

# AN ACCURATE AND EFFICIENT ALGORITHM FOR PEPTIDE AND PTM IDENTIFICATION BY TANDEM MASS SPECTROMETRY

KANG NING  
ningkang@comp.nus.edu.sg

HOONG KEE NG  
nghoongk@comp.nus.edu.sg

HON WAI LEONG  
leonghw@comp.nus.edu.sg

*Department of Computer Science, School of Computing, National University of Singapore, Computing 1, Singapore 117590*

Peptide identification by tandem mass spectrometry (MS/MS) is one of the most important problems in proteomics. Recent advances in high throughput MS/MS experiments result in huge amount of spectra. Unfortunately, identification of these spectra is relatively slow, and the accuracies of current algorithms are not high with the presence of noises and post-translational modifications (PTMs). In this paper, we strive to achieve high accuracy and efficiency for peptide identification problem, with special concern on identification of peptides with PTMs. This paper expands our previous work on PepSOM with the introduction of two accurate modified scoring functions:  $S_{\lambda}$  for peptide identification and  $S_{\lambda}^*$  for identification of peptides with PTMs. Experiments showed that our algorithm is both fast and accurate for peptide identification. Experiments on spectra with simulated and real PTMs confirmed that our algorithm is accurate for identifying PTMs.

## 1. Introduction

Peptide identification by tandem mass spectrometry (MS/MS) is a very challenging problem that receives wide attention by computational biologists. Huge amount of MS/MS spectra generated requires the most taxing of resources to process. With the presence of noises and post-translational modifications (PTMs) that further complicate the problem, current algorithms for peptide identification are not very accurate.

Approaches for peptide identification can be categorized into database search algorithms [1-3] and *de novo* algorithms [4-7]. The former return peptide sequences that match the parent mass of the experimental spectrum via some scoring functions. Their accuracies largely depend on the completeness of the database, and the process is usually slow. Additionally, they generally do not perform well for peptides sequences not already known as well as peptides with PTMs. On the other hand, *de novo* algorithms interpret peptide sequences from spectrum data purely by analyzing the intensity and correlation of the peaks in the spectrum data. They can retrieve tags from spectrum with high accuracy [3], and the process is very fast (always within a minute). However, their performance quickly deteriorates in the presence of noises and PTMs.

Striving for high efficiency and identification accuracy for peptide identification by MS/MS with the presence of PTMs is an essential issue, which is especially important for experts in the analysis of results in the “wet laboratory”. This paper focuses on this issue.

Recently, the InsPecT algorithm [8] was proposed, which first generates a set of highly accurate tags from spectrum, and then use these tags to filter peptide sequences in database. As *de novo* is imperfect, multiple tags are produced for each spectrum. The accuracy of InsPecT depends on the quality of the tags but even in the context of up to a dozen modifications, it performs reasonably well. Another interesting aspect of InsPecT is that it uses automata to search for peptide sequences. For a batch of spectrum data, the process can be very quick (about 10 ms per spectrum).

For algorithms based on tags for coarse filtering, though using tags can achieve

reasonable efficiency, the quality of candidates is very dependent on the quality of tags which in turn are highly dependent on the quality of the spectra.

Identification of the PTMs is another important problem. Most of current algorithms that are able to identify PTMs [1-3] first define a set of possible PTMs before peptide identification. This approach is limited by the number of predefined PTMs. The InsPecT algorithm takes another approach, which allows blind PTM search (no specified PTM). This approach can discover virtually every possible PTMs in the peptides, but may have high false positive rate.

Previously, we proposed the PepSOM algorithm [9] which can achieve high efficiency for peptide identification by database search based on SOM and MPRQ. However, the accuracies of the PepSOM results are not very satisfactory. This is because after candidate peptides are retrieved from database, they are scored and ranked by SPC, which is not an accurate scoring function especially on noisy spectra and spectra with PTMs. Apparently, comparing candidate peptides with experimental spectrum alone is not enough, so in this paper candidate peptides are also compared with highly-reliable tags generated from experimental spectrum by our *de novo* algorithm.

## 2. Computational Model and Algorithm

In this section, we briefly formulate the problem of peptide identification by MS/MS, and describe a *de novo* algorithm to generate multi-charge strong tags, and mention PepSOM. Due to space constraints, refer to [9-11] for more details. Then we introduce the scoring functions to score and rank candidate peptides (with PTMs).

### Problem Formulation and Multi-charge Strong Tags

To introduce multi-charge strong tags, we first define some general terms. In tandem mass spectrometry, a peptide sequence  $\rho$  will be fragmented into a spectrum  $S$ , which is composed of many peaks  $\{p_1, p_2, \dots, p_n\}$ . Each of the peaks  $p_i$  is represented by its *intensity*( $p_i$ ) and mass-to-charge ratio  $mz(p_i)$ . If peak  $p_i$  is not noise, then it will represent a fragment ion of  $\rho$ . We say that peak  $p_i$  is a *support peak* for the fragment  $q$  and we say that the fragment  $q$  is supported by the peak  $p_i$ . A 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 the problem of peptide identification by MS/MS, the input is the mass spectrum  $S$ , and the output is the putative peptide sequence  $P$  from which the spectrum is generated. The theoretical spectrum completely characterizes all possible peaks for a peptide by considering all ion types. On the contrary, experimental spectrum seldom completely characterizes all possible peaks for peptides, and it contains many noises.

Recently, we proposed the GST-SPC algorithm [11] which was shown to generate high quality tags (called *multi-charge strong tags*, or simply *tags*). In the first phase, GST-SPC computes a set of all tags. Then GST-SPC tries to link these tags by their mass differences, and computes a peptide sequence that is optimal with respect to shared peaks count (SPC) from all peptides derived from tags. Since previous results show that the tags generated by GST-SPC are accurate, in this paper we use multi-charge strong tags generated by the first phase of GST-SPC in scoring the candidate peptides.

### Binning of Peaks

Binning is performed to convert peptides (transformed to theoretical spectra) in database to high-dimensional vectors in vector space.

A spectrum is divided into fixed intervals by mass-to-charge ratios; within each interval the peak with the highest intensity is chosen. To further improve the performance of binning, we incorporated noise removal and scoring of bins after binning. For the full details, refer to [9]. Proper values of tolerance used in binning can preserve accuracies,

while decreasing the computational cost greatly, especially for noisy spectra. For the ion trap datasets in this paper, mass tolerance  $m_t^*$  is set to be 0.5 Da and the mass range of bin  $m_{bin}$  is set to be 0.25 Da. With the process of binning and noise removal, only those significant bins (peaks) are kept, resulting in better accuracy and efficiency.

### SOM and Multi-Point Range Query

Self-organizing map (SOM) [12] is used to transform high-dimensional vectors to 2D points on a plane. In our algorithm, spectrum similarity could be transformed to vector similarity and then to 2D metric similarity. Subsequently, MPRQ [13, 14] is used for multi-point similarity query on the plane to efficiently identify candidates.

For peptide identification, once the theoretical spectra for the peptide sequences in the database are mapped as 2D points on a plane by SOM, we transform the query (experimental) spectra into query points in plane and proceed to query. MPRQ algorithm also accepts as input a parameter  $d$  that controls the radius of the search distance. The larger the value of  $d$ , the more candidate peptides will be returned. MPRQ can efficiently process multiple input points *simultaneously* during query, effectively performing configurable multi-spectra similarity search on database of known peptides.

### Scoring and Ranking

To achieve high accuracy in peptide identification, the most important step is the scoring and ranking of candidate peptides results from database search. We had shown in [9] that by using SPC alone for scoring of candidate peptides results in low identification accuracy. Therefore, here we also compared the candidate peptides with tags generated by GST-SPC. This approach combines the comparison of candidate peptides with experimental spectrum and also with tags.

We now introduce *SPC score* and *S<sub>tag</sub> score*. SPC score is computed as the number of peaks of the same mass-to-charge ratios (within tolerance) between experimental and theoretical spectrum of the candidate peptide, over the number of peaks in experimental spectrum. The S<sub>tag</sub> score, which measures the similarity of candidate peptide to tags, is computed as the ratio of candidate peptide that can match one or more tags at the correct position (within the range of [0,100] Da), over the length of the candidate peptide. For example, given the candidate peptide “VAQLEQVYIR” and two tags “VAK” and “IVYLR” starting from mass of 0 Da and 550 Da, respectively. If we do not allow mismatch, then S<sub>tag</sub> is computed as (3+4)/10=0.7; if we allow up to one mismatch, then S<sub>tag</sub> is computed as (3+5)/10=0.8. To score and rank candidate peptides, we define and use a scoring function  $S_\lambda$  which is a weighted sum of SPC and S<sub>tag</sub> scores.

$$S_\lambda = w_1 \cdot \text{SPC} + w_2 \cdot \text{S}_{\text{tag}} \quad (1)$$

The weights are derived empirically. We selected a large amount of (experimental spectrum, peptide) pairs with high confidence ( $X_{\text{corr}} \geq 2.5$ ) from the ISB dataset (details in Table 1). We then compute their SPC and S<sub>tag</sub> scores and tried different combinations of these two scores. We found that  $w_1 = 0.1$  and  $w_2 = 0.9$  give discriminative results (details omitted due to space limit), and the results were normalized this way.

For PTM identification, it is observed that because of peptide fragmentation such as loss of water and ammonia, PTMs such as phosphorylation, as well as the errors introduced by the mass spectrometer ion detector, *mass shifts* in spectra are very common. Specifically, each PTM corresponds to a set of shifted peaks in experimental spectrum. And highly possible PTMs should have strong support represented by such a set of mass shifts. Here, we use a *modified SPC scoring function* (SPC\*) that can better handle sets of mass shifts in spectra for identification of peptides with PTMs.

To illustrate the mass shifts by PTMs, Fig. 1 shows an example of an experimental spectrum which is identified to be  $I^{+43}\text{TFYEDR}$  (with PTM) by [15]. We compared it with theoretical spectra (with  $\Delta^R$ ) for two peptides,  $I^{+43}\text{TFYEDR}$  and  $\text{ITFYEDR}$ . The

intensity of the peaks in theoretical spectrum is not known, and we assume b-ion and y-ion peaks to have higher intensity than peaks of other ion types.

Comparison between theoretical spectrum of  $I^{+43}$ TFYEDR and that of ITFYEDR clearly shows a set of peak shifts, corresponding to the PTM on amino acid ‘‘I’’. It is hard to see from Fig. 1 which theoretical spectrum is similar to experimental spectrum without computation of SPC score. The SPC score for  $I^{+43}$ TFYEDR is 0.287, while that for ITFYEDR is 0.154. A big difference is observed between them. Using the above spectrum example, we performed database search, and the top candidate peptides are ITFYEDR and LTFYEEV. The tag generated by GST-SPC algorithm is ‘‘TFYED’’. We computed the  $S_\lambda$  score for these two peptides as  $S_\lambda = 0.1*0.154 + 0.9*0.71 = 0.654$  for ITFYEDR, and  $S_\lambda = 0.1*0.021 + 0.9*0.57 = 0.515$  for LTFYEEV.

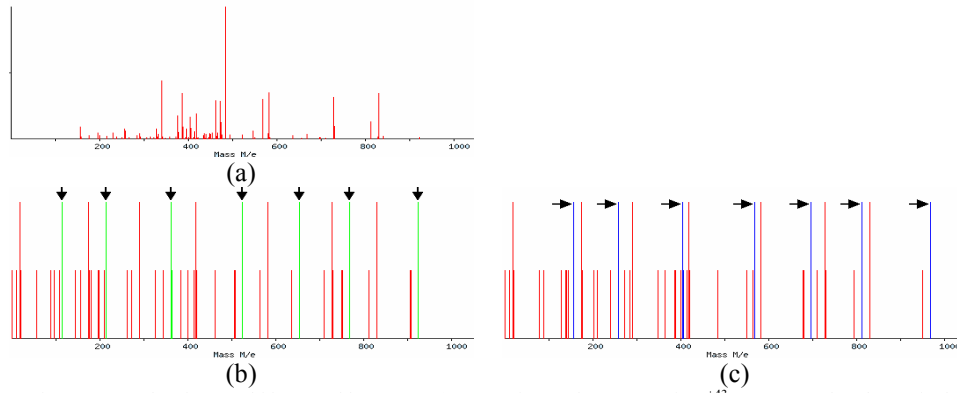


Fig. 1. Example of mass shift caused by PTMs. (a) Experimental spectrum for  $I^{+43}$ TFYEDR, (b) Theoretical spectrum for ITFYEDR, (c) Theoretical spectrum for  $I^{+43}$ TFYEDR. The shifted b-ion peaks are indicated with arrows in (b) and (c).

At each cleavage site, we assume each of  $i*m_{bin}$  Da for all  $-60 \leq i*m_{bin} \leq 60$  (60 Da was determined empirically) as a putative mass shift. We define  $SPC_{i,j}$  as the SPC score between experimental and theoretical spectrum of candidate peptide  $P$ , where we assume a mass shift of  $i*m_{bin}$  Da at cleavage site  $j$  of  $P$ . It is easy to see that  $SPC_{0,j}$  is the SPC score of experimental and theoretical spectrum without mass shift at cleavage site  $j$ . If the largest  $SPC_{i,j}$  for cleavage site  $j$  is obtained with  $i > 0$ , then this cleavage site  $j$  is a putative PTM site with mass shift of  $i*m_{bin}$  Da, with the *PTM score* of

$$S_{PTM}(j) = (SPC_{i,j} - SPC_{0,j}) \quad (2)$$

If  $S_{PTM}(j)$  is greater than a threshold  $T_{PTM}$  (determined empirically to be 0.1), then we say that this putative PTM site is significant, and we identify this as a PTM in the peptide. In  $S_{PTM}(j)$  function, a series of mass shifts introduced by a single PTM is regarded as a whole event, which is more realistic.

Take the same example above, we computed the  $S_{PTM}$  score at cleavage sites for  $I^{+43}$ TFYEDR. Since ‘‘TFYED’’ is a tag, only the cleavage sites after ‘‘I’’ ( $j = 1$ ) and before ‘‘R’’ ( $j = 5$ ) were analyzed. Results are that  $S_{PTM}(1) = 0.133$  with mass shift of 43 Da, and  $S_{PTM}(5) = 0$  with mass shift of 0. Since  $S_{PTM}(1)$  is above  $T_{PTM}$  while  $S_{PTM}(5)$  is below it, we thus identify this PTM correctly.

We further define  $SPC_{\{i_1..i_q\},\{j_1..j_q\}}$  as the SPC score between experimental and theoretical spectrum of identified peptide  $P$ , where a mass shift of  $\{i_1*m_{bin}..i_q*m_{bin}\}$  Da matches with cleavage site  $\{j_1..j_q\}$  of  $P$ , in which each  $S_{PTM}(j)$  is greater than  $T_{PTM}$ . The corresponding  $S_{PTM}^*$  is defined as

$$S_{PTM}^* = \sum_{j=1}^K (SPC_{\{i_1..i_q\},\{j_1..j_q\}} - SPC_{\{i_1..i_q\},\{0..0\}}) | (S_{PTM}(j) > T_{PTM}) \quad (3)$$

in which  $K$  is the length of the peptide.  $S_{\text{PTM}}^*$  indicates the significance of PTMs in a spectrum. Though we have considered multiple PTMs in  $\text{SPC}_{\{j_1..j_q\}}$ , experiments show that there is usually not more than one PTM per spectrum. Taking PTMs into consideration, the modified  $S_\lambda$  score is then defined as

$$S_\lambda^* = w_1 \cdot \text{SPC}_{\{i_1..i_q\}, \{j_1..j_q\}} + w_2 \cdot S_{\text{tag}} \quad (4)$$

which can be used for identification of peptides with PTMs. The weight  $w_1$  and  $w_2$  are again determined empirically. To derive the weights, we selected a large amount of (experimental spectrum, peptide) pairs of peptides with PTMs of high confidence ( $p$ -value of 0.05 or better, computed the same way as in InsPecT [3]) from ISB datasets. We then computed their  $\text{SPC}_{\{i_1..i_q\}, \{j_1..j_q\}}$  and  $S_{\text{tag}}$  scores and tried many different combinations of these two scores. Similar to that used for  $S_\lambda$  scoring function, the results (details not shown) indicate that  $w_1 = 0.3$  and  $w_2 = 0.7$  give discriminative results.

Take the same example as illustrated above; we have computed the  $S_\lambda^*$  score for the spectrum against the two candidate peptides ITFYEDR and LTFYEEV. Results show that ITFYEDR has score  $S_\lambda^* = 0.3 \cdot 0.154 + 0.7 \cdot 0.71 = 0.543$  while  $I^{+43}\text{TFYEDR}$  has score  $S_\lambda^* = 0.3 \cdot 0.287 + 0.7 \cdot 0.71 = 0.583$ ; and LTFYEEV has score  $S_\lambda^* = 0.3 \cdot 0.021 + 0.7 \cdot 0.57 = 0.405$  while  $L^{+43}\text{TFYEEV}$  (the best PTM identified on sequence LTFYEEV) has score  $S_\lambda^* = 0.3 \cdot 0.144 + 0.7 \cdot 0.57 = 0.442$ . These results show that  $I^{+43}\text{TFYEDR}$  has the best  $S_\lambda^*$  score, indicating that  $S_\lambda^*$  score is discriminative.

Apparently,  $S_\lambda^*$  is much more expensive in terms of computation than  $S_\lambda$ . However, after coarse filtering, we only need to consider a small number of candidate peptides; computing  $S_\lambda^*$  on these limited set of candidate peptides is still acceptable.

### Our Algorithm

In this paper, we use a peptide identification algorithm that is a combination of database search technique and *de novo* technique. It has the following steps: (i) both peptides in database and experimental spectra are first converted to high-dimensional vectors via binning; (ii) the vectors are mapped to 2D plane with SOM; (iii) candidate peptides are then selected from database with MPRQ; and (iv) these candidate peptides are scored and ranked (fine filtered) by a scoring function that compares them with the experimental spectrum as well as multi-charge strong tags generated by a *de novo* algorithm [11]. Steps (i)-(iii) are coarse filtering steps, in which spectra similarity is transformed to vector similarity and then to 2D points metric distance similarity. These steps are similar to those in PepSOM [9]. Step (iv) is a fine filtering step in which the candidate peptides are scored and ranked by comparing them with experimental spectrum and tags generated by the GST-SPC algorithm. At the end of step (iii), if we assume that there is no PTM in the spectrum and want to perform fast peptide identification, then  $S_\lambda$  scoring function is used. Otherwise  $S_\lambda^*$  scoring function is used for identification of peptides with PTMs.

### 3. Experiments

Experiments were performed on a 3.0 GHz PC with 1.0 GB main memory running Linux. Our algorithm is implemented in C++ and Perl. SOM\_PAK [16] was the SOM implementation used. For analysis and comparison, we had selected established algorithms with freely available software: two database search algorithms, Sequest [1] and InsPecT [3]; as well as two *de novo* algorithms, Lutefisk [7] and PepNovo [5].

Spectrum datasets (query datasets) were obtained from Open Proteomics Database [17], PeptideAtlas [18] and Institute for Systems Biology [19]. All of the experimental mass spectra were ion trap data having low mass resolution. As the statistical evaluation of the correlation of spectrum and peptide is still a difficult open problem, we treated

Sequest result with  $Xcorr \geq 2.5$  as ground truth, which is considered reliable.

The PeptideAtlas spectrum dataset A8\_IP were obtained from Human Erythroleukemia K562 cell line. Electrospray ionization source of an LCQ Classic ion trap mass spectrometer (ThermoElectron, San Jose, CA) was used, and DTA files were generated from MS/MS spectra using TurboSequest. All 44 spectra that were identified with  $Xcorr \geq 2.5$  were chosen. For OPD, the spectrum dataset used was opd00001\_ECOLI, *Escherichia coli* spectra 021112.EcoliSol 37.1(000). The spectra were obtained from *E. coli* HMS 174 (DE3) cell, which is grown in LB medium until  $\sim 0.6$  abs (OD 600). The spectra were generated by the ThermoFinnigan ESI-Ion Trap “Dexa XP Plus” and the sequences for these spectra were validated by Sequest. The ISB dataset was generated using an ESI source from a mixture of 18 proteins, obtained from ion trap mass spectrometry, and consists of spectra of up to charge 3. Most importantly, these ISB datasets were annotated by a few algorithms [8, 15] to be free of PTMs (refer to [http://www.systemsbiology.org/extra/protein\\_mixture.html](http://www.systemsbiology.org/extra/protein_mixture.html)).

The databases that we have used contain peptides from the respective protein sequences dataset. Specifically, *E. coli* K12 protein sequences for OPD datasets, IPI HUMAN protein sequences for PeptideAtlas dataset and human plus control protein mixture for ISB dataset were used. As the number of protein sequences were very large for PeptideAtlas (60,090) and ISB (88,374) datasets, we used only the protein sequences corresponding to spectra identified with  $Xcorr \geq 2.5$  (our ground truth). However, the sizes of databases were still very large because of many peptides. The parameters for the generation of databases, the query datasets and theoretical spectra are shown in Table 1.

Table 1. Parameters for the generation of databases and theoretical spectra.

Parameters	Values		
	PeptideAtlas	OPD	ISB
No. of protein sequences	31	4,279	3,553
Total database size	9,421	494,049	1,248,212
Query size	44	202	995
Fragments mass tolerance	0.5 Da		
Parent mass tolerance	1.0 Da		
Modifications	-		
Charge	+2, +3		
Ion type	a, b, y, $-H_2O$ , $-NH_3$		
Missed cleavages	0		
Protease	Trypsin		
Mass range	0-5000 Da		

To compare the different algorithms, the following accuracy measures were used:

$$Recall = \frac{\#correct}{|\rho|} \quad (5)$$

$$Precision = \frac{\#correct}{|P|} \quad (6)$$

where  $\#correct$  is the number of correctly identified amino acids. For two amino acids in the correct peptide  $\rho$  and the respective identification result  $P$ , only if their positions do not have a difference of more than 100 Da (determined empirically) and they are of the same amino acids (except (I, L), as well as (K, Q)), do they contribute one count to  $\#correct$ . A high *Recall* being that the algorithm recovers a large portion of the correct peptide. For a fair comparison with algorithms like PepNovo that only outputs the highest scoring tags (subsequences), we also use a *Precision* measure, which measures how many of the results are correct. Note that these recall and precision measures are different from sensitivity and specificity measures used in PepSOM paper, since there is a position constraint on amino acids in recall and precision measures, rather than only using LCS to measure  $\#correct$  in sensitivity and specificity in PepSOM.

## Experiments on Peptide Identification

In this subsection, we performed experiments using peptides without PTMs. Firstly, we analyzed the quality of the tags generated by the GST-SPC algorithm. We measured the ratio of completely correct tags in the results, as well as recall and precision of the tags. Results are shown in Table 2. Note that we had only analyzed the quality of tags on ISB spectra in our previous study [11]. By also measuring OPD and PeptideAtlas datasets, we empirically proved the accuracy of tags on a variety of datasets.

Table 2. Statistical results on the quality of the generated tags. “No. of tags per spectrum” shows the average number of tags generated per spectrum. “No. of complete correct per spectrum” measures the average number of tags identified that are completely correct (i.e. identified with 100% precision). “Complete correct accuracy” is the ratio of “completely correct tags” to number of tags on average. The recall and precision results are obtained from tags by the GST-SPC algorithm.

Datasets	Query Size	Average Peptide length	No. of Tags per Spectrum	No. of Complete Correct per Spectrum	Complete Correct Accuracy	Recall	Precision
OPD	202	10.14	7.42	6.01	0.81	0.43	0.43
PeptideAtlas	44	10.02	9.76	6.83	0.70	0.40	0.36
ISB	995	19.37	6.19	4.61	0.74	0.36	0.32

From Table 2, we observed that more than 1/3 of the amino acids in real peptide sequences (recall) can be correctly identified by tags. Also, when the tags are generated, more than 70% of the tags are completely correct, meaning that the tags generated are reliable. Since each tag is at least one amino acid in length, it can also be observed that a significant amount of tags are overlapping. For more reliable results in the following experiments, only non-overlap tags with high scores (determined by GST-SPC) are used.

Secondly, we investigate the quality of candidate peptides identified by MPRQ and SOM. We analyzed the search distance  $d$  on the accuracy of search results on datasets of different sizes. Note that similar spectra that overlap on the same 2D point can be losslessly retrieved by our algorithm since it has built an index for these overlapping spectra. The candidate peptides are scored and ranked by SPC score only. First-rank peptide represents the peptide with theoretical spectrum that has the highest SPC score against the experimental spectra. Best-match peptide is the peptide among all candidates that match with the “real” peptide with the highest precision (recall).

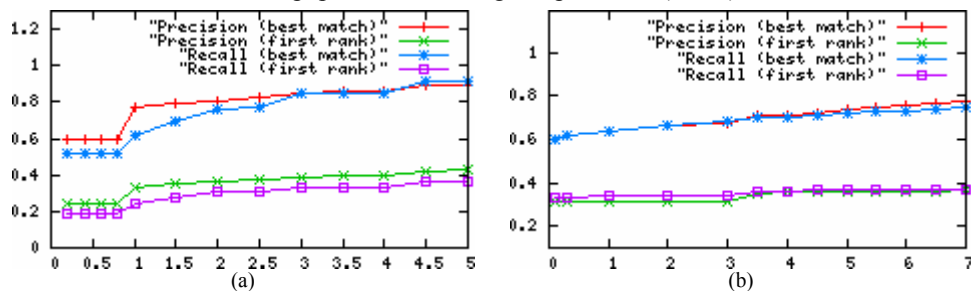


Fig. 2. The effects of increasing distance  $d$  on recall and precision on (a) PeptideAtlas dataset (b) ISB dataset.

Analysis of the search distance  $d$  (Fig. 2) show that the recall and precision of best-match peptides are much higher than those for first-rank peptides, indicating that (i) SPC score alone is not a good scoring function; and (ii) a properly designed scoring function can improve the identification accuracies significantly. The PeptideAtlas, OPD and ISB datasets are datasets of increasing sizes. Results also indicate that for larger datasets, the search distance should also be bigger to achieve high recall and precision. Therefore, for datasets of large sizes, we used larger search distances. Specifically, we used  $d=1.0$  for PeptideAtlas datasets,  $d=2.5$  for OPD datasets and  $d=3.5$  for ISB datasets. Fig. 2(b) also shows that as search distance  $d$  increases, more candidate peptides are returned increasing both recall and precision. However, when  $d$  grows beyond the similarity clusters in the

SOM, not so relevant candidates returned bring down the recall and precision values.

The average search time per spectrum is less than 11 ms. This is comparable to InsPecT (which average 10 ms per spectrum with default settings, but based on smaller database), one of the fastest database search algorithms.

Another important question is, among the candidate sequences, what is the ratio of them being identical to the real peptide sequences. We have found that when we consider all of the candidates, the “complete correct accuracy” is much higher; for OPD dataset it is 69.5%, PeptideAtlas 63.1% and ISB 65.3%. And if we allow up to two amino acids difference from real peptide sequences, the ratios increase to 80.1%, 85.3% and 78.6% respectively. Therefore, given a good scoring function, the peptide identification accuracy can be significantly increased. As the size of the candidate sequences generated by our algorithm is rather small (refer Table 6), we believe these high ratios indicate good performance of the SOM and MPRQ as coarse filtering.

Table 3. Comparison of different algorithms on the accuracies of peptide identification. In each column, the “Precision / Recall” values are listed.

Datasets	Database Size	Query Size	InsPecT	Lutefisk	PepNovo	Our Algorithm
OPD	494,049	202	0.580 / 0.542	0.101 / 0.006	0.232 / 0.186	<b>0.582 / 0.603</b>
PeptideAtlas	9,421	44	0.801 / 0.389	0.149 / 0.057	0.275 / 0.128	<b>0.521 / 0.457</b>
ISB	1,248,212	995	0.584 / 0.621	0.011 / 0.022	0.548 / 0.561	<b>0.594 / 0.695</b>

Subsequently, we compared our algorithm with other algorithms. For our algorithm,  $S_\lambda$  scoring function is used, and the results are based on peptides with the best score. The results with first rank given by these algorithms were used for analysis.

We observe from Table 3 that both precision and recall of our algorithm are better than Lutefisk and PepNovo (both *de novo* algorithms). This is reasonable since *de novo* algorithms do not utilize any information from databases. But even when comparing their results with the quality of tags generated by our algorithm (Table 2), we notice that the quality of tags generated by our algorithm is better than peptide identification results by Lutefisk, and comparable with that by PepNovo. Although InsPecT has higher precision, our results outperform InsPecT in recall. Specifically, for the OPD dataset, both the algorithms have precision of about 0.58, but our algorithm has higher recall. For the PeptideAtlas dataset, the precision of our algorithm is much worse than that of InsPecT, but the recall is 17% better. For the ISB dataset, both InsPecT and our algorithm have similar precision, but recall of our algorithm is higher. This means that our algorithm can identify more portion of the real peptide.

We have also observed that by scoring peptide candidates using  $S_\lambda$ , both precision and recall consistently increase (last column of Table 3), compared with only using SPC score (Fig. 2). This proves the superiority of  $S_\lambda$  scoring function.

### Experiments on PTM Identification

PTM identification is of great importance to current mass spectrum analysis. To analyze PTMs, we first performed experiments on experimental spectra *in silico* with artificially added PTMs (we call these simulated PTMs). We selected spectra from ISB datasets that are annotated to be free of PTMs. For every peptide, the PTM that we had artificially added is phosphorylation for every amino acids involved. In the corresponding experimental spectrum, we shifted every peak that corresponds to the respective peptide fragment according to the restricted ion types  $\Delta^R$ . Summary of modifications:

Modification	Amino acid involved	Context	Mass difference (Da)
Phosphorylation	T,S,Y	PTM	+79.97

Our algorithm is not designed specifically for phosphorylation per se, but can also be easily applied to detect other types of PTMs. These can be shown in experiments on the detection of PTMs on real datasets, using ISB spectra [19] that contain PTMs but are



distinct from the “ISB dataset” we used above which do not. It was found that there are PTMs in these ISB datasets [15], and their identifications (called UCSD annotation) are found at [http://www.systemsbio.org/extra/UCSD\\_supplemental\\_identifications.txt](http://www.systemsbio.org/extra/UCSD_supplemental_identifications.txt). There are 551 spectra with at least one PTM identified by InsPecT from a total of 2,799 ISB spectra. The results of the “UCSD annotation” were treated as ground truth, since they are annotated by InsPecT with an annotation of p-value 0.05 or better, indicating reliability.

The UCSD datasets contains those ISB spectra with PTMs identified by InsPecT. However, for analysis and comparison purpose, we have also applied different algorithms on other ISB datasets for possible identification of new PTMs in these spectra. ISB spectra that are different from previously described datasets were selected; we refer to this dataset and our annotations as “NN annotation” dataset. This dataset contains 3,000 spectra. Again, we treat PTMs identified with p-value 0.05 (computed the same way as in InsPecT) or better as ground truth.

$S_{\lambda}^*$  was used to identify peptides with PTMs. Peptide identification accuracy is measured as the percentage of candidate peptides that contain the exact original (unmodified) peptide. PTM identification accuracy is measured as the percentage of results in which the best-score PTM (Eq. 3) identification is *correct*, where PTM identification is defined as correct if (i) the original peptide is identified correctly *and* (ii) the PTM site and the value of mass shift are identified correctly. For instance, a peptide (with PTM) “AS<sup>+80</sup>RK” is identified correctly, if “ASRK” is identified correctly, *and* the PTM site and PTM mass shift (+80 Da) after “S” are identified correctly.

Firstly, we have analyzed the accuracies of PTM identification on simulated PTMs. We used tags of specific lengths for analysis. The results on ISB spectra with simulated PTMs are shown in Table 4.

Table 4. Accuracies (%) of PTM identification from simulated spectra by tags of different lengths. The columns with Top  $k = 1, 2, 3, 4$  represent the (peptide / PTM) identification accuracies for top- $k$ . “No limit” means that the best-score tags are used without any length limit. “Filtration ratio” is computed as the number of candidates after tag filtration over the number of candidates after MPRQ. “Time” is the total time to identify the peptides and PTMs for 995 spectra. Results without using tags are also shown.

Database Size	Query Size	Tag length	Top 1	Top 2	Top 3	Top 4	All	Filtration Ratio	Time (s)
1,248,212	995	3	46.7 / 30.2	50.1 / 36.3	62.6 / 40.5	69.2 / 46.5	<b>71.3 / 60.1</b>	0.0148	65.6
		4	56.9 / 34.6	40.5 / 25.6	44.4 / 32.6	51.0 / 39.0	<b>63.3 / 50.0</b>	0.0021	67.5
		No limit	46.8 / 32.9	52.0 / 36.1	58.3 / 43.3	64.4 / 50.1	<b>72.8 / 59.1</b>	0.0491	66.6
		No tag	31.7 / 26.4	35.5 / 26.6	41.1 / 35.2	46.9 / 39.5	<b>56.7 / 40.8</b>	–	70.7

From the results above, it can be observed that sequence tags of length 3 and 4 are able to further filter out candidates from the results of SOM and MPRQ. With reduced candidates, the accuracy for PTM identification increased. Compared with results without tags, the percentages of search results that contain the exact correct peptide are significantly higher. For example, for filtration with tags of length 3, about 46.7% to 71.3% of original peptides are identified correctly. Increase filtration tags length to 4 decreases peptide identification accuracies, but using best-score tags without any length limit do not show such decrease. PTM identification accuracies show similar patterns. These indicate that although longer tags may have lower recall, the best-score tags are of high recall, regardless of their length.

We notice that the process time based on tags of length 4 is greater than that on tags of length 3. We think that though longer tags may filter out more candidates, which makes later scoring faster, the filtering step itself is more time consuming than those based on tags of length 3, so that the total time is longer. Also, we notice that the processing time of tags without length limit is shorter than that of the processing time of tags with length 4. Since we have observed that the best-score tags are of average length

$> 4$  (details not shown), this indicates that the best-score tags (without any length limit) can filter out even more candidates, which makes the scoring step faster.

The InsPecT algorithm (with blind PTM search) is also applied on these spectra with simulated PTMs. Results show that both the peptide and PTM identification accuracies are not as high as our algorithm. In all of the results (10 identifications per spectrum) given by InsPecT, the peptide identification accuracy is around 50%, while the PTM identification accuracy is approximately 33%.

We observed that by comparing candidate peptides with tags, a large ratio of candidate peptides that do not match with any tags will be filtered. We have also observed that the filtration ratio is small. For instance, the filtration ratio for tags with length 3 is 0.0148; for length 4 is 0.0021. This indicates that tags can further reduce the number of candidate peptides for further careful examination by  $S_{\lambda}^*$ .

Experiments on the identification of PTMs on real ISB spectra with ‘‘UCSD annotation’’ were also performed (Fig. 3(a)). Since experiments on simulated PTMs (Table 4) show that best-score tags with no length limits have the best accuracies, we used them here. Again, we treated PTMs identified with p-value 0.05 or better as ground truth. Results show that the filtration ratio of our algorithm is 0.062. The peptide identification accuracies are 42.0%, 45.7%, 48.2%, 50.6% and 55.5% for Top 1, 2, 3, 4 and all candidates, respectively; and the PTM identification accuracies are 31.6%, 33.1%, 34.8%, 40.2% and 41.8% for Top 1, 2, 3, 4 and all candidates, respectively. These values are slightly smaller than those on simulated spectra, and we think this is due to the different PTM types in real spectrum.

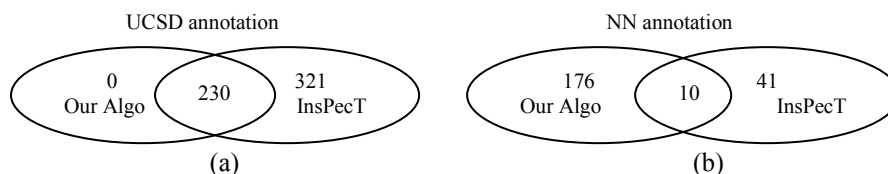


Fig. 3. The number of PTMs identified by our algorithm (no length limit on tags, Top 1 result) and InsPecT on ‘‘UCSD annotation’’ and ‘‘NN annotation’’ datasets. (a) There are 230 PTMs that both algorithms identified on ‘‘UCSD annotation’’. (b) There are 10 PTMs that both algorithms identified on ‘‘NN annotation’’.

Apart from the ‘‘UCSD annotations’’ on 2,799 spectra, we have also examined other ISB spectra in ‘‘NN annotation’’ on 3,000 spectra (Fig. 3(b)). Again, note that the UCSD datasets contains those ISB spectra with PTMs identified by InsPecT. On the other hand, the NN datasets contains spectra that are not subject to PTM identification by any algorithms before. We applied our algorithm and InsPecT algorithm (with blind PTM search) on these spectra; PTMs identification with p-value 0.05 or better are treated as ground truth. On these ‘‘NN annotation’’ spectra, InsPecT has identified 51 PTMs (among 78 peptide identifications with or without PTMs of p-value 0.05 or better) while our algorithm (using tags with no length limit) has identified 186 PTMs. Among those identified PTMs, 10 are identified by both algorithms. Interestingly, though InsPecT algorithm has identified 321 more PTMs in the UCSD annotation, in the NN annotation, 135 more peptides were identified using our algorithm. We think that this is because of the variance of the ‘‘UCSD annotation’’ and ‘‘NN annotation’’ datasets. Even though ‘‘NN annotation’’ dataset is selected randomly from ISB dataset, the complex nature of the ISB data itself makes such a big difference.

Listed below are some novel PTMs annotated (predicted) by our algorithm with high PTM scores (and low p-values) in ‘‘NN annotation’’ ISB spectra. Some of these annotations have experiment support (see References column), others are completely novel annotations. A full list of these novel PTM annotations will be provided upon request.

Table 5. Some representative novel PTMs identified by our algorithm.

Spectra (sergei digest)	Peptides with PTMs	$S_{PTM}$	p-value	Notes and References
A_full_01.3541.3553.2	NFYFQCFNSG <sup>37</sup> LDSVLIADVPIEES	0.254	0.0207	
A_full_02.2185.2189.3	G <sup>40</sup> IIWGEDTLMEYLENPKK	0.197	0.0148	
A_full_03.3951.3951.2	DVPDARKC <sup>33</sup> ACASHVAKVA	0.563	0.0059	Also annotated by Unimod [20] without verification
A_full_05.1020.1020.2	LLKF <sup>71</sup> GQEV	0.349	0.0001	[21]
A_full_07.1737.1741.3	A <sup>16</sup> TAQADVMMETPDELQAAVWEK	0.197	0.0062	
B_full_03.0832.0834.2	NALS <sup>78</sup> GNQNLEVVQLRLY	0.333	0.0007	

### Efficiency

One of the most important aspects of our algorithm is that it is very fast. Table 6, reproduced from [9], explains. The coarse filtering rate is very low as we only need to compare each spectrum against the candidate peptides identified by MPRQ. Compared to the tandem cosine coarse filter used in [22] that filters to around 0.5% of the database, our algorithm has a better filtering efficiency.

Table 6. Candidates size, average candidate size and coarse filtering rate. ‘‘Candidates size’’ is the combined total results from coarse filtering of the database using the query size as input query points for the MPRQ algorithm. ‘‘Average Candidate Size’’ is the average peptide sequence candidates for each spectrum (query). ‘‘Coarse Filtering Rate’’ is computed by ‘‘average candidate size’’ over the database size.

Database	Database Size (peptides)	Query Size	Candidates Size	Average Candidate Size	Coarse Filtering Rate
OPD	494,049	202	68,610	339.7	0.069%
PeptideAtlas	9,421	44	654	14.9	0.158%
ISB	1,248,212	995	101,443	102.0	0.008%

After database search, the scoring of candidate peptides by  $S_\lambda$  scoring function is approximately 5 seconds per spectrum, while  $S_\lambda^*$  scoring function needs about 20 seconds for each spectrum. As comparison, for InsPecT the running time of blind search of PTMs is approximately 1 second per spectrum per megabyte of database (as stated in InsPecT documentation, and verified by our experiments). Relatively, our algorithm is very efficient on PTM identifications.

The program for our algorithm is available upon request.

## 4. Conclusion and Future Work

This paper focused solely on the peptide identification problem, striving to achieve high identification accuracy and efficiency for peptide identification, especially for peptides with PTMs. An algorithm that transforms spectra similarity to similarity of vectors, and then to metric similarity (distance) of 2D points on a plane was used. The vectors are input to SOM to produce an indexable map in which MPRQ could use to find candidate peptides efficiently. Candidate peptides are fine-filtered with proposed scoring functions ( $S_\lambda$  for peptide identification and  $S_\lambda^*$  for identification of peptides with PTMs), which compare each of them with experimental spectrum and highly reliable tags generated by our GST-SPC algorithm.

Experiments lent strong support to the fact that by using  $S_\lambda$  scoring function that take into consideration score based on tags, the accuracies (precision and recall of the results) of our algorithm are high, yet still maintaining efficiency, especially for large batch processes. By using  $S_\lambda^*$  scoring function that take into consideration of mass shifts caused by PTMs, our algorithm can accurately identify peptides with PTMs. The novel PTMs that are predicted by our algorithm with high scores are interesting for manual verifications later in wet laboratories.

Recently, we noticed an algorithm Popitam (<http://www.expasy.org/tools/popitam>) that has similar scheme as ours. In Popitam, the scoring function is based on genetic

programming (machine learning), which are quite different from our scoring function. Comparison of the two algorithms may be of interest in the future.

## References

- [1] 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," *Journal of the American Society for Mass Spectrometry*, vol. 5, pp. 976-989, 1994.
- [2] D. N. Perkins, D. J. C. Pappin, D. M. Creasy, et al., "Probability-based protein identification by searching sequence databases using mass spectrometry data," *Electrophoresis*, vol. 20, pp. 3551-3567, 1999.
- [3] A. Frank, S. Tanner, and P. Pevzner, "Peptide Sequence Tags for Fast Database Search in Mass Spectrometry," *International Conference on Research in Computational Molecular Biology (RECOMB)*, 2005.
- [4] V. Dancik, T. Addona, K. Clauser, et al., "De novo protein sequencing via tandem mass-spectrometry," *Journal of Computational Biology*, vol. 6, pp. 327-341, 1999.
- [5] A. Frank and P. Pevzner, "PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling," *Analytical Chemistry*, vol. 77, pp. 964 -973, 2005.
- [6] B. Ma, K. Zhang, C. Hendrie, et al., "PEAKS: Powerful Software for Peptide De Novo Sequencing by MS/MS," *Rapid Comm in Mass Spectrometry*, vol. 17, pp. 2337-2342, 2003.
- [7] J. A. Taylor and R. S. Johnson, "Sequence database searches via de novo peptide sequencing by tandem mass spectrometry," *Rapid Comm in Mass Spec*, vol. 11, pp. 1067-1075, 1997.
- [8] S. Tanner, H. Shu, A. Frank, et al., "InsPecT: Fast and accurate identification of post-translationally modified peptides from tandem mass spectra," *Analytical Chemistry*, vol. 77, pp. 4626-4639, 2005.
- [9] K. Ning, H. K. Ng, and H. W. Leong, "PepSOM: An Algorithm for Peptide Identification by Tandem Mass Spectrometry based on SOM," *Genome Informatics*, vol. 17, pp. 194-205, 2006.
- [10] H. K. Ng, K. Ning, and H. W. Leong, "A New Approach for Similarity Queries of Biological Sequences in Databases," *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2007.
- [11] K. Ning, K. F. Chong, and H. W. Leong, "De novo Peptide Sequencing for Multi-charge Mass Spectra based on Strong Tags," *Asia Pacific Bioinformatics Conference*, 2007.
- [12] T. Kohonen, *Self-Organizing Maps*, 3rd ed: Springer, 2001.
- [13] H. K. Ng and H. W. Leong, "Path-Based Range Query Processing Using Sorted Path and Rectangle Intersection Approach," *International Conference on Database Systems for Advanced Applications (DASFAA)*, pp. 184-189, 2004.
- [14] H. K. Ng, H. W. Leong, and N. L. Ho, "Efficient Algorithm for Path-Based Range Query in Spatial Databases," *Int'l Database Engg & Applications Symp (IDEAS)*, pp. 334-343, 2004.
- [15] D. Tsur, S. Tanner, E. Zandi, et al., "Identification of post-translational modifications by blind search of mass spectra," *Nature Biotechnology*, vol. 23, pp. 1562 - 1567, 2005.
- [16] T. Kohonen, J. Hynninen, J. Kangas, et al., "SOM\_PAK: The Self-Organizing Map Program Package," *Technical Report A31*, pp. FIN-02150 Espoo, 1996.
- [17] J. T. Prince, M. W. Carlson, R. Wang, et al., "The need for a public proteomics repository," *Nature Biotechnology*, vol. 22, pp. 471-472, 2004.
- [18] F. Desiere, E. W. Deutsch, N. L. King, et al., "The PeptideAtlas Project," *Nucleic Acids Research*, vol. 34, pp. D655-D658, 2006.
- [19] A. Keller, S. Purvine, A. I. Nesvizhskii, et al., "Experimental protein mixture for validating tandem mass spectral analysis," *Omics*, vol. 6, pp. 207-212, 2002.
- [20] S. Kim, S. Na, J. W. Sim, et al., "MODi: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra," *Nucleic Acids Research*, vol. 34, pp. W258-63, 2006.
- [21] K. Kubota, T. Yoneyama-Takazawa, and K. Ichikawa, "Determination of sites citrullinated by peptidylarginine deiminase using <sup>18</sup>O stable isotope labeling and mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 19, pp. 683-8, 2005.
- [22] S. R. Ramakrishnan, R. Mao, A. A. Nakorchevskiy, et al., "A fast coarse filtering method for peptide identification by mass spectrometry," *Bioinformatics*, vol. 22, pp. 1524-1531, 2006.