Computing Post-translation Modification using FTMS

Wei Shen[1] **Wing-Kin Sung**[1,2] **Siu Kwan SZE**[2]
[1]*Department of Computer Science, School of Computing, National University of Singapore*
[2]*Genome Institute of Singapore*
Email :shenwe@comp.nus.edu.sg,, ksung@comp.nus.edu.sg, szen@gis.a-star.edu.sg

ABSTRACT: Post translational modifications (PTMs) discovery is an important problem in proteomic. In the past, people discover PTMs by Tandem Mass Spectrometer based on "bottom-up" strategy. However, such strategy suffers from the problem of failing to discover all PTMs. Recently, due to the improvement in proteomic technology, Taylor *et al.* proposed a database software to discover PTMs with "topdown" strategy by FTMS, which avoids the disadvantages of "bottom-up" approach. However, their proposed algorithm runs in exponential time, requires a database of proteins, and needs prior knowledge about PTM sites. In this paper, a new algorithm is proposed which can work without a protein database and can identify modifications in polynomial time. Besides, no prior knowledge about PTM sites is needed.

1 INTRODUCTION

Protein is a sequence of amino acids. Recent advance in mass spectrometry technology allows us to recover the sequence of a protein [11]. Moreover, some proteins will undergo a process called post-translational modification (PTM). This process modifies some amino acids in a protein and changes its function. One well-known example is the methylation of histones. This process changes the function of histones and affects the formation of chromatin [3,7,16]. It in turn affects the gene regulation activity. Hence, it is important to have some methods to identify the post-translational modification of a protein. Generally, there are two classes of methods for locating post-translational modification. The first approach is based on the bottom-up spectrum [4,21]. In this case, protein is first digested into a collection of peptides with about 10 amino acid residues. Then their peptide masses got from the experiment are matched against the list of peptide masses expected from the protein sequence. The non-matching masses could imply the post-translational modifications. Those peptides are further fragmented to generate the "tandem mass spectrum" which is then used to identify the peptide and to localize its modification. Normally, peptides are identified by matching the experimental spectrum against the theoretical spectra corresponding to the peptides in a database. There are several different algorithms, such as Peptide Sequence Tag [9], SEQUEST [2], and Mascot [12]. Sequence Tag searches peptides in the database by allowing partial peptide mass unmatched. The latter two, which were originally used to identify unmodified peptides, can be used to identify modification by taking more than one possible amino acid molecular weight into account, depending on the modification considered [1,8]. However, such approaches generate more answers and the modified peptides identified are less certain. Another algorithm is based on de novo peptide sequencing [14]. It uses a new notion of spectral similarity that allows one to identify related spectra considering the multiple modifications. But the results show that this method is not successful due to the limitation of de novo sequencing.

Although the bottom-up approach is widely used, it may miss some modifications since the coverage of peptide fragments got from the digestion is not 100%. Even worst, the bottom-up approach becomes more unreliable when we study large protein. When the protein size is big, the number of fragments increases. The common spurious peptide mass can be mistaken to be a modified peptide mass. In contrast, these problems can be solved by using top-down tandem spectrum [5,15,18,19].

In top-down protein sequencing, instead of digesting the modified protein into peptides, the modified protein is analyzed directly by ECD-FTMS[10,17,20,22]. Many copies of the modified protein are fragmented by ECD and each copy is charged and broken randomly into two pieces at some peptide bond. The resulting set of fragment ions are passed into a FTMS machine and their masses per charge ratios are measured and generated a spectrum (see Figure 1.1 for an example). The spectrum consists of a lot of peaks, each of which is probably produced by many copies of the same fragment ion. The position of the peak in the spectrum represents the mass to charge ratio of the corresponding fragment ion. The height of the peak indicates its intensity. Given the spectrum, the computational problem is to identify all the modifications.

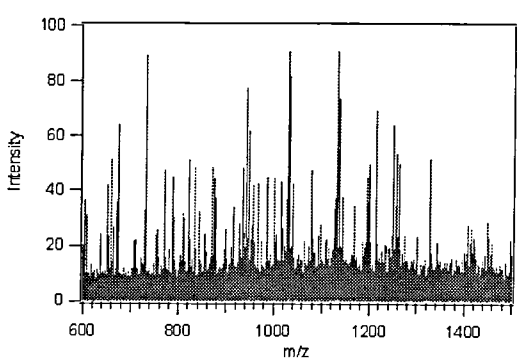

Figure 1.1: FTMS Spectrum

The only previous work is by Taylor et al [13,19] and they suggested identifying modifications using database-searching approach. They first construct a database that contains the intact proteins with different combinations of modifications. However, there are exponential possible combinations of modifications. To reduce the database size, the included modifications need to satisfy some prior biology knowledge. Then, the database is searched to

identify a modified protein that best matches the spectrum. The limitation of the database-searching algorithm is that it is based on the prior biology knowledge. If modifications occur at some unknown sites, their method may not work. This paper presents a dynamic programming algorithm to identify modifications without any prior knowledge. Thus, by using our method, novel modification sites can be discovered. More importantly, our algorithm runs in polynomial time instead of exponential time.

The rest of the paper is organized as follows: Section 2 details the PTM problem. Section 3 gives a dynamic programming algorithm to solve the problem. Lastly, Section 4 shows the experimental results.

2 PRELIMINARY

Consider a FTMS spectrum M of a particular post-translational modification of a certain protein H. This section describes the problem of determining the most likely posttranslational modified protein form H^m of H that best fits the spectrum M.

2.1 The Ion Mass Calculation

Amino acids consist of 20 different types. We use A to denote the alphabet of the 20 amino acids. For any amino acid, $wt(a)$ is denoted to be its monoisotopic mass. The maximum and the minimum masses among all amino acid types are 186.08 Dalton and 57.02 Dalton respectively. Suppose there are t possible types of modifications for a certain protein. Including the non-modification case, there are $t+1$ types of modifications in total. We use Γ to denote the alphabet of the $t+1$ types of modifications. For any $m \in \Gamma$, $wt(m)$ is denoted as the mass of this modification. The maximum and the minimum masses among all types of modifications are $m_{\max}$ Dalton and 0 Dalton respectively.

In total, the maximum and the minimum masses of a modified amino acid are $186.08 + m_{\max}$ Dalton and 57.02 Dalton, respectively.

2.2 FTMS Spectrum

In the experiment, every fragment cleaved from H^m can have different charged states and generate a few different peaks in the spectrum. Fortunately, each isotopic cluster in the FTMS can be assigned a charge (z) based on the one Dalton inter-peak spacing (1/z) [6]. We can preprocess the FTMS spectrum and convert all peaks of different charged states into single charged equivalents. Furthermore, every isotopic cluster is represented by a peak at the monoisotopic mass. Its intensity is the sum of the intensities of all peaks in the corresponding isotopic clusters. Therefore, from now on, every fragment ion is assumed to be single charged and its peak is at its monoisotopic mass. In other word, a spectrum can be represented by $M = \{(x_i, y_i) | 1 \le i \le num\}$ where num is the total number of peaks in M. Below, we describe the computation of the mass for every fragment ion of a protein. Consider a peptide sequence $H = a_1 a_2 a_3 \mathrm{K}\, a_n$. We denote $wt(H) = \sum_{1 \le i \le n} wt(a_i)$. Because of the extra H_2O, the actual mass of H is $wt(H)+18.01$.

As shown in Figure 2.1, FTMS fragments the peptide H into five different types of ions. The ions can be classified into two groups: the N-terminal group and the C-terminal group. The N-terminal group contains a-ion, b-ion and c-ion while the C-terminal group contains y-ion and z-ion

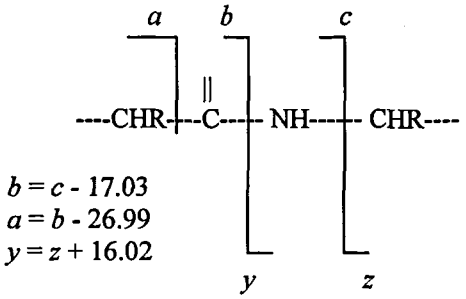

Figure 2.1: Fragment Ions of ECD

Consider the ith prefix of H, which is $a_1 a_2 \mathrm{K}\, a_i$. Let x be $wt(a_1 a_2 \mathrm{K}\, a_i)$. Then, the corresponding masses of the a-ion, b-ion and c-ion in the N-terminal group are x-26.99, x, and x+17.03 respectively . We denote $N(x)$ as $\{x - 26.99,\ x,\ x + 17.03\}$. Similarly, for the ith suffix $a_i a_{i+1} \mathrm{K}\, a_n$ of H, let $x = wt(a_i a_{i+1} a_{i+2} \mathrm{K}\, a_n)$ be its mass. The corresponding masses of the y-ion and z-ion in the C-terminal group are x +18.01 and x + 1.99 respectively. We denote $C(x)$ as $\{x+18.01,\ x+1.99\}$.

Based on the above equations, ideally, the spectrum of the protein H should have a list of peaks whose masses are belonging to

$$L(H) = \Upsilon_{1 \le i \le n-1} \{N(wt(a_1 a_2 \Lambda\, a_i)) \cup C(wt(a_{i+1} a_{i+2} \Lambda\, a_n))\} \quad (1)$$

Now, H is modified and let $H^m = a'_1 a'_2 \mathrm{K}\, a'_n$ be the resultant modified peptide where each a'_i is the residue formed after a_i is modified by m_i. Note that $wt(a_i') = wt(a_i) + wt(m_i)$. Then, $wt(H^m)$ can be defined similarly and the actual mass of H^m equals $wt(H^m)+18.01$. In addition, in the ideal case, the spectrum of the peptide H^m should have a list of peaks whose masses are belonging to $L(H^m)$.

Given a modified peptide H^m, let $M = \{(x_i, y_i) | 1 \le i \le num\}$ be the experimental FTMS spectrum of H^m with num peaks where, for the ith peak (x_i, y_i), x_i is its mass (position) and y_i is its intensity (height) in the spectrum.

Ideally, we expect M contains a list of peaks whose masses belong to $L(H^m)$. Since the experimental data is not accurate, the positions of the peaks may be shifted by a little bit. Let δ >0 be the error of the spectrometer. Due to the high accuracy of FTMS, we assume δ <0.5 in this paper. For any peak (x,y) of M and $w \in L(H^m)$, if

$|w-x| \le \delta$, we say that the peak (x,y) is explained by w. Denote $L_M(H^m)$ to be the set of all possible peaks in M that can be explained by some w in $L(H^m)$, that is

$$L_M(H^m) = \{(x_i,y_i) \in M \mid \exists w \in L(H^m) \text{ such that } |w-x_i| \le \delta\} \quad (2)$$

2.3 Modification Identification Problem

It is obviously that the more and the higher peaks in M are explained by $L_M(H^m)$, the higher the chance that M is the spectrum generated by H^m. Here, we use a simple function

$$G(M,H^m) = \sum_{(x_i,y_i) \in L_M(H^m)} y_i . \quad (3)$$

Note that the bigger the value $G(M,H^m)$, the more likely that H^m is the correct modified protein for H.

The problem is summarized as follows: Consider a protein sequence $H = a_1 a_2 a \mathrm{K}\, a_n$. The mass of H is $W = wt(H) + 18.01$. Let W^m be the mass after H is modified and δ is the error bound of the mass spectrometer. We would like to compute the modified peptide $H^m = a'_1\, a'_2 \mathrm{K}\, a'_n$ such that (1) every a'_i is the residue formed after a_i is modified by some $m_i \in \Gamma$, (2) $|wt(H^m)+18.01-W^m| \le \delta$ and (3) $G(M,H^m)$ is maximized.

3 ALGORITHM

3.1 Dynamic Algorithm

Consider a protein sequence $H = a_1 a_2 \Lambda\, a_n$ of weight $W = wt(H)+18.01$. Suppose H is modified. Through mass spectrometry, we obtain an experimental spectrum M for the modified protein and also deduce that the mass of the modified protein is W^m. Now, among exponential possible modification combinations of the protein H, this section describes a dynamic programming solution to find an optimal modification combination H^m which maximizes $G(M,H^m)$.

Let F(M, H) = max{ $G(M,H^m)$ | H^m is some modification of H such that $|wt(H^m)+18.01-W^m| \le \delta$ }. Let the first i and the last j modifications of H be $P_i = a_1'a_2'\Lambda\, a_i'$ and $S_j = a_{n-j+1}'a_{n-j+2}'\Lambda\, a_n'$. For any $k \le i$, the set of peaks corresponding to $a_1'a_2'\Lambda\, a_k'$ is $N(wt(a_1'a_2'\Lambda\, a_k')) \cup C(W^m - 18.01 - wt(a_1'a_2'\Lambda\, a_k'))$. Similarly, for any $k \ge n-j+1$, we can show that the set of peaks corresponding to $a_k'a_{k+1}'\Lambda\, a_n'$ is $N(W^m - 18.01 - wt(a_k'a_{k+1}'\Lambda\, a_n')) \cup C(wt(a_k'a_{k+1}'\Lambda\, a_n'))$. Let $L_M(W^m,P_i,S_j)$ be the union of all these peaks. Then, we define

$$G(M,W^m,P_i,S_j) = \sum_{(x,y) \in L_M(W^m,P_i,S_j)} y . \quad (4)$$

Let $D = W^m - W$. For $0 \le i+j+1 \le n$ and $0 \le q_1+q_2 \le D$, let $T[i,q_1,j,q_2] = \max\{ G(M,W^m,P_i,S_j) \mid P_i$ and S_j are the modifications of $a_1 \Lambda\, a_i$ and $a_{n-j+1} \Lambda\, a_n$ respectively such that $wt(P_i) = wt(a_1 \Lambda\, a_i) + q_1$ and $wt(S_j) = wt(a_{n-j+1} \Lambda\, a_n) + q_2 \}$.

Lemma 1. *The score of the optimal modification of H with respect to M, that is, F(M,H) is*

$$\max_{|D'-D| \le \delta} \max_{0 \le i \le n, 0 \le q \le D', m \in \lambda} T[i,q,n-i-1,D'-q-m]. \quad (5)$$

Lemma 2. $T[i,q_1,j,q_2]$ *satisfies the following equations.* ($|p_i+q_1-s_j-q_2| \le 186.08 + m_{\max}$)

- *Basis:* T[0,0,0,0] = 0
- *Recurrence: For* $i > 0$, $0 \le j \le n-i-1$, $q_1 \ge 0$, $0 \le q_2 \le D-q_1$ *we have the following recursive function.*

$$T[i,q_1,j,q_2] = \max \left(m \in \Gamma \max \begin{cases} T[i-1,q_1-wt(m),j,q_2] + score(p_i+q_1, s_j+q_2) \\ \quad \text{if } p_{i-1}+q_1-wt(m) < s_j+q_2 \\ T[i,q_1,j-1,q_2-wt(m)] + score(\overline{s_j+q_2}, \overline{p_i+q_1}) \\ \quad \text{if } p_i+q_1 \ge s_{j-1}+q_2-wt(m) \end{cases} \right) \quad (6)$$

where

$$\bar{v} = W^m - v; \quad (7)$$

$$p_i = wt(a_1 a_2 \ldots a_i); \quad (8)$$

$$s_j = wt(a_{n-j+1} a_{n-j+2} \ldots a_n) + 18.01 \quad (9)$$

$$score(x,y) = G(L_M(N(x) \cup C(W^m - 18.01 - x)) - L_M(N(W^m - y) \cup C(y - 18.01))) \quad (10)$$

By Lemma 1,

$$F(M,H) = \max_{|D'-D| \le \delta} \max_{0 \le i \le n, 0 \le q \le D'} T[i,q,n-i-1,D'-q-m] \quad (11)$$

Hence, the target H^m can be computed as follows: first, evaluate all the entries $T[i,q_1,j,q_2]$, where $|p_i+q_1-s_j-q_2| \le 186.08+m_{\max}$ and $q_1+q_2 \le D'+\delta$, based on the above recursive formula; then, among all $|D'-D| \le \delta$, $0 \le i \le n$, $0 \le q \le D'$, and $m \in \Gamma$, find the entry $T[i,q,n-i-1,D'-q-m]$ with maximum value; finally, by backtracking, we can recover the target H^m. Figure 3.1 shows the pseudo code of the algorithm

Lemma 3: The algorithm can compute the optimal solution of the protein modification problem in $O\left(len \times \min(len, \frac{(186.08+m_{\max}) \times 2}{57.02}) \times \left(\frac{D}{\delta}\right)^2 \times \frac{\delta}{\Delta} \right)$ time.

In practice, there are still something can be done to improve the above algorithm. The following sections will introduce the several tips to accelerate the algorithm.

3.2 Change of Backtracking Algorithm

As mentioned before, ECD normally breaks about 50% of

the amino acid bonds of a protein, which means that there are still a lot of bonds not fragmented. However, cleaving the protein backbone between each modification site is critical to achieve complete modification identification and allocation[18]. If a lot of PTM sites are not broken from ECD, we cannot uniquely identify the locations of the PTM sites and many possible solutions can be generated. For example, consider a protein

Input: Total tested modification $D = W^m - W$, A peak list of the spectrum, modification list Γ, sequence of tested protein H(length is *len*), calibration of the spectrum Δ, and error bound δ of the spectrum.
Output: the maximum scored H^m of modification masses D' such that $|D'-D| \le \delta$.

1. Initialize all $T[i,j,k,l] = -\infty$; Let $T[0,0,0,0] = 0$
2. for i from 0 to *len*-1 step 1 do
3. for j from 0 to D step Δ do
4. for k from 0 to *len*-i-1 step 1 do
5. for l from 0 to $D-j$ if $|p_i + j - s_k - l| \le 186.08 + m_{\max}$ step Δ do
6. if $p_i + j < s_k + l$ for $m \in \Gamma$ such that $wt(m) + j + l \le D$
7. $T[i+1, wt(m)+j, k, l] = \max\begin{cases} T[i+1, wt(m)+j, k, l] \\ T[i,j,k,l] + score(p_{i+1} + wt(m) + j, s_k + l) \end{cases}$
8. else for $m \in \Gamma$ such that $wt(m) + j + l \le D$
9. $T[i,j,k+1,l+wt(m)] = \max\begin{cases} T[i,j,k+1,l+wt(m)] \\ T[i,j,k,l] + score(\overline{s_{k+1} + l + wt(m)}, \overline{p_i + j}) \end{cases}$
10. Compute the best $T[i,j,k,l]$ for all i, j, k, l and the $m \in \Gamma$ satisfying $i = len - k - 1$ and $|j + l + wt(m) - D| \le \delta$
11. Use back tracking to construct the H^m

Figure 3.1

$H = a_1 a_2 \Lambda\, a_i \Lambda\, a_j \Lambda\, a_k \Lambda\, a_n$ and assume a_j is modified. Suppose ECD does not cleave at any site between a_i and a_k. Since we have no knowledge on the amino acids between a_i and a_k, a normal backtracking routine will report (k-i+1) possible solutions where the modification occurs at amino acid a_x for $i \le x \le k$. When there are more amino acids and more modifications occurring between a_i and a_k, the possible cases will grow exponentially and it is inefficient to backtrack all possible solutions.

To solve this problem, we change the backtracking algorithm. Instead of tracing all the solutions, we just report that the modification occurs in a certain range. Using the above example, the modified backtracking algorithm will just output one solution and report there is one modification between a_i and a_k. This is realized as follows:

Consider $H = a_1 a_2 \Lambda\, a_n$ and a spectrum M of the modified H. Let

$$p_i = wt(a_1 a_2 ... a_i) \quad (8)$$

$$s_j = wt(a_{n-j+1} a_{n-j+2} ... a_n) + 18.01 \quad (9)$$

$$L_M^P(i,q) = L_M(N(p_i + q) \cup C(W^m - 18.01 - p_i - q)) \quad (12)$$

$$L_M^S(j,q) = L_M(N(W^m - s_j - q) \cup C(s_j + q - 18.01)) \quad (13)$$

To help the modified backtracking, when we fill in the table T, we need to maintain the parent pointers using the following two steps.

1. If $p_i + q_1 < s_j + q_2$, we set

$$T[i+1, q_1 + wt(m), j, q_2]\text{'s parent} = \begin{cases} T[i,q_1,j,q_2] \\ \quad When \quad L_M^P(i,q_1) \ne \phi \\ T[i,q_1,j,q_2]\text{'s parent} \\ \quad When \quad L_M^P(i,q_1) = \phi \end{cases}$$

2. If $p_i + q_1 \ge s_j + q_2$, we set

$$T[i,q_1,j+1,q_2 + wt(m)]\text{'s parent} = \begin{cases} T[i,q_1,j,q_2] \\ \quad When \quad L_M^S(j,q_2) \ne \phi \\ T[i,q_1,k,q_2]\text{'s parent} \\ \quad When \quad L_M^S(j,q_2) = \phi \end{cases}$$

The above parent pointers ensure that we only trace back to $T[i,q_1,j,q_2]$ entry where the mass $p_i + q_1$ or $s_j + q_2$ can be explained by some peaks in the spectrum. Our modified backtracking algorithm will trace back based on these parent pointers. Thus, we can avoid generating many solutions through backtracking and improve the efficiency.

3.3 Changed the Modification Mass Storing Method in the Table Element

In the above algorithm, we only constraint that $0 \le q_1 + q_2 \le D$ in $T[i,q_1,j,q_2]$. However, in most cases, only several modification mass values in the range from 0 to D are feasible. So, it is waste of space and time to construct and fill a table $T[i,q_1,j,q_2]$ for all q_1, q_2 such that $0 \le q_1 + q_2 \le D$.

Thus we change the way to store modification mass values such that q_1 and q_2 only represent the meaningful value. We do this through the following steps:

1. Construct a mass array E such that, for any mass m, $E[m] = 1$ if m is equal to the sum of some modification masses; otherwise $E[m] = 0$. The E array can be constructed in $O\left(\frac{D}{\Delta}\right)$ time.
2. Among all possible masses $0 \le m \le D$, let m_1, m_2, ..., m_n be masses such that $E[m_i]=1$ and $E[D-m_i]=1$. Let F be an array such that $F[1]=m_1$, $F[2]=m_2$, ..., $F[n]=m_n$.

3. Now we can construct table $T[i,q_1,j,q_2]$ with $0 \le q_1, q_2 \le n-1$ and the modification masses can be got from $F[q_1]$ and $F[q_2]$.

For example, in histone, the possible modifications are methylation, phophorylation, ADP ribosylation, biotinylation and ubiquitination. So, the set of possible modification masses is {14.02, 42.01, 79.96, 541.06, 226.08, 8560.62}. If the total modification mass is 93.98 Dalton, we conclude that the only possible modification combination is methylation+phophorylation, which means that the possible values for q_1 and q_2 are either 93.98, 79.96, 14.02 or 0. If we use the original storing method, the table will have all the elements with q_1 and q_2 from 0 to 93.98. By using the new way to store modification masses, we have $F = \{0, 14.02, 79.96, 93.98\}$ and $n = 4$. Thus we can construct table $T[i,q_1,j,q_2]$ with $0 \le q_1, q_2 \le 3$.

4 EXPERIMENT RESULT

We use histones to test our algorithm. There are six types of modifications which can affect the amino acids in the histone sequences. They are methylation(14.02 Dalton), acetylation(42.01 Dalton), phophorylation(79.96 Dalton), ADP ribosylation(541.06 Dalton), biotinylation(226.08 Dalton) and ubiquitination(8560.62 Dalton). Among them, methylation has three status, mino-, di-, or trimethylation. Thus, including "no modification", there are 9 elements in Γ.

To compare with the database searching method [13], we first construct an artificial data set which is similar to the experimental data stated in [13] to test our program. The tested histone is H4 with 112 Dalton above its unmodified mass and the known modification locations are positions 1, 16, and 20. Below figure graphically shows the modifications.

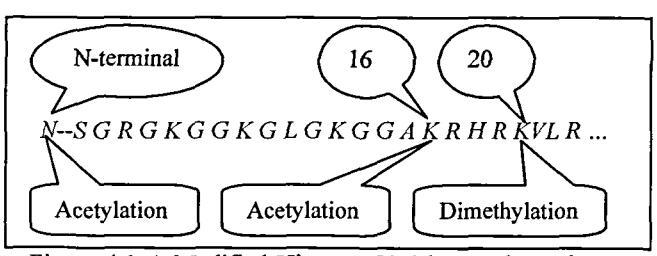

Figure 4.1 A Modified Histone H4. The numbers above the sequence show the positions of the modified amino acids in the histone and the modification type is remarked below the sequence.

In [13], the authors did an ECD/FTMS experiment on H4 and they reported all matched peaks in their webpage. We generate an artificial ECD/FTMS spectrum by randomly introducing 100% noise peaks into the spectrum. By the algorithm, we discover there is an acetylation at N-terminal, an acetylation at position 16 and two methylations (or one di-methylation) at positions 20-21. The uncertainty at positions 20-21 is because of the loss of important peaks resulted from the modification site. Below figure visualizes the modifications.

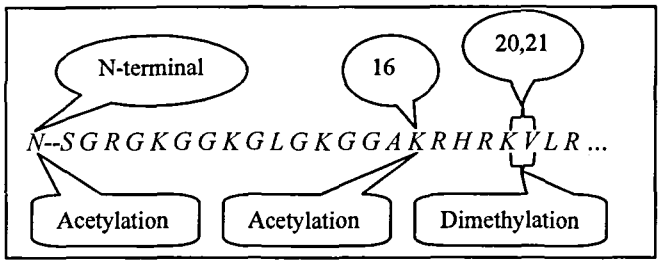

Figure 4.2: Result Got by Our Algorithm

To test the robustness of the algorithm, we gradually delete some matched peaks from the original spectrum. Table 4.1 shows the results.

Deleted site *i*	Modification allocation	# of solutions	# of solutions (original backtracking)
19	N-terminal(Ac), 16(Ac), 19-21(2Me)	1	6
19 to 18	N-terminal(Ac), 16(Ac), 18-21(2Me)	1	10
19 to 17	N-terminal(Ac), 16(Ac), 17-21(2Me)	1	15
19 to 16	N-terminal(Ac), 16-21(2Me+1Ac)	1	90
19 to 15	N-terminal(Ac), 15-21(2Me+1Ac)	1	147

Table 4.1. This table shows the modification allocation when we delete the peaks generated by the cleavage after *i*th amino acid from the spectrum. The 2nd last column shows the number of solutions reported by our algorithm. The final column shows the number of solutions reported if we use the original backtracking method.

Table 4.1 shows that the algorithm can discover the modifications even when more important peaks are deleted. More importantly, our algorithm only report one solution. If we use the original backtracking method, many solutions are reported. Note that the number of solutions increases exponentially when more and more correct peaks are deleted.

We should note that our algorithm does not require any prior knowledge of the modification site. If such knowledge is available, a better solution can be obtained. For example, in Figure 4.2, if we have the prior knowledge that V could not be modified by methylation, we can conclude that the position 21 is not modified while position 20 is modified by a dimethylation.

Besides, we got a real spectrum for histone H2A to test our algorithm. Based on the literature, the only known modification for H2A is acetylation and it occurs at the N-terminal. By running our program on the real spectrum, we report that there is an acetylation before the 6th amino acid. The following figure visualizes the result.

We have investigated why the algorithm fails to find the exact location of the modification. After checking the spectrum, we found that the spectrum has no peak generated by the cleavage of the first five amino acids of H2A.

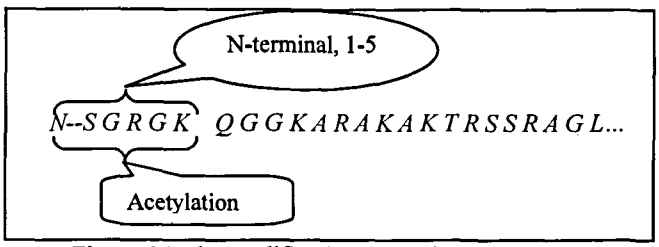

Figure 4.3 The modification allocation of H2A

5 CONCLUSION

This paper proposed a dynamic programming algorithm via a "top-down" mass spectrometry to solve the post translational modifications (PTMs) problem. There are many advantages of this new method. First, our method can work without a protein database. Second, no prior knowledge of the modification sites in the protein is required. Last but not least, it can identify the modifications in polynomial time, which is very efficient compared to the widely used database searching method.

There are several possible future works. First, current work needs to know the set of modification types. We would like to explore if it is possible to detect PTM sites without knowing the modification types in advance. Second, the current work did not explore the intensity pattern of the FTMS such as the intensity relationship between different ions. We would like to utilize those intensity patterns to give a better scoring function to improve the performance of the algorithm. Finally, we hope to do further experiments to test the performance of our algorithm.

REFERENCES

[1] D.M. Creasy, J.S. Cottrell. Error-tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* 2, 1426-1434, 2002.

[2] J.K. Eng, A.L. McCormack, J.R. III. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976-989, 1994.

[3] O. Fernandez-Capetillo, S.K. Mahadevaiah, A. Celeste, P.J. Romanienko, R.D. Camerini-Otero, W.M. Bonner, K. Manova, P. Burgoyne, A. Nussenzweig. H2AX is required for chromatin remodeling and inactivation of sex chromosomes in male mouse meiosis. *Dev. Cell.*, 4, 497-508, 2003.

[4] S.B. Ficaaro, M.L. McCLeland, P.T. Stukenberg, D.J. Burke, M.M. Ross, J. Shabanowitz, D.H. Hunt, F.M. White. Phosphoproteome analysis by mass spectrometry and its application to Saccharomyces cerevisiae. *Nat. Biotechnol*, 20, 301-305, 2002.

[5] Y. Ge, B.G Lawhorn, M. ElNaggar, E. Strauss, J.H. Park, T.P. Begley, F.W. McLafferty. Top Down Characterization of Larger Proteins (45 kDa) by Electron Capture Dissociation Mass Spectrometry. *J. Am. Chem. Soc.*, 124, 672-678, 2002.

[6] D.M. Horn, R.A. Zubarev, F.W. McLafferty. Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. *J. Am. Soc. Mass Spectrom.*, 11, 320-332, 2000.

[7] N.J. Krogan, M. Kim, A. Tong, A. Golshani, G Cagney, V. Canadien, D.P. Richards, B.K. Beattie, A. Emili, C. Boone, A. Shilatifard, S. Buratowski, J. Greenblatt. Methylation of Histone H3 by Set2 in Saccharomyces cerevisiae Is Linked to Transcriptional Elongation by RNA Polymerase II. *Mol. Cell. Biol.*, 23, 4207-4218, 2003.

[8] M.J. MacCoss, C.C. Wu, J.R. III. Yates. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.*, 74, 5593-5599, 2002.

[9] M. Mann, M. Wilm, Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390-4399, 1994.

[10] F.W. McLafferty. High-Resolution Tandem FT Mass Spectrometry above 10 kDa. *Acc. Chem. Res.*, 27, 379-386. 1994.

[11] A. Pendey, M. Mann. Proteomics to study genes and genomes. *Nature*, 405, 823-826, 2000.

[12] D. Perkins, D. Pappin, D. Creasy, J. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551-3567, 1997.

[13] J.J. Pesavento, Y.B. Kim, G.K. Taylor, N.L. Kelleher. Shotgun Annotation of Histone Modifications: A New Approach for Streamlined Characterization of Proteins by Top Down Mass Spectrometry. *J. Am. Chem. Soc.* 126, 3386-3387, 2004.

[14] P.A. Pevzner, V. Dancik, C.L. Tang. Mutation-tolerant protein identification by mass spectrometry. *J. Comp. Biol.*, 7, 777-787, 2000.

[15] G.E. Reid, S.A. McLuckey. 'Top down' protein characterization via tandem mass spectrometry. *J. Mass Spectrom.* 37, 663-675, 2002.

[16] H. Santos-Rosa, R. Schneider, A.J. Bannister, J. Sherriff, B.E. Bernstein, N.C. Emre, S.L. Schreiber, J. Mellor, T. Kouzarides. Active genes are tri-methylated at K4 of histone H3. *Nature*, 419, 407-411, 2002.

[17] S.D.H. Shi, M.E. Hemling, S.A. Carr, D.M. Horn, I. Lindh,; F.W. McLafferty. Phosphopeptide /Phosphoprotein Mapping by Electron Capture Dissociation Mass Spectrometry. *Anal. Chem.*, 73, 19-22, 2001.

[18] S.K. Sze, Y. Ge, H. Oh, F.W. McLafferty. Top-down mass spectrometry of a 29-kDa protein for characterization of any posttranslational modification to within one residue. *Proc. Natl. Acad. Sci. U.S.A.* 99, 1774-1779. 2002.

[19] G.K.Taylor, Y.B. Kim, A.J. Forbes, F. Meng, R. McCarthy, N.L. Kelleher. Web and Database Software for Identification of Intact Proteins Using "Top Down" Mass Spectrometry. *Anal. Chem.* 75, 4081-4086, 2003.

[20] E.R. Williams. Tandem FTMS of Large Biomolecules. *Analytical Chemistry News & Features*, 179A-185A. 1998.

[21] M.R. Wilkins, E. Gasteiger, A.A. Gooley, B.R. Herbert, M.P. Molloy, P.A. Binz, K. Ou, J.C. Sanchez, A. Bairoch, K.L. Williams, D.F., Hochstrasser. High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.*, 289, 645-657, 1999.

[22] R.A. Zubarev, N.L. Kelleher, F.W. McLafferty. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *J. Am. Chem. Soc.*, 120, 3265-3266. 1998.