ALGORITHM FOR PEPTIDE SEQUENCING BY TANDEM MASS SPECTROMETRY BASED ON BETTER PREPROCESSING AND ANTI-SYMMETRIC COMPUTATIONAL MODEL

Kang Ning and Hon Wai Leong

Department of Computer Science, National University of Singapore
Block S15, 3 Science Drive 2, Singapore 117543
{ningkang, leonghw}@comp.nus.edu.sg

Peptide sequencing by tandem mass spectrometry is a very important, interesting, yet challenging problem in proteomics. This problem is extensively investigated by researchers recently, and the peptide sequencing results are becoming more and more accurate. However, many of these algorithms are using computational models based on some unverified assumptions. We believe that the investigation of the validity of these assumptions and related problems will lead to improvements in current algorithms. In this paper, we have first investigated peptide sequencing without preprocessing the spectrum, and we have shown that by introducing preprocessing on spectrum, peptide sequencing can be faster, easier and more accurate. We have then investigated one very important problem, the anti-symmetric problem in the peptide sequencing problem, and we have proved by experiments that model that simply ignore anti-symmetric or model that remove all anti-symmetric instances are too simple for peptide sequencing problem. We have proposed a new model for anti-symmetric problem in more realistic way. We have also proposed a novel algorithm which incorporate preprocessing and new model for anti-symmetric issue, and experiments show that this algorithm has better performance on datasets examined.

1. INTRODUCTION

Peptide sequencing by mass spectrometry (referred to as “peptide sequencing” in the following part) is the process of interpreting peptide sequence from the mass spectrum. Peptide sequencing is an important problem in proteomics. Currently, though high throughput mass spectrometers has generated huge amount of spectra, the peptide sequencing these spectrum data is still slow and not accurate. Algorithms for peptide sequencing can be categorized into database search algorithms [1-3] and *de novo* algorithms [4-6]. The database search algorithms are suitable for known sequences already existing in the database. However, they do not have good performance for novel sequences not available in database. For these peptide sequences, the *de novo* algorithms are the methods of choice. *De novo* algorithms interpret peptide sequences from spectrum data purely by analyzing the intensity and correlation of the peaks in the spectrum.

Though current extensive research in *de novo* peptide sequencing helps to improve the accuracies, there are still many obstacles for both *de novo* and database search approaches, which make further improvement of the accuracies of peptide sequencing difficult. Among these obstacles, preprocessing to remove the noises from spectrum before peptide sequencing, as well as the anti-symmetric problem, are two of the very important issues.

Preprocessing to remove noisy peaks

A peak in spectrum is noisy if it does not correspond to a peptide fragment, but a contaminant in mass spectrometers, experiment environments, etc. Since most spectra contain a significant amount of noises, and noisy peaks may mislead interpretation; therefore, preprocessing to remove noisy peaks from the spectrum is necessary.

The anti-symmetric problem

A peak p_i is anti-symmetric if there can be different fragment ion interpretations for p_i, otherwise, p_i is symmetric. There is an anti-symmetric problem in spectrum S if S has one peak p_i which is anti-symmetric. For the spectrum graph G [4] used to represent spectrum, a path in G is called anti-symmetric if there are no two vertices (fragment ion interpretations) on this path which represent the same peak; otherwise, we say that this path has the anti-symmetric problem. The anti-symmetric problem is common in peptide sequencing. Currently there are generally two approaches to the anti-symmetric problem. One approach is to ignore the anti-symmetric problem [6]; and another is to apply the “strict” *anti-symmetric rule* that require each peak to be represented by at most one vertex (fragment ion interpretation) on a path in the spectrum graph G [4; 7; 8]. The “strict” anti-symmetric rule is used in many peptide sequencing algorithms, but whether applying this rule is realistic is doubtful.

In this paper, we will address computational model to remove noise peaks from spectrum. This model also includes the method for introduction of “pseudo peaks” into the spectrum to improve peptide sequencing accuracies. We have also proposed the restricted anti-symmetric model for the anti-symmetric problem. We have then proposed a novel peptide sequencing algorithm which incorporate these two computational models.

2. ANALYSIS OF PROBLEMS AND CURRENT ALGORITHMS

In this section, we will analyze the presence of noises in the spectrum, as well as the difference between the algorithms that use preprocesses and those which do not use them. We will also investigate how significant is the anti-symmetric problem in peptide sequencing by mass spectrum, and how current algorithms cope with this problem.

2.1. General Terminologies

We first define some general terms. Through mass spectrometer, or tandem mass spectrometer, a peptide P=($a_1a_2…a_n$), where each of $a_1,…,a_n$ is one of the amino acids, is fragmented into a spectrum S with maximum charge of α. The parent mass of the peptide P is given by $M = m(\rho) = \sum_{j=1}^{l} m(a_j)$. Consider a peptide prefix fragment ρ_k = ($a_1a_2…a_k$), for $k \leq n$, the prefix mass is defined as $m(\rho_k) = \sum_{j=1}^{k} m(a_j)$. Suffix masses are defined similarly. We always express a fragment mass in experimental spectrum using the *PRM* (prefix residue mass) representation, which is the mass of the prefix fragment. Mathematically, for a fragment q with mass $m(q)$, we define $PRM(q) = m(q)$ if q is a prefix fragment (such as {b-ion}); and we define $PRM(q) = M - m(q)$ if q is a suffix fragment (such as {y-ion}). A spectrum S is composed of many peaks {p_1, p_2 … p_l}. Each of the peaks p_i is represented by its intensity intensity(p_i) and mass-to-charge ratio mz(p_i). If peak p_i is not noisy peak, then it will represent a fragment ion of P. Each peak p_i can be characterized by the ion-type, that is specified by (t, h, z)∈($\Delta_t \times \Delta_h \times \Delta_z$), where Δ_z is the set of charges of the ions, Δ_t is the set of basic ion-type, and Δ_h is the set of neutral losses incurred on the ion. In this paper, we restrict our attention to the set of ion-types Δ^R=($\Delta_t \times \Delta_h \times \Delta_z$), where Δ_z ={1,2,…,α}, Δ_t = {a-ion, b-ion, y-ion} and Δ_h = {∅, –H_2O, –NH_3}. Suppose the (t, h, z)-ion of the fragment q (prefix or suffix fragment) produces an observed peak p_i in the experimental spectrum S that has a mass-to-charge ratio of mz(p_i), then m(q) can be computed using a shifting function, *Shift*, defined as follows:

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

where δ(t) and δ(h) are the mass shift associated with ion-type t and the neutral-loss h, respectively. In addition, we define m(p,(t,h,z))=m(q). We say that peak p_i is a support peak for the fragment q and has ion-type (t, h, z), and that the fragment q is supported by the peak p_i. The peak p_j is a support peak for the peak p_i if both of them are support peaks for the same fragment q.

In peptide sequencing, if two peaks correspond to two consecutive prefix/suffix ions, then we say they are connected. Formally, if p_i and p_j are peaks in S, and they correspond to a (t_i, h_i, z_i) ion and a (t_j, h_j, z_j) ion respectively, then p_i and p_j are said to be connected with a mass tolerance of m_t if $|m(p_i,(t_i,h_i,z_i))-m(p_j,(t_j,h_j,z_j))|<m_t$. The presence of connected peaks is the basis of sequencing algorithms.

In the problem of peptide identification by tandem mass spectrometry, the input includes the mass spectrum S, the set of possible ion types Δ and the parent mass M (and for database search algorithms, a database of peptides). The output is the putative peptide sequence P of that matches with S better than any other peptides.

In this paper, we have specially concerned on multi-charge spectra, which are spectra with charge greater than 1. This is because multi-charge spectra (1) are vastly in existence as the results of ion trap mass spectrometry experiments, (2) usually contain many multi-charge peaks and (3) contains many noisy peaks. Therefore, multi-charge spectra are suitable for our analysis of computational models in this paper.

To account for the different ion-types in spectrum, especially for multi-charge spectrum, we introduced the concept of the extended spectrum S^{α}_{β} [9] where α is the maximum charge of the spectrum S, and β is the largest charge considered for extension. In the *extended spectrum* S^{α}_{β} , for each peak $p_j \in S$ and ion-type (z, t, h)∈({1,2,…,β}×Δ_t×Δ_h), we generate a pseudo-peak denoted by (p_j, (z, t, h)) with a corresponding *assumed* fragment mass. We then introduce an extended spectrum graph, denoted by $G_d(S^{\alpha}_{\beta})$, for the extended spectrum S^{α}_{β}, where d is the "connectivity". For simplicity, we first define $G_1(S^{\alpha}_{\beta})$, the extended spectrum graph for S^{α}_{β} with connectivity 1. Each vertex $v=(p_i,(t,h,z))$ in this graph represents a peak $(p_i,(t,h,z))$ in the extended spectrum S^{α}_{β}, namely, the (t,h,z)-ions for the peak p_i. There is a directed edge between two vertices if their mass difference is equal to the mass of 1 amino acid. We also define the theoretical spectrum *$TS^{\alpha}_{\beta}(P)$* that completely characterizes the set of all possible peaks for a peptide assuming that the ions can take charge *$1,2,...,\beta$*. Note that by comparison of *theoretical spectrum* with experimental spectrum, the theoretical *upper bounds* for different measurements on peptide sequencing results can be calculated [9]. Another useful measure is the SPC, The *shared peaks count* (*SPC*) between the experimental spectrum S and a peptide *P* is defined as the number of peaks in *S* that has the same mass-to-charge ratio (mz) as those in *TS*(*P*), the theoretical spectrum of *P*.

2.2. Datasets

All of the experiments in this paper use the spectra selected with different charges from (a) Amethyst data set from Global Proteome Machine (GPM) [10] and (b) the data set from Institute for Systems Biology (ISB) [11]. The GPM dataset are MS/MS spectra obtained from QSTAR, from both MALDI and ESI sources. The ISB dataset was generated using ESI source from a mixture of 18 proteins, obtained from Ion-Trap, and consists of spectra of up to charge 3. In contrast to the GPM datasets, the ISB datasets are of low quality.

We have selected spectra with corresponding peptide sequences validated by Xcorr score > 2.5. Table 1 listed the number of spectra and the number of peaks per spectrum for different charges of GPM and ISB spectra.

Table 1. The number of spectra, and the number of peaks per spectrum. The results are based on the GPM and ISB datasets of different charges.

Charge	No. Spectrum		No. peaks per spectrum	
	GPM	ISB	GPM	ISB
1	756	16	48.2	226.6
2	874	489	46.9	221.3
3	454	490	42.6	230.7
4	207	-	46.8	-
5	37	-	46.1	-
Total	2328	995	46.5	226.0

Each GPM spectrum has between 20-50 peaks (usually high quality peaks) and an average of about 40 peaks. In contrast, each ISB spectrum has between 50~300 peaks and an average of 150 peaks. Moreover, for the corresponding peptide sequences, GPM sequences have average lengths of 14.5 amino acids, and ISB sequences have average length of 15.0.

2.3. Problems Analysis

Since binning is the general prerequisites for spectra data preprocess, in this section, we have first analyzed the methods for binning of the peaks in the spectrum, and then discuss preprocessing to remove noisy peaks from while introduce "pseudo peaks" into spectrum. Then we have analyzed of the anti-symmetric problem.

- **Binning of peaks in spectrum**

Binning discretizes the mass to charge ratios of the peaks to a series of bins of equal sizes. Each bin contains a single peak. The binning idea is already embedded in [12; 13] for mass spectrum alignment. In [12; 13], the peaks of the spectrum are packed into many bins of same sizes, and the spectrum is transformed to a sequence of 0s and 1s. Recently, a database search algorithm COMET [14] is proposed which uses the bins (usually of size 1 Da) for their correlations and statistical analysis (Z-score) for accurate peptide sequencing by database search (spectrum comparison).

The important parameters considered in binning include the size of the bins, the number of supporting peaks, as well as the intensities of the peaks. The lemma below shows that connected peaks remain connected after binning if we adjust the mass tolerance properly.

Lemma 1. Suppose two peaks *pi* and p_j are connected with a mass tolerance m_t, and p_i* and p_j* are bins corresponding to p_i and p_j, then p_i* and p_j* are connected with a mass tolerance of $z*m_{bin}+m_t$, where m_{bin} is the bin size, and z is the maximum possible charge state.

Proof: Suppose p_i and p_j correspond to a (t_i, h_i, z_i) ion and a (t_j, h_j, z_j) ion respectively, then there exists an amino acid A such that

$$m(A)-m_t<|m(p_i,(t_i,h_i,z_i))-m(p_j,(t_j,h_j,z_j))|< m(A)+m_t \qquad (2)$$

Note that

$$|mz(p_i)-mz(p_i*)|<m_{bin}/2,\ |mz(p_j)-mz(p_j*)|<m_{bin}/2 \qquad (3)$$

From (2), we have

$$|m(p_i*,(t_i,h_i,z_i))-m(p_i,(t_i,h_i,z_i))|<z*m_{bin}/2, \qquad (4)$$
$$|m(p_j*,(t_j,h_j,z_j))-m(p_j,(t_j,h_j,z_j))|<z*m_{bin}/2,$$

Thus

$$|m(p_i*,(t_i,h_i,z_i)-m(p_j*,(t_j,h_j,z_j))| \qquad (5)$$
$$<|m(p_i*(t_i,h_i,z_i))-m(p_i,(t_i,h_i,z_i)|$$

$$+|m(p_i,(t_i,h_i,z_i))-m(p_j,(t_j,h_j,z_j))|$$
$$+|m(p_j*,(t_j,h_j,z_j))-m(p_j,(t_j,h_j,z_j))|$$
$$<z*m_{bin}+|m(p_i,(t_i,h_i,z_i))-m(p_j,(t_j,h_j,z_j))|$$

Similarly,

$$|m(p_i,(t_i,h_i,z_i))-m(p_j,(t_j,h_j,z_j))|- z*m_{bin} \qquad (6)$$
$$<|m(p_i*,(t_i,h_i,z_i)-m(p_j*,(t_j,h_j,z_j))|,$$

Hence, combined with (1), we have

$$m(A)-m_t-z*m_{bin} \qquad (7)$$
$$<|m(p_i*,(t_i,h_i,z_i)-m(pj*,(t_j,h_j,z_j))|$$
$$<m(A)+m_t+z*m_{bin}$$

That is,

$$||m(p_i*,(t_i,h_i,z_i)-m(p_j*,(t_j,h_j,z_j))|-m(A)|<m_t+z*m_{bin} \qquad (8)$$

Hence, p_i* and p_j* are connected with a mass tolerance of $z*m_{bin}+m_t$. Proved.

Therefore, it is clear that given the proper value of tolerance, the binning can preserve the accuracies. The binning method makes the removal of noises easier, and also makes sequencing faster and potentially more accurate, especially for noisy spectrum.

- **Preprocessing to remove noisy peaks and introduce pseudo peaks**

Noisy peaks exist in every spectrum, but how to distinguish them from "true" peaks is not an easy problem. The first step is to analyze the spectrum data and find the patterns of noisy peaks. To this end, we have analyzed most abundant ion type: {b-ion, ∅, 1}, {b-ion, ∅, 2}, {b-ion, -H_2O, 1}, {b-ion, -NH_3, 1}, {y-ion, ∅, 1}, {y-ion, ∅, 2}, {y-ion, -H_2O, 1}, {y-ion, -NH_3, 1}, and assume those peaks not of these ion types noises. The analysis is done on binned GPM dataset and ISB dataset. The experimental spectrum and theoretical spectrum for the corresponding sequence is compared, and peaks in experimental spectrum that can be matched with certain ion types are counted. The "content of peaks" for specific ion type is defined as the ratio of "number of peaks" (in experimental spectrum) of that ion type, over total number of peaks in experimental spectrum. The number of peaks and the contents of peaks of different ion types are analyzed, with average results in Table 2.

From Table 2, we can see that noisy peaks comprise a significant portion of the peaks in the experimental spectrum. For GPM datasets, 80% of the peaks are noisy peaks, and the most abundant ion types – the b- and y-ion types, only compose 6% and 5% of the peaks. For ISB datasets, 83% of the peaks are noisy peaks, and the most abundant ion types - the b- and y-ion types, only compose 4% and 5% of the peaks. ISB spectra have more noisy peaks, and peptide sequencing for these spectra are more difficult.

Table 2. The average contents of different types of peaks in GPM and ISB spectra. The symmetric peaks are just counted once for total content measures.

Ion type	**No. of peaks (Avg)**		**Content of peaks (Avg)**	
	GPM	**ISB**	**GPM**	**ISB**
b-ion, ∅, 1	2.5	11.2	0.07	0.05
b-ion, ∅, 2	0.3	3.55	0.01	0.02
b-ion, -H_2O, 1	0.6	1.83	0.01	0.01
b-ion, -NH_3, 1	0.3	1.83	0.01	0.01
y-ion, ∅, 1	1.6	6.7	0.07	0.04
y-ion, -H_2O, 1	0.3	1.1	0.01	0.01
y-ion, -H_2O, 1	0.3	3.6	0.01	0.02
y-ion, -NH_3, 1	0.3	2.0	0.01	0.01
Noises	26.0	157.3	0.80	0.83
Total	32.2	189.1	1.00	1.00

Further analysis of the noisy peaks indicates that there are more noisy peaks in the middle part (according to mass to charge ratios) of the spectrum, than those at the two ends of the spectrum. Also, most of the noisy peaks have some features in common, such as low intensity and few other ions (b-, y-, loss of water or ammonia, for example) support.

For some famous algorithms such as Lutefisk [6], there are no such preprocessing to remove noises. PEAKS [15] and PepNovo [5] are two famous algorithms that have implemented preprocesses. In PEAKS, the noise level of the spectrum is estimated, and the intensities of all the peaks in the spectrum are reduced by this noise level. Then all the peaks with zero or negative intensities are removed. In PepNovo, preprocessing is applied to remove or downgrade peaks that have low intensity, and do not appear to be b- or y-ions. Recently, the AUDENS algorithm has been proposed [16]. The algorithm has a flexible preprocessing module which screens through the peaks in the spectrum, and distinguishes between signal and noise peaks.

Previous preprocessing for peptide sequencing by mass spectrometry only considered how to remove noisy peaks. However, some fragment ions are not represented by any of the peaks. Appropriate introduction of "pseudo peaks" into spectrum may help in interpretation of these fragment ions, and increase the sequencing accuracies. The idea of "pseudo peaks" is first described in PEAKS [15]. PEAKS assumes that peaks are at every place in the spectrum, and those which are not present in the actual spectrum are peaks with 0 intensities. It is proven that appropriate introduction of "pseudo peaks" can partially solve the problem of missing edges in the spectrum graph approach [15].

In our preprocessing computational model, we have remove noisy peaks from, removal as well as the introduction of pseudo peaks into spectrum. Notice that though the process is similar to previous work, the computation model is different.

- **The anti-symmetric problem**

We have mentioned that there are two approaches to the anti-symmetric problem: 1) ignore the anti-symmetric problem and 2) apply "strict" anti-symmetric rule. In the following part, we show that since both of the approaches are based on unverified assumptions, they do not reflect the nature of real spectrum.

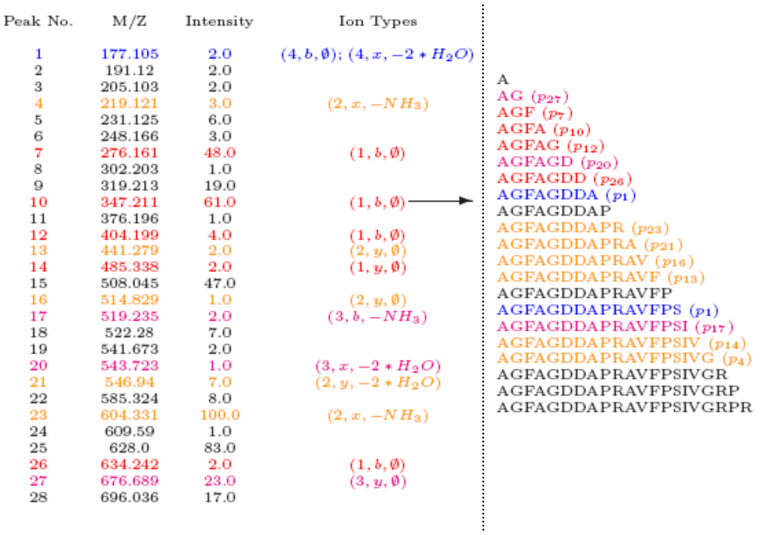
First we give part of a real spectrum from GPM datasets (Fig 1). Note that peak no. 1 has multiple annotations. If we just ignore this peak, then there are two peptide fragments that we cannot interpret (*AGFAGDDA* and *AGFAGDDAPRAVFPS*), while the peptide itself has 21 amino acids. Therefore, we see that the simple model which apply strict anti-symmetric rule may miss some interpretations of peptide fragments.

To analyze the significance of the anti-symmetric problem in peptide sequencing, we have generated the theoretical spectrum of known peptide sequences. We have analyzed most abundant ion types {b-ion, ∅, 1}, {b-ion, ∅, 2}, {b-ion, -H_2O, 1}, {b-ion, -NH_3, 1}, {y-ion, ∅, 1}, {y-ion, ∅, 2}, {y-ion, -H_2O, 1}, {y-ion, -NH_3, 1}, and assume there is no noise. Two peaks are said to be overlap if their mass difference is within threshold (default of 0.25 Da). Note that each of such overlapping peaks is equivalent to a symmetric peak.

Results on selected GPM and ISB spectrum datasets are shown in Table 3. The "average numbers" are the average number of symmetric peaks for theoretical spectrum of one peptide sequence, and the "average ratios" are computed as "average numbers", over average number of peaks in theoretical spectrum.

It is obvious that the instances of overlaps (within threshold, 0.25 Da) are quite common. For the overlaps of b- and y-ions in GPM datasets, there is one overlap instance in about 5 peptide sequences, or in about 67 amino acids. The overall overlap instances are even more common, one instance in about 0.36 sequences, or about 5 amino acids. The ISB datasets has a little bit less overlaps, but overall, there is still more than one instance in 0.35 sequences, or in 4 amino acids.

Note that we have not considered peaks with higher charges (z≥3). But previous research [9] has found that there is significant amount of higher charge (z≥3) peaks in high-charge spectra. It is nature that the number of overlapping instances will increase when we have



Peak No.	M/Z	Intensity	Ion Types
1	177.105	2.0	$(4, b, \emptyset); (4, x, -2 * H_2O)$
2	191.12	2.0	
3	205.103	2.0	
4	219.121	3.0	$(2, x, -NH_3)$
5	231.125	6.0	
6	248.166	3.0	
7	276.161	48.0	$(1, b, \emptyset)$
8	302.203	1.0	
9	319.213	19.0	
10	347.211	61.0	$(1, b, \emptyset)$
11	376.196	1.0	
12	404.199	4.0	$(1, b, \emptyset)$
13	441.279	2.0	$(2, y, \emptyset)$
14	485.338	2.0	$(1, y, \emptyset)$
15	508.045	47.0	
16	514.829	1.0	$(2, y, \emptyset)$
17	519.235	2.0	$(3, b, -NH_3)$
18	522.28	7.0	
19	541.673	2.0	
20	543.723	1.0	$(3, x, -2 * H_2O)$
21	546.94	7.0	$(2, y, -2 * H_2O)$
22	585.324	8.0	
23	604.331	100.0	$(2, x, -NH_3)$
24	609.59	1.0	
25	628.0	83.0	
26	634.242	2.0	$(1, b, \emptyset)$
27	676.689	23.0	$(3, y, \emptyset)$
28	696.036	17.0	

A
AG (p_{27})
AGF (p_{7})
AGFA (p_{10})
AGFAG (p_{12})
AGFAGD (p_{20})
AGFAGDD (p_{26})
AGFAGDDA (p_{1})
AGFAGDDAP
AGFAGDDAPR (p_{23})
AGFAGDDAPRA (p_{21})
AGFAGDDAPRAV (p_{16})
AGFAGDDAPRAVF (p_{13})
AGFAGDDAPRAVFP
AGFAGDDAPRAVFPS (p_{1})
AGFAGDDAPRAVFPSI (p_{17})
AGFAGDDAPRAVFPSIV (p_{14})
AGFAGDDAPRAVFPSIVG (p_{4})
AGFAGDDAPRAVFPSIVGR
AGFAGDDAPRAVFPSIVGRP
AGFAGDDAPRAVFPSIVGRPR

Fig 1. Example of a real spectrum (left) with its corresponding peptide (right). The ion types are represented by (t, h, z)∈(Δt×Δh×Δz), as defined above. In the bracket after the peptide fragment is the corresponding peak number.

considered high-charge peaks, and more ion types. Therefore, “strict” anti-symmetric rule is not realistic.

Table 3: The average numbers and ratios of overlapping instances for different kinds of overlaps. Results on all of the GPM and ISB data.

Overlapping Types	GPM datasets		ISB datasets	
	Average number	Average Ratio	Average number	Average Ratio
b-ion, ∅, 1←→ y-ion, ∅, 1	0.213	0.015	0.154	0.011
b-ion, ∅, 1←→ y-ion, ∅, 2	0.203	0.015	0.173	0.012
b-ion, ∅, 1←→ y-ion, -H_2O, 1	0.307	0.023	0.307	0.023
b-ion, ∅, 1←→ y-ion, -NH_3, 1	0.199	0.014	0.129	0.008
y-ion, ∅, 1←→ b-ion, ∅, 2	0.094	0.006	0.110	0.008
y-ion, ∅, 1←→ b-ion, -H_2O, 1	0.095	0.006	0.220	0.014
y-ion, ∅, 1←→ b-ion, -NH_3, 1	0.090	0.006	0.199	0.012
b-ion, ∅, 2←→ y-ion, ∅, 2	0.336	0.024	0.331	0.024
b-ion, ∅, 2←→ y-ion, -H_2O, 1	0.152	0.000	0.128	0.000
b-ion, ∅, 2←→ y-ion, -NH_3, 1	0.255	0.017	0.340	0.021
y-ion, ∅, 2←→ b-ion, -H_2O, 1	0.143	0.010	0.124	0.008
y-ion, ∅, 2←→ b-ion, -NH_3, 1	0.000	0.000	0.000	0.000
b-ion, -H_2O, 1←→ y-ion, -H_2O, 1	0.213	0.015	0.154	0.011
b-ion, -H_2O, 1←→ y-ion, -NH_3, 1	0.125	0.009	0.269	0.018
y-ion, -H_2O, 1←→ b-ion, -NH_3, 1	0.099	0.007	0.075	0.005
b-ion, -NH_3, 1←→ y-ion, -NH_3, 1	0.213	0.015	0.154	0.011
All	**2.735**	**0.192**	**2.864**	**0.196**

Experiments were also performed with random introduction of noises into theoretical spectrum. Results (details not shown) indicate that there is a significant increase in the number of overlap instances. Therefore, ignoring the anti-symmetric problem is also not realistic, especially for noisy spectra.

In Lutefisk [6], the anti-symmetric problem is assumed not exist, and a peak can be annotated as different ion types. In the Sherenga algorithm [4], only one ion type is possible for each peak, but the exact algorithm that solve the anti-symmetric algorithm is not described. The dynamic programming algorithm for solving anti-symmetric problem is described in [7; 8], and suboptimal algorithm that gives the suboptimal results for the anti-symmetric problem is shown in [17].

Since our experiments have shown that neither of the two approaches (assumptions) to the anti-symmetric problem is realistic, the simple models based on these assumptions may be the obstacles for improvements of current algorithms. Therefore, we have proposed a more realistic computational model for anti-symmetric problem.

3. NEW COMPUTATIONAL MODELS AND ALGORITHM

We propose a new algorithm that is based on two new computational models: 1) preprocessing that can remove noisy peaks from, while introduce pseudo peaks into, the spectrum; and 2) new anti-symmetric model that is more flexible and realistic to the anti-symmetric problem.

3.1. Preprocessing to remove noisy peaks and introduce pseudo peaks

First, the binning process is applied on the peaks in the spectrum. The masses of amino acids are at least of 1.0 Da difference (except for (I, L) and (Q, K), which cannot be distinguished by any *de novo* peptide sequencing algorithm without isotope information). We thus set the value of mass tolerance m_t to be 0.5 Da, and the bin size m_{bin} to be 0.25 Da (according to Lemma 1). With the process of binning, later processes can be even more accurate (lemma 1 shows that there is no loss of accuracy) as well as more efficient because less peaks are considered.

After binning, the pseudo peaks are introduced into every empty bin, and each of them are of 1/10 intensity (empirically determined) of the lowest intensity in original spectrum.

After binning the peaks and introduction of pseudo peaks, the support scores are computed for every bin (peak). Here, we transform each of the bins (peaks) into vertices (ion interpretations) in the extended spectrum graph $G_I(S^{\alpha}_{\beta})$, and then score each of the vertices. Define $N_{support}(v_i)$ as the number of v_j ($v_j \neq v_i$), where $PRM(v_j)=PRM(v_i)$. Define the intensity function as $f_{intensity}(v_i)=\max(0.01, \log_{10}(intensity(v_i)))$, where $\log_{10}(intensity(v_i))$ is normalized, so that $f_{intensity}(v_i)$ cannot be less than 0. Let L be the total number of incoming and outgoing edges for v_i, and a_j be the amino acid for the edge (v_i,v_j) (or (v_j ,v_i)). Then $\sum||(PRM(v_j)-PRM(v_i)|-mass(a_j)|/L$ is the average mass error for v_i. To avoid "divide-by-zero" error in calculating the weight function, we define error function as $f_{error}(v_i)=\max(0.05, \sum||(PRM(v_j)-PRM(v_i)|-mass(a_j)|/L)$. The definition ensure that $f_{error}(v_i)$ is larger than 0.05, a reasonably small error value. Then the score of vertex v_i in $G_I(S^{\alpha}_{\beta})$ is defined as

$$w(v_i) = \frac{N_{\text{support}}(v_i) + f_{\text{intensity}}(v_i)}{f_{\text{error}}(v_i)} \qquad (9)$$

For each bin, the support score is computed and ranked.

Some of the actual peaks that are highly likely to be noises are deleted, and some of the pseudo peaks highly likely to represent ion types are kept. Using this method, we can 1) prune out noises in the spectrum and 2) introduce meaningful peaks into the spectrum. So we may create better spectrum graph to process. Based on the analysis of the scores of peaks in the spectrum (details not shown here), the lowest 20% bins in scores ranking, or those bins with scores less than 1% of the highest ones are filtered out.

3.2. The Anti-symmetric Problem

Since there are a significant ratio of peaks in spectrum that can be (correctly) annotated as different ion types, the anti-symmetric rule should not be strictly followed. Otherwise, there is loss of information. However, since there are still quite some noisy peaks after preprocess, peptide sequencing that ignores anti-symmetric problem may also be misled by noisy peaks, and thus not preferable. Therefore, it would be better if a more flexible and less strict anti-symmetric rule is applied on the spectrum for the anti-symmetric problem.

We have proposed the *restricted anti-symmetric model*. In this model, restricted number (*r*) of peaks can have different ion types. It is easy to observe that the current two approaches for anti-symmetric problem can be described by this model. The approach that ignores the anti-symmetric problem is the one with *r*=number of peaks, and the approach that apply the "strict" anti-symmetric rule is the one with *r*=0.

The restricted anti-symmetric model is based on the extended spectrum graph $G_i(S^{\alpha}_{\beta})$ model using *multi-charge strong tags* [18]. Multi-charge strong tags are highly reliable tags in the spectrum graph $G_i(S^{\alpha}_{\beta})$. A *multi-charge strong tag* of ion-type (z*, t, h) $\in \Delta^R$ is a maximal path $\langle v_0, v_1, v_2, \ldots, v_r \rangle$ in $G_1(S^{\alpha}_{\alpha}, \{\Delta^R\})$, where every vertex v_i is of a (z*, t, h)-ion, in which *t* and *h* should be the same for all vertices, and *z** can be different number from $\{1,\ldots\alpha\}$.

The principle of the *restricted anti-symmetric model* is that if a multi-charge strong tags (tag) T_i in $G_i(S^{j}_{k})$ is of high score, and on this tag, the number (*r*) of overlapping instances (an instance is represented as two vertices of different ion type for the same peak) is within certain tolerance (half of the length of tag), then T_i is a good tag in $G_i(S^{\alpha}_{\beta})$, and it is selected for subsequent process.

It is easy to see that preprocessing and restricted anti-symmetric models can be applied on any *de novo* peptide sequencing algorithms to improve the accuracies (details in experiments). Below we describe our novel algorithm based on these two models.

3.3. Novel Peptide Sequencing Algorithm

Our novel algorithm (GST-SPC*) is based on the previously proposed GST-SPC algorithm [18] which has good performance. GST-SPC algorithm has two phases. In the first phase, the GST-SPC algorithm computes a set of tags - the set of all multi-charge strong tags (corresponding to tags of maximal length in extended spectrum graph) - and this leads to an improvement in the sensitivity that can be achieved. In the second phase, the GST-SPC algorithm try to link these tags, and computes a peptide sequence that is optimal with respect to shared peaks count (SPC) from all sequences that are derived from tags. The GST-SPC performs comparable to or better than other *de novo* sequencing algorithms (Lutefisk and PepNovo), especially on multi-charge spectra.

In the GST-SPC* algorithm, before peptide sequencing, all of the peaks of the spectrum are binned, with each bin of the mass range m_{bin} (0.25 Da). The pseudo peaks are introduced into every empty bins. Bins (transformed to vertices in extended spectrum graph) that have very low scores or low support rank are filtered out. Based on the analysis of the peaks in the spectrum, lowest 20% bins, as well as those bins with support scores less than 5% of the highest ones are filtered out.

In GST-SPC algorithm, we note that all of the tags can have their SPC computed before deriving the paths in the spectrum. So in GST-SPC* algorithm, after tags are generated in the extended spectrum graph $G_1(S^{\alpha}_{\beta})$, we have filtered out the tags that violate the "*restricted anti-symmetric rule*". For the restricted anti-symmetric model on tags, we restricted *r* to be at maximum half the length of that tag. We have then computed the SPC for those "good" tags. Then a variant of width first search algorithm is applied on $G_1(S^{\alpha}_{\beta})$ to find paths from v_0 to v_M, so that these paths have high SPC, and they are consistent with *restricted anti-symmetric model*. Since the number of tags is small, the algorithm is efficient. A flowchart of the whole algorithm is illustrated in Fig 2.

4. EXPERIMENTS

4.1. Experiment Settings

All of the experiments in this paper are performed on a PC with 3.0 GHz CPU and 1.0 GB memory, running Linux system. Our algorithm is implemented in Perl. We have also selected Lutefisk [6], PepNovo [5] and PEAKS [15], three modern and commonly used algorithms with freely available implementations (online portal for PEAKS), for analysis and comparison. The best results given by different algorithms are used for comparison.

For measurement of the sequencing performance, we have adopted the following measurements: Sensitivity and Positive Predictive Value (PPV).

$$\text{Sensitivity} = \#\,\text{correct} / |\,\rho\,| \quad (10)$$
$$\text{PPV} = \#\,\text{correct} / |\,P\,| \quad (11)$$
$$\text{Tag-Sensitivity} = \#\,\text{tag-correct} / |\,\rho\,| \quad (12)$$
$$\text{Tag-PPV} = \#\,\text{tag-correct} / |\,P\,| \quad (13)$$

where #correct is the "number of correctly sequenced amino acids" and #tag-correct is "the sum of lengths of correctly sequenced tags (of length > 1)". #correct 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. The tag-sensitivity accuracy take into consideration of the continuity of the correctly sequences amino acids. For a fair comparison with algorithms as PepNovo that only outputs highest scoring tags, we also use PPV and tag-PPV measures, which indicate how much of the results are correct.

Upper Bound on Sensitivity: Given a spectrum S and the correct peptide sequence ρ, let $U(S_\beta^\alpha, \{d\})$ denote the *theoretical upper bound on sensitivity* that can be attained by any algorithm using the extended spectrum graph $G_d(S_\beta^\alpha)$, namely using the extended spectrum S_β^α and a connectivity d. The bound $U(S_\beta^\alpha, \{d\})$ is computed as the maximum number of amino acids that can be identified from $G_d(S_\beta^\alpha)$ with all of ion types in Δ, over the length of ρ. PepNovo and Lutefisk which considers charge of up to 2 are bounded by $U(S_2^\alpha, \{2\})$ and there is a sizeable gap between $U(S_2^5, \{2\})$ and $U(S_5^5, \{2\})$. This bound was introduced in [18] for the analysis of the multi-charge spectra. In this paper, we have also computed this bound to evaluate the performance of different algorithms.

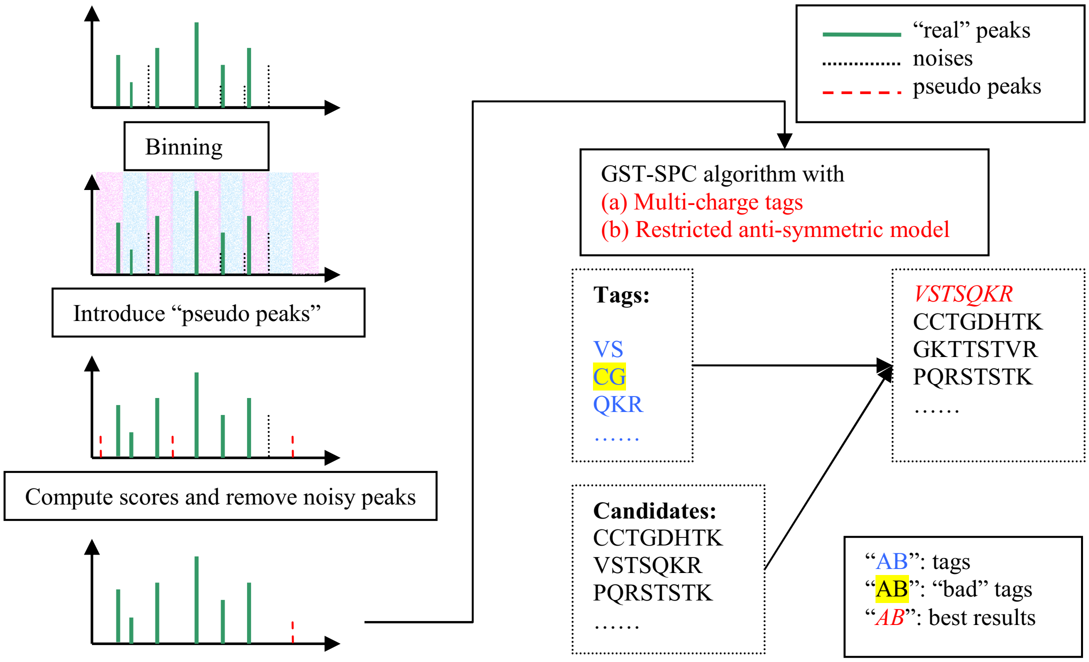


Fig 2. Flowchart of the whole algorithm. "bad" tags are tags that violate the restricted anti-symmetric model.

4.2. Results

We have first analyzed the performance of preprocessing method, and compared the results of Lutefisk, PepNovo, PEAKS and GST-SPC. We have also compared these results with theoretical *upper bounds* on sensitivity, to measure how good the results of these algorithms are compared to optimal ones. The GPM and ISB spectra are categorized by charges (given by spectrum data). The results are shown in Table 4.

From results, we have observed that preprocessing to remove the noises can effectively increase the sequencing accuracies. Compared with the results from original GST-SPC without preprocess, both of the PPV and sensitivity accuracies increase by about 8% for GPM datasets, and about 5% for ISB datasets after preprocess. This difference is probably due to the fact that ISB spectrum has more noises in it than GPM spectrum, so after preprocessing to filter out noises, ISB spectra still have more noises. Such accuracies are much superior to results from Lutefisk algorithm, especially on spectrum with high charges (z≥3). The novel algorithm outperforms the PepNovo algorithm on GPM dataset; and for ISB dataset, the accuracies are closer. Interestingly, when compared with PEAKS, we have discovered that though PEAKS's results on spectra with charge 1 and 2 are comparable with our results, they are better than our results on multi-charge spectrum. This is because PEAKS also has a preprocessing step to remove noisy peaks and introduce pseudo peaks, again prove that such preprocessing in necessary. As can be found later, when we have used new anti-symmetric model, the accuracies of our algorithm are improved, and there is almost no difference between them. Compared with theoretical upper bounds, we can see that there is still much room for improvements.

We have then performed analysis on restricted anti-symmetric model. All of the results based on GST-SPC algorithm are preprocessed. The results based on restricted anti-symmetric model (GST-SPC*) are compared with the results based on strict anti-symmetric rule (strict anti-symmetric) and results from GST-SPC which ignores anti-symmetric issue (no anti-symmetric). The results are shown in Table 5.

Table 5. The results based on the restricted anti-symmetric model, compared with other models. The accuracies in cells are represented in a (PPV/sensitivity [tag PPV/tag sensitivity]) format.

Dataset	**No. of spectrum**	**GST-SPC (no anti-symmetric)**	**GST-SPC (strict anti-symmetric)**	**GST-SPC***
GPM				
Charge 1	756	0.395/0.381 [0.131/0.130]	0.394/0.399 [0.144/0.142]	0.398/0.342 [0.144/0.145]
Charge 2	874	0.334/0.385 [0.142/0.160]	0.348/0.386 [0.130/0.158]	0.345/0.408 [0.151/0.159]
Charge 3	454	0.312/0.327 [0.077/0.091]	0.320/0.342 [0.078/0.090]	0.332/0.351 [0.079/0.096]
Charge 4	207	0.230/0.229 [0.043/0.042]	0.238/0.238 [0.043/0.041]	0.241/0.239 [0.046/0.045]
Charge 5	37	0.195/0.190 [0.020/0.027]	0.197/0.195 [0.026/0.025]	0.208/0.201 [0.028/0.029]
Total	2328	**0.345/0.360 [0.116/0.146]**	**0.344/0.364 [0.123/0.155]**	**0.347/0.375 [0.129/0.158]**
ISB				
Charge 1	16	0.390/0.473 [0.120/0.132]	0.386/0.486 [0.121/0.132]	0.393/0.491 [0.161/0.160]
Charge 2	489	0.411/0.398 [0.096/0.072]	0.414/0.397 [0.090/0.076]	0.434/0.421 [0.119/0.121]
Charge 3	490	0.408/0.496 [0.101/0.145]	0.426/0.528 [0.115/0.156]	0.419/0.531 [0.117/0.164]
Total	995	**0.409/0.447 [0.109/0.120]**	**0.419/0.464 [0.118/0.112]**	**0.427/0.475 [0.119/0.141]**

Table 4. The performance of preprocess. The accuracies in cells are represented in a (PPV/sensitivity) format. "-" means that the value is not available by the algorithm, and "*" shows the average values based on charge 1 and charge 2 spectra.

Dataset	**No. of spectrum**	**Upper Bound (Sensitivity)**	**Lutefisk**	**PepNovo**	**PEAKS**	**GST-SPC (without preprocess)**	**GST-SPC (with preprocess)**
GPM							
Charge 1	756	0.44	0.261/0.258	0.322/0.186	0.402/0.375	0.369/0.378	0.395/0.381
Charge 2	874	0.52	0.243/0.241	0.316/0.215	0.449/0.437	0.321/0.365	0.334/0.385
Charge 3	454	0.38	0.111/0.113	-	0.329/0.323	0.291/0.291	0.312/0.327
Charge 4	207	0.36	0.065/0.063	-	0.279/0.297	0.219/0.226	0.230/0.229
Charge 5	37	0.29	0/0	-	0.270/0.329	0.192/0.191	0.195/0.190
Total	2328	**0.41**	**0.203/0.202**	**0.319/0.202***	**0.392/0.381**	**0.312/0.336**	**0.345/0.360**
ISB							
Charge 1	16	0.55	0.127/0.130	0.630/0.769	0.481/0.486	0.370/0.464	0.390/0.473
Charge 2	489	0.54	0.033/0.034	0.481/0.445	0.481/0.486	0.360/0.347	0.411/0.398
Charge 3	490	0.49	0.002/0.002	-	0.481/0.486	0.360/0.453	0.408/0.496
Total	995	**0.51**	**0.019/0.020**	**0.486/0.455**	**0.481/0.486**	**0.360/0.401**	**0.409/0.447**

Table 5 shows that the restricted anti-symmetric model has superior accuracies. Compared with the results from algorithms which ignores anti-symmetric problem (no anti-symmetric), the application of restricted anti-symmetric model can improve the accuracies by about 5%, and this is probably due to the fact that restricted anti-symmetric model can remove some “bad” tags. About 2% to 5% improvements is observed when compared with the results from strict anti-symmetric model, this is consistent with the results of significance of the anti-symmetric problem in Table 3. The results also show a great improvement in tag PPV and tag sensitivity by using the restricted anti-symmetric rule, especially on ISB datasets. This may also be caused by the restricted anti-symmetric model that removes the “bad” tags.

Compare the results in Table 5 with those from Table 4, we have also observed that by the use of restricted anti-symmetric rule in GST-SPC*, the peptide sequencing results are more accurate. The results of GST-SPC* are closer to accuracies of PepNovo (charge and 2) and PEAKS, and significantly better than results of Lutefisk. We also note that these results are still about 20% (charge 1 and charge 2 spectra) to 50% (charge 5 spectra) less than the theoretical upper bounds of the accuracies given in [9].

We have then computed the number of results that are 100% match with the correct peptide sequences (sensitivity=1 and PPV=1). Results show that all of these algorithms output more than 5% of 100% match results. For our novel algorithm which introduces pseudo peaks, the problem that many of the missing fragmentations do not have enough peaks support still exists. We think that better scoring function can help to improve the ratio of 100% match results.

We have also applied preprocessing and restricted anti-symmetric model on other algorithms. We have selected PepNovo algorithm in this experiment. PepNovo takes input as the preprocessed spectra by our preprocessing model, and output the tags. We have then rescored and ranked these tags according to the restricted anti-symmetric model. We refer to this method based on preprocessing and restricted anti-symmetric model as PepNovo*.

Table 7. The performance of preprocessing and anti-symmetric model on PepNovo. The accuracies in cells are represented in a (PPV/sensitivity) format.

Dataset	**No. of spectrum**	**PepNovo**	**PepNovo with preprocess**	**PepNovo***
GPM				
Charge 1	756	0.322 / 0.186	0.320 / 0.190	0.330 / 0.201
Charge 2	874	0.316 / 0.215	0.319 / 0.221	0.333 / 0.221
Total	1630	**0.319 / 0.202**	**0.321 / 0.212**	**0.331 / 0.220**
ISB				
Charge 1	16	0.630 / 0.769	0.635 / 0.791	0.645 / 0.791
Charge 2	489	0.481 / 0.445	0.480 / 0.445	0.488 / 0.445
Total	505	**0.486 / 0.455**	**0.485 / 0.417**	**0.489 / 0.425**

The results show that by using preprocess, the accuracies of PepNovo can be improved, but not much. By using preprocessing and restricted anti-symmetric model together, the accuracies can be further improved. We believe that preprocessing and restricted anti-symmetric model can be applied on other algorithms and also improve their accuracies.

Table 6. Sequencing results of Lutefisk, PepNovo, GST-SPC and our novel algorithm. The accurate subsequences are labeled in italics. “M/Z”means mass to charge ratio, “Z”means charge, and “-” means there is no result.

M/Z	**Z**	**Real**	**Lutefisk**	**PepNovo**	**GST-SPC**	**GST-SPC***
1219.8	**2**	**VAQLEQVYIR**	[170.11]E***LEKVYLR***	GL***QLEQVYLR***	AVE***IEQVYIR***	VAAGKE***IEQVYIR***
1397.9	**2**	**ELEEIVQPIISK**	[242.14]***EEL***AVG[LP]***LSK***	***EELVKPLLSK***	***EIEEI***A[101.02]QH***ISK***	***EIEEI***GIIG***PISK***
1644.9	**2**	**PAAPAAPAPAEKTPVKK**	[AP]***AAPA***[HS]AP[198.14]***PA***AA[CS]	***AAPA***DFEAMTNLPK	***APAAPAPA***[56.06]APAMTKVPK	***APAAPAPA***F[51.14]APADHAAAP[8.00]***KK***
1838.8	**3**	**SSYSLSGWYENIYIR**	[172.09]L[303.17][243.13][NP][MT]***LYLR***	-	***SS***IYI[27.30]IIEPCE***IYIR***	***SS***IYI[27.30]IIEPCE***IYIR***
1936.1	**4**	**SIRVTQKSYKVSTSGPR**	[199.14][PW][259.10]L[250.14]***KVSTSGPR***	-	VVIS***VTQK***[63.847]W***KVSTSGPR***	VVCP***VTQQ***[95.80]PG***KVSTSGPR***
2101.1	**4**	**KIETRDGKLVSESSDVLPK**	[243.09]***LV***R[TY]YT***SESS***AE[PV]R	-	IKQHTHECY***SESSDVIPK***	IKQHTHECY***SESSDVIPK***
3752.0	**5**	**LPPGEQCEGEEDTEYMTPSSRPLRPLDTSQSSR**	-	-	***IP***VPAQV[1944.68]GRSPVQIC***SR***	***IP***VVGQV***E***[2025.98]GRSPVIKC***SR***
2359.0	**5**	**CDKDLDTLSGYAMCLPNLTR**	-	-	AFCDYA[417.18]RNQKIRCP***TR***	AFC***DID***[423.17]RNQKIRCP***TR***

In Table 6, we have listed a few “good” interpretations of the GST-SPC* algorithm, on which Lutefisk does not provide good results. It is interesting to note that more and longer peptide fragments are correctly sequenced by the novel algorithm - the power of preprocessing and the restricted anti-symmetric rule.

In these interpretations, we observe that the novel algorithm that incorporates preprocessing and restricted anti-symmetric model can predict more and longer fragments of the correct peptides than Lutefisk, PepNovo and original GST-SPC. For example, for the peptide sequence “*PAAPAAPAPAEKTPVKK*”, the two tags “*APAAPAPA*” and “*KK*” are both interpreted correctly only by this novel algorithm.

Efficiency: The GST-SPC* algorithm can process a GPM spectrum (fewer peaks) in about 8 seconds, and 20 seconds for an ISB spectrum (many peaks). This is a little bit faster than the original GST-SPC algorithm, but slower than Lutefisk algorithm (within 10 seconds for these spectra) and PepNovo (about 10 to 15 seconds for these spectra) algorithm. This is because preprocessing can reduce the number of peaks, but the restricted anti-symmetric rule cause the increase of time. For PEAKS algorithm, the average processing time is 0.3 second per spectrum on the powerful computation facility of *peaks online* (http://www.bioinfor.com:8080/peaksonline). Because of preprocess, the space needed by GST-SPC* is less than the original GST-SPC algorithm. The novel algorithm used approximately 20 MB memory to process a GPM spectrum, and about 50 MB memory to process an ISB spectrum, in which most of the space is used for store the extended spectrum graph.

5. CONCLUSIONS

In this paper, we have addressed two important issues in peptide sequencing. The first one is preprocessing to remove noisy peaks from spectrum, and introduce pseudo peaks into spectrum at the same time. We have shown by experiments that there is a significant portion of noisy peaks in the spectrum, and our preprocessing method, which removes noisy peaks and introduce pseudo peaks, can make peptide sequencing more efficient and more accurate. The second issue is about the anti-symmetric problem. We have shown that both strict anti-symmetric rule and no consideration of anti-symmetric problem are not realistic, and we have proposed a restricted anti-symmetric model. Both models can help improve accuracies of *de novo* algorithms, and the novel GST-SPC* algorithm that incorporates these models is shown to have high performance on datasets examined.

However, there are still gaps between accuracies of this algorithm and the theoretical upper bounds. The algorithm can be improved by using better scoring function (rather than SPC), better preprocessing method, and more adaptable anti-symmetric model. We are currently working on these aspects, and preliminary results are encouraging.

The peptide sequencing problem is a very interesting problem in bioinformatics, and there are many other problems in peptide sequencing, such as peptide sequence assembly. We will apply our computational models on some of these interesting problems in the future.

References

1. Eng, J.K., McCormack, A.L. and John R. Yates, I. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, *JASMS*, **5**, 976-989.
2. Perkins, D.N., Pappin, D.J.C., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data, *Electrophoresis*, **20**, 3551-3567.
3. Tanner, S., Shu, H., Frank, A., Mumby, M., Pevzner, P. and Bafna., V. (2005) Inspect: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra, *Anal Chem*, **77**, 4626--4639.
4. Dancik, V., Addona, T., Clauser, K., Vath, J. and Pevzner, P. (1999) De novo protein sequencing via tandem mass-spectrometry, *J. Comp. Biol.*, **6**, 327-341.
5. Frank, A. and Pevzner, P. (2005) PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling, *Anal. Chem.*, **77**, 964 -973.
6. Taylor, J.A. and Johnson, R.S. (1997) Sequence database searches via de novo peptide sequencing by tandem mass spectrometry, *Rapid Commun Mass Spectrom.*, **11**, 1067-1075.
7. Chen, T., Kao, M.-Y., Tepel, M., Rush, J. and Church, G.M. (2001) A Dynamic Programming Approach to De Novo Peptide Sequencing via Tandem Mass Spectrometry, *Journal of Computational Biology*, **8**, 325-337.

8. Lu, B. and Chen, T. (2004) Algorithms for de novo peptide sequencing via tandem mass spectrometry, *Drug Discovery Today: BioSilico*, **2**, 85-90.

9. Chong, K.F., Ning, K., Leong, H.W. and Pevzner, P. (2006) Modeling and Characterization of Multi-Charge Mass Spectra for Peptide Sequencing, *Journal of Bioinformatics and Computational Biology*, **4**, 1329-1352.

10. Craig, R. and Beavis, R.C. (2004) TANDEM: matching proteins with mass spectra, *Bioinformatics*, **20**, 1466-1467.

11. Keller, A., Purvine, S., Nesvizhskii, A.I., Stolyar, S., Goodlett, D.R. and Kolker, E. (2002) Experimental protein mixture for validating tandem mass spectral analysis, *OMICS*, **6**, 207-212.

12. Pevzner, P.A., Dancik, V. and Tang, C.L. (2000) Mutation-tolerant protein identification by mass-spectrometry, *International Conference on Computational Molecular Biology (RECOMB 2000)*, 231–236.

13. Tsur, D., Tanner, S., Zandi, E., Bafna, V. and Pevzner, P.A. (2005) Identification of Post-translational Modifications via Blind Search of Mass-Spectra. *IEEE Computer Society Bioinformatics Conference (CSB) 2005*.

14. Keller, A., Eng, J., Zhang, N., Li, X.-j. and Aebersold, R. (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats, *Molecular Systems Biology*, doi:10.1038/msb4100024.

15. Ma, B., Zhang, K. and Liang, C. (2005) An Effective Algorithm for the Peptide De Novo Sequencing from MS/MS Spectrum, *Journal of Computer and System Sciences*, **70**, 418-430.

16. Grossmann, J., Roos, F.F., Cieliebak, M., Lipták, Z., Mathis, L.K., Müller, M., Gruissem, W. and Baginsky, S. (2005) AUDENS: A Tool for Automated Peptide de Novo Sequencing, *J. Proteome Res.*, **4**, 1768 -1774.

17. Lu, B. and Chen, T. (2003) A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry, *J Comput Biol.*, **10**, 1-12.

18. Ning, K., Chong, K.F. and Leong, H.W. (2007) De novo Peptide Sequencing for Multi-charge Mass Spectra based on Strong Tags, *Fifth Asia Pacific Bioinformatics Conference (APBC 2007)*.