *b*-ions or *y*-ions and strong tags must have at least 2 edges.)

Figure 5 shows the two strong tags obtained for the spectrum given in Figure 4.

To help the search for good strong tags, we define a *weight function* that is used to score vertices and strong tags. The weight of vertex $v_i \in G_1(S_\beta^\alpha)$ is defined as

$$w(v_i) = \frac{f_{support}(v_i) + f_{loss}(v_i) + f_{intensity}(v_i)}{f_{tolerance}(v_i)} \tag{5}$$

- $f_{support\text{-}ion}(v_i)$ is a function of the number of v_j, with v_j having a different ion-type as v_i, but for the same subsequence
- $f_{loss}(v_i)$ is a function of the number of v_j, with $(PRM(v_i) - PRM(v_j))$=17 or 18,
- $f_{intensity}(v_i)$ is a function of ($\log_{10}$(intensity of the peak for which v_i represent)),
- $f_{tolerance}(v_i) = (1/N)$ ($\sum$ | PRM(v_j) – PRM(v_i) – mass(a_k) |), where N is the total number of incoming and outgoing edges for v_i, and a_k is the amino acid for each edge (v_i, v_j) or (v_j, v_i).

For a strong tag T=(v_1, v_2, ..., v_r), the weight $W(T)$ of the strong tag T is just the sum of weight of the vertices in T, namely, $W(T) = \sum_{v_i \in T} w(v_i)$. Obviously, we are interested in finding a set of "best" strong tags, namely, tags that optimizes the weight $W(T)$. The spectrum graph $G_1(S_\beta^\alpha)$ is a DAG that may consist of several disjoint components. For each disjoint component C, we use a depth-first search (DFS) algorithm to compute a best strong tag for component C. We let *BST* denote the set of "best" strong tags from each of the components C in the spectrum graph.

bond	$^{+1}$y	$^{+1}$y*	$^{+1}$b	$^{+1}$b*
^{S}1	1807.0	1790.0	130.0	113.0
^{I}2	1693.9	1676.9	243.1	226.1
^{R}3	1537.8	1520.8	**399.2**	382.2
^{V}4	1438.8	1421.7	**498.3**	481.3
^{T}5	1337.7	1320.7	**599.3**	582.3
^{Q}6	1209.7	1192.6	**727.4**	710.4
^{K}7	1081.6	1064.5	**855.5**	838.5
^{S}8	994.5	977.5	942.5	925.5
^{Y}9	831.5	814.4	1105.6	1088.6
^{K}10	**703.4**	686.3	1233.7	1216.7
^{V}11	**604.3**	587.3	1332.8	1315.7
^{S}12	**517.3**	500.2	1419.8	1402.8
^{T}13	**416.2**	**399.2**	1520.8	1503.8
^{S}14	**329.2**	312.2	1607.9	1590.8
^{G}15	**272.2**	255.1	1664.9	1647.9
^{P}16	**175.1**	158.1	1761.9	1744.9

Figure 4. Theoretical spectrum for the peptide sequence "SIRVTQKSYKVSTSGPR", with parent mass of 1936.05 Da. "y" and "b" indicates y- and b-ions, "+1", "+2" indicates charge 1 and 2, and "*" indicates ammonia loss. Bold numbers are peaks present in experimental spectrum.

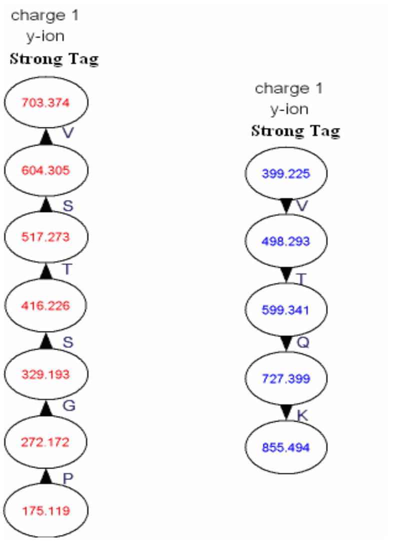


Figure 5. Example of strong tags in the spectrum graph for spectrum in Figure 4. There are 2 strong tags. Vertices (small ovals) represent fragment masses, and edges (small triangles) represent amino acids whose mass is the same as the mass difference of the vertice

3.2 The GBST Algorithm

We have developed a simple de novo peptide sequencing algorithm based on strong tag that we call the *Greedy Best Strong Tag (GBST)* algorithm which uses the strong tags in the spectrum graph. The *GBST* algorithm starts by computing the set *BST* of best strong tag as described in Section 3.1. After the BST is compute, the algorithm proceeds to find the best peptide sequence that can be obtained by "linking up" the strong tags in BST. We first build the strong tag graph $G_d(BST)$, where the vertices are the strong tags in *BST*, and we have an edge (u, v) from the *tail* vertex u of the tag T_u to the head vertex v of the tag T_v if $PRM(v)$ is larger than $PRM(u)$ by the total mass of d' amino acids, where $d' \le d$. (We use d=2.) Compared to the spectrum graph G, the strong tag graph $G_d(BST)$ is very small – only $|BST|$ vertices and the number of edges is also small since we only connect strong tags in a head-to-tail manner. A path in $G_d(BST)$ is called a strong tag path since the vertices are strong tags. For a strong tag path $P = (T_1, T_2, \ldots, T_q)$, we define the weight $W(P)$ of the path P to be the sum of the weight of the strong tags in P, namely, $W(P) = \sum_{T_i \in P} W(T_i)$. The final step in the *GBST* algorithm is to use a DFS algorithm to compute the "best" strong tag path from v_0 to v_M in the graph $G_d(BST)$.

3.3 Experiments on Algorithms

The experimental data are selected from GPM spectrum datasets [8]. We have selected spectra data with different characteristics (average peak intensities, charges, etc.) for analysis. We have applied our algorithm on these spectrum data. For these spectrums, we have also compared our results with those of the Lutefisk [7] and PepNovo [5]. For comparison of prediction results, we have defined two accuracy measures:

- *Sensitivity* = #correct / $|\rho|$
- *Specificity* = #correct / $|P|$

where #correct is the "*number of correctly sequenced amino acids*". The number of correctly sequence amino acids is computed as the *longest common subsequence* (lcs) of the correct peptide sequence ρ and the sequencing result P. *Sensitivity* indicates the quality of the sequence with respect to the correct peptide sequence and a high sensitivity means that the algorithm recovers a large portion of the correct peptide. For fair comparison with algorithms like PepNovo that only outputs the highest scoring tags (subsequences) we also use the specificity measure.

Table 1: Results of GBST, compared with Lutefisk and PepNovo on GPM spectra. Results show that GBST is generally comparable and sometimes better, especially for multi-charge spectra. (*based on +1 and +2 spectra).

Charge	**Number of spectrum**	**Lutefisk**	**PepNovo**	**GBST**
1	756	0.261 / 0.258	0.322 / 0.186	***0.296 / 0.315***
2	874	0.243 / 0.241	0.316 / 0.215	***0.297 / 0.326***
3	454	0.111 / 0.113	-	***0.262 / 0.285***
4	207	0.065 / 0.063	-	***0.190 / 0.222***
5	37	0 / 0	-	***0.165 / 0.223***
All	**2328**	**0.203 / 0.202**	**0.319 / 0.202***	***0.278 / 0.304***

In the experiments, we have only run PepNovo on spectra with charge 1 and 2 (since it only handles charge 1 and 2), and compared the results with our algorithm. In Table 1, the accuracy values are represented in a (specificity/sensitivity) format.

Experiments results show that our algorithm generally perform comparable to or better than Lutefisk [7] and PepNovo [5]. This is obvious for multi-charge spectra. The relatively high specificity accuracy of our algorithms shows that our sequencing results have high signal-to-noise ratio, which are comparable with results of Lutefisk and PepNovo. The higher sensitivity accuracy shows that our algorithms can sequence more correct amino acids than Lutefisk and PepNovo.

4 Conclusion

Multi-charge spectra have not been adequately addressed by many *de novo* sequencing algorithms. In this paper, we give a characterization of multi-charge spectra and use it to analyze multi-charge spectra from GPM. Our results clearly show why existing algorithms do not perform well on multi-charged spectra. We also present a simple *de novo* sequencing algorithm (called *GBST* algorithm) which makes use of this model to predict sequences of such spectra. Our *de novo* algorithm not only works well for multi-charge spectra, but it still performs well on singly-charges spectra.

Acknowledgements:

The authors would like to thank researchers at UCSD for providing us with experimental data and PepNovo program for our comparison. This work was partially supported by the National University of Singapore under grant R252-000-199-112.

References

1. D. N. Perkins, D. J. C. Pappin, D. M. Creasy and J. S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20:3551-3567, 1999.
2. J. K. Eng, A. L. McCormack and I. John R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *JASMS*, 5:976-989, 1994.
3. M. Mann and M. Wilm. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, 66:4390-4399, 1994.
4. V. Dancik, T. Addona, K. Clauser, J. Vath and P. Pevzner. De novo protein sequencing via tandem mass-spectrometry. *J. Comp. Biol.*, 6:327-341, 1999.
5. A. Frank and P. Pevzner. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Anal. Chem.*, 77:964 -973, 2005.
6. B. Ma, K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie. PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS. *Rapid Communications in Mass Spectrometry*, 17:2337-2342, 2003.
7. J. A. Taylor and R. S. Johnson. Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem.*, 73:2594-2604, 2001.
8. R.Craig, J.P. Cortens and RC. Beavis. Open source system for analyzing, validating, and storing protein identification data. *J Proteome Res.*, 3:1234-1242, 2004.