PROTEIN MODIFICATION AND PEPTIDE IDENTIFICATION FROM MASS SPECTRUM

SHEN WEI

(B.Eng., Shanghai JiaoTong University)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

2005

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Sung Wing Kin, who provided me with years of professional guidance and personal example during my master study at NUS. His invaluable motivation, advice and comments have the largest immediate influence on this thesis.

Genome Institute of Singapore has facilitated my research by providing precious data. Among all the staffs there who have given me helpful assistance, I would like to present my special thanks to Dr. SZE Siu Kwan.

I gratefully acknowledge the financial support of National University of Singapore in the form of my research scholarship. Besides, I would also like to express my gratitude for the excellent environment and facilities provided by NUS.

My heartfelt thanks go to my friends for their constant love and support. Yin Hainan, Qian Bo and Shi Yijing have each given me years of friendship and have done more for me than I could ever hope to repay.

Last but not least, I would like to express my sincerest thanks to my parents. Their love and understanding are my impetus to do the research during my graduate studies.

TABLE OF CONTENTS

SUMMARY

Proteome is the complete set of proteins produced by the genome. It is much more complex than either the genome or the transcriptome. Moreover, protein products can not be accurately predicted from genome by decoding genomic sequences. As a result, proteomics, the large-scale study of the proteome is a growing research area in the post genomic era. The determination of the amino acid sequence of a protein is the first step toward the structure and the function of the protein and it is a crucial requirement for the success of proteomics. In this thesis we study two problems related to protein sequencing via mass spectrum.

First, we discuss the protein post translational modifications (PTMs) identification via “top-down” mass spectrometry. In literature, database searching method is used to identify the modification. In this thesis, we propose a dynamic programming algorithm to solve this problem. Compared with the widely used database searching method, our new algorithm has several advantages. First, our method can work without a protein database. Second, there is no prior knowledge of the modification sites in the protein needed. Last but not the least, it can identify the modifications in polynomial time, which is very efficient compared to the widely used database searching method.

Second, we discuss the de novo peptide sequencing problem. There are two kinds of algorithms to automate peptide sequencing in literature. One is the database searching method and another is de novo peptide sequencing. Scoring function is an important component for both methods. In literature, not a lot has been done to incorporate the intensity pattern into the scoring function. We propose a new de novo peptide sequencing algorithm DTseq, which uses an intensity-based scoring function. The scoring function is based on two competing models. One of them is a decision tree probability model which fully explores the factors that influence the intensity pattern in the spectrum. The decision tree model estimates the likelihood of certain observed intensity given the local chemical and physical attributes of the fragment. Besides, a random probability model is used to estimate the probability that certain peak is actually a noise peak in the spectrum. To test our algorithm, we compare DTSeq with two best de novo peptide sequencing algorithms: Peaks and PepNovo. The results show that DTSeq performs best among all the three algorithms. It can obtain the longest maximum subsequence of predicted peptide as well as the highest prediction accuracy.

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation

Human genome[I] contains the complete set of genes required to build a functional human being. Nowadays, large quantities of deoxyribonucleic acid (DNA) have been sequenced, cataloged, and annotated. However, this information is not enough to infer biological function because the genome is only one source of information [8, 24]. The transcription of genes is the first stage of gene expression and is followed by the translation of messenger RNA to produce proteins.

Proteome is the complete set of proteins produced by the genome. It is much more complex than either the genome or the transcriptome[II]. Moreover, protein products can not be accurately predicted from genome by decoding genomic sequences. This is because each protein can be chemically modified in different ways after synthesis, which cannot be deduced from gene sequence. The modifications add chemical state to the basic protein sequence and cause the change in the protein function and cell signaling. In addition, the proteome is also very dynamic. It varies considerably in different circumstances due to different patterns of gene expression and different patterns of protein modification. As a result, proteomics, the large-scale study of the proteome is a growing research area in the post genomic era.

[I] The entire complement of genetic material in a chromosome set.

[II] The full complement of activated genes, mRNAs, or transcripts in a particular tissue at a particular time.

A key requirement for the success of proteomics is the ability to identify unambiguous proteins in complex mixtures. The determination of the amino acid sequence of a protein is the first step toward the studying of the structure and the function of this protein. Moreover, some proteins will undergo a process called post-translational modifications (PTMs). 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[12, 23, 40]. It in turn affects the gene regulation activity. Hence, it is important to have some methods to get the protein sequence and identify the post-translational modification of a protein. Recently, mass spectrometry (MS) has become the method of choice for the rapid identification of proteins and the characterization of post-translational modification[34].

1.2 FTMS and LC/MS/MS

1.2.1 Fragmentation

Generally in mass spectrometry experiments, proteins or peptides break along their backbones between successive amino acids during the stage of fragmentation. A protein P is a sequence of n amino acids, $P = a_1 a_2 \cdots a_n$, the single breakage along the protein's backbone results in a prefix fragment $a_1 a_2 \cdots a_i$ (N-terminal fragment) and suffix fragment $a_{i+1} a_{i+2} \cdots a_n$ (C-terminal fragment). Since the fragments retain a charge, they are also called the fragment ions and they can be detected by a mass spectrometer.

The fragmentation results are shown in the mass spectrum (Figure 1.1). The spectrum consists of many peaks, each of which is generated by many copies of one fragment ion. The position of the peak represents the mass/charge ratio of the corresponding fragment ion, and the height of the peak indicates the relative intensity of the fragment ion. As a result, different peptides/proteins usually produce different spectra. Then the task is to use the spectrum to determine its sequence or to identify the post-translational modifications. This step is an indispensable process and many researches have been done to automate it.

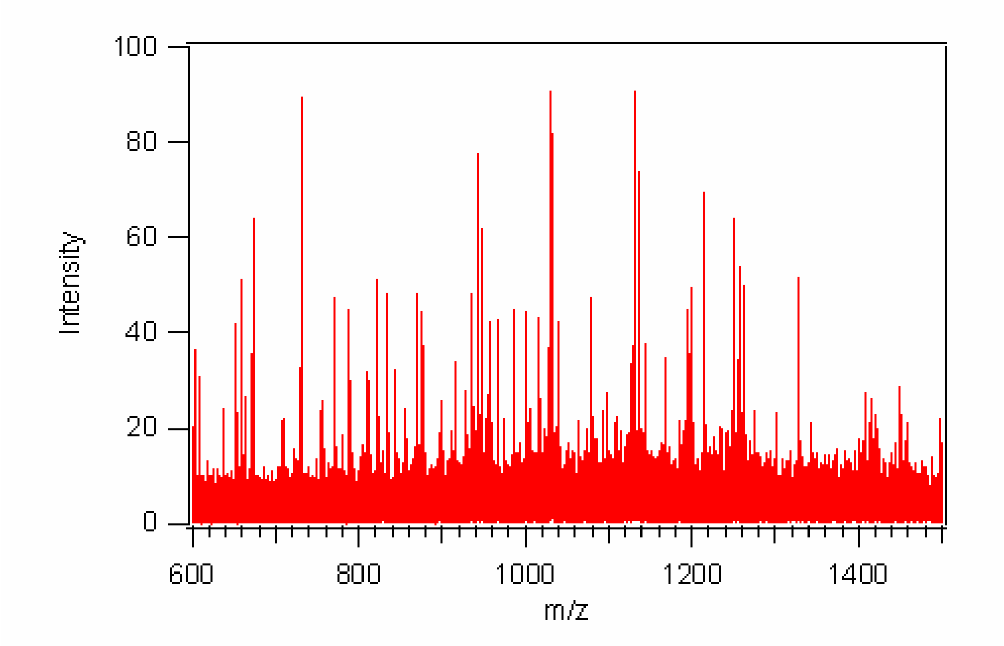


Figure 1.1: Mass spectrum

1.2.2 FTMS

There are two kinds of mass spectrometry used in the thesis. The first is the Fourier transform MS (FTMS)[33, 48]. This kind of spectrum is used to solve the top-down post-translational modification identification in proteins. To reach the goal, generally capture dissociation (ECD)[16, 42, 51] is used to cleave the whole protein. The most important property of ECD is that it can cleave any amino acid bonds except for the N-terminal side of proline in the protein sequence. Through nonergodic dissociation, ECD induces much more general backbones and derives extensive sequence information without loss of posttranslational modifications from proteins. In general, ECD can cut about 50% of the amino acid bonds of a protein sequence. Figure 1.2 illustrates the 5 possible types of fragment ions got from ECD and the most frequently appearing types of ions are the c-ion and z-ion. Fourier transform mass spectrum (FTMS) is then used to show the fragmentation result of the protein. FTMS has high precision in measuring mass/charge ratio. Another advantage of FTMS is that it can measure the masses larger than 10kDa.

$b = c - 17.03$
$a = b - 26.99$
$y = z + 16.02$

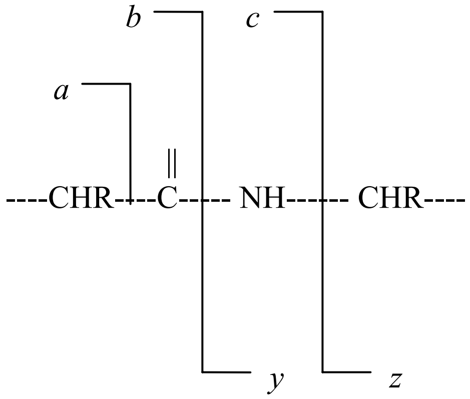


Figure 1.2: Fragment ions of ECD

1.2.3 LC/MS/MS

Another kind of spectrum is the LC/MS/MS spectrum which is used to solve the peptide sequencing problem in the thesis. In an MS/MS experiment, a mixture of proteins is first digested into peptides by enzymes such as trypsin and the masses of the intact peptides are determined, producing a ‘peptide mass fingerprint’ of the sample. Trypsin only cuts the amino acid bonds C-terminal to Lysine (K) or Arginine (R). Then a different procedure called tandem mass spectrometry is used to test the unknown peptide in the spectrum. In this step, the charged peptides are fragmented and ionized by methods such as collision induced dissociation (CID). During the CID process, peptide bonds are broken and one peptide is divided into two fragments. Fragments retaining the ionizing charge after CID have their mass/charge ratio measured by the mass spectrometer. There are usually six types of fragment ions (Figure 1.3) got by a single cleavage along the peptide's backbone directly. Among them, the b-ion and y-ion are the most frequently appeared ions in the spectrum. Besides, by some neutral losses (chemical group such as H_2O and NH_3), $b-H_2O$, $b-NH_3$, $y-H_2O$ and $y-NH_3$ are produced. However, these types of ions are much less observed than the b-ion and y-ion.

$b = c - 17$
$a = b - 28$
$y = z + 17$
$x = y + 26$

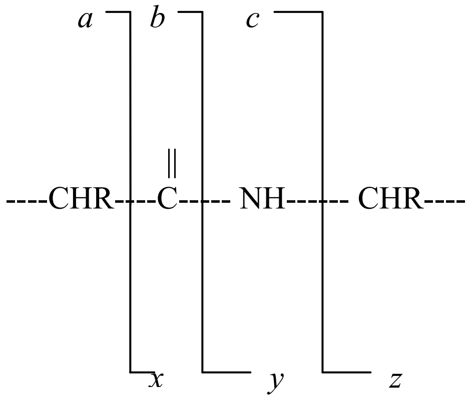


Figure 1.3: Fragment ions of CID

1.3 Organization of the thesis

In this thesis, we consider two problems related to protein sequencing. In the first problem, we propose a dynamic programming algorithm to identify the post translational modifications (PTMs) with a “top-down” strategy using FTMS. In the second problem, we propose a new probability model which fully considers the chemical and physical factors that influence the intensity pattern for de novo peptide sequencing algorithm via LC/MS/MS spectrum. The rest of the thesis is organized as follows: In Chapter 2, we introduce the protein post translational modifications (PTMs) problem; In Chapter 3, we introduce the peptide sequencing problem; Then we go to the conclusion in Chapter 4.

Chapter 2

PTMs IDENTIFICATION BY TOP DOWN MASS SPECTROMETRY

2.1 Related Work

Generally, there are two classes of methods for locating post-translational modification. The first approach is based on the bottom-up spectrum[14, 49]. 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 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[32], Sequest[10], and Mascot[35]. 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[5, 30]. However, such approaches generate more answers and the modified peptides identified are less certain. Another algorithm is based on de novo peptide sequencing[37]. 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 spectrum [39, 43, 46].

In top-down protein sequencing, instead of digesting the modified protein into peptides, the modified protein is analyzed directly by ECD-FTMS, theoretically allowing the entire sequence available for examination and giving a more complete characterization of the protein and the associated post-translational modifications.

After the spectrum is constructed, some algorithms can be applied to identify the modification. The only previous work is by Pesavento et al. and they suggested identifying modifications using database-searching approach [36]. They first construct a protein 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 knowledge of PTMs sites. If modifications occur at some unknown sites, their method may not work.

Thus we propose a new way to solve the problem. the contributions of our algorithm are as follows:

1. In the database searching method, first all possible modified protein forms are listed in the database. When the protein size and the number of possible modifications increase, the number of possible modified protein forms grows exponentially. By dynamic programming, our method can localize the modification sites and determine the modification types in polynomial time.
2. The modification can be identified without any prior knowledge about PTMs sites. In database searching method putative modification sites are needed based on prior knowledge. Thus by using our method novel modification sites can be discovered.

The rest of this chapter is organized as follows: Section2 details the PTMs problem. Section3 gives a dynamic programming algorithm to solve the problem. Lastly, Section4 shows the experimental results.

2.2 Problem Definition

Let H^m be the possible post-translational modified protein form of certain protein H and M be the spectrum got by the fragmentation of the sample of modified H. We use all the fragment masses of H^m to match the peaks in M. Intuitively, the more high intensity peaks are matched the more likely H^m is the correct post-translational modified protein for H that generates spectrum M. In this section we will give a clear picture of this problem.

2.2.1 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 $a \in A$, $wt(a)$ 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 Σ to denote the alphabet of the t+1 types of modifications. For any $m \in \Sigma$, $wt(m)$ is denoted as the mass of this modification. The maximum modification mass is $m_{\max}$ Dalton and the minimum modification mass is 0.

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

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 (e) based on the one Dalton inter-peak spacing (1/e) [17]. 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 ion is assumed to be single charged and its peak is at its monoisotopic mass. In other word, a spectrum can be represented by $M = \{(x_i, y_i) \mid 1 \le i \le num\}$ where *num* is the total number of peaks in *M*. Below, we describe the calculation of mass for every fragment ion of a protein.

Consider a protein sequence $H = a_1a_2a_3 \ldots a_n$. We denote $wt(H) = \sum_{1 \le i \le n} wt(a_i)$. Because of the extra H$_2$O, the actual mass of H is $wt(H) + 18.01$.

As shown in Figure 1.2, ECD fragments the protein 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.

Consider the *i*th prefix of H, which is $a_1a_2a_3 \ldots a_i$. Let x be $wt(a_1a_2a_3 \ldots a_i)$. 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:

$$N(x) = \{ x - 26.99,\ x,\ x + 17.03 \}$$

Similarly, for the *i*th suffix $a_ia_{i+1}a_{i+2} \ldots a_n$ of H, let $x = wt(a_ia_{i+1}a_{i+2} \ldots a_n)$ 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:

$$C(x) = \{ x + 18.01,\ x + 1.99 \}$$

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

$$L(H) = \bigcup_{1 \le i \le n-1} \left(N\left(wt(a_1 a_2 \ldots a_i)\right) \cup C\left(wt(a_{i+1} a_{i+2} \ldots a_n)\right) \right) \tag{2.1}$$

Now, H is modified and let $H^m = a'_1 a'_2 \ldots a'_n$ be the resultant modified protein 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 protein H^m should have a list of peaks whose masses are belonging to $L(H^m)$.

Given a modified protein H^m, after fragmentation, let $M = \{(x_i, y_i) \mid 1 \le i \le num\}$ be the experimental corresponding FTMS spectrum of H^m with *num* peaks where, for the *i*th 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 chapter. 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 $\overline{\overline{L(H^m)}}$ to be the set of all possible peaks in *M* that can be explained by some *w* in $L(H^m)$, that is:

$$\overline{\overline{L(H^m)}} = \{(x_i, y_i) \in M \mid \textit{there is } w \in L(H^m) \textit{ such that } |w - x_i| \le \delta\}. \tag{2.2}$$

2.2.2 Modification Identification Problem

W^m is the tested mass of the modified protein and the unmodified protein mass W can be calculated since we know the protein sequence. Based on the information we will try to determine the types and locations of modifications whose masses are summed up to W^m - W.

It is obviously that the more and higher peaks in M are explained by $L(H^m)$, the higher chance that M is the spectrum generated by H^m. In another word, the more and higher peaks in $\overline{\overline{L(H^m)}}$, the more likely H^m is the expected modified protein for H. Here, we use a simple function to evaluate the matching, that is, for any $\overline{\overline{L}}$ as a list of matched peaks:

$$G\left(\overline{\overline{L}}\right) = \sum_{(x_i, y_i) \in \overline{\overline{L}}} y_i \tag{2.3}$$

Note that the bigger the value $G(\overline{\overline{L}})$, the more likely that H^m is the correct modified protein for H.

The problem is summarized as follows:

Consider a protein sequence $H = a_1a_2a_3 \ldots a_n$, 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 $H^m = a'_1 a'_2 \ldots a'_n$

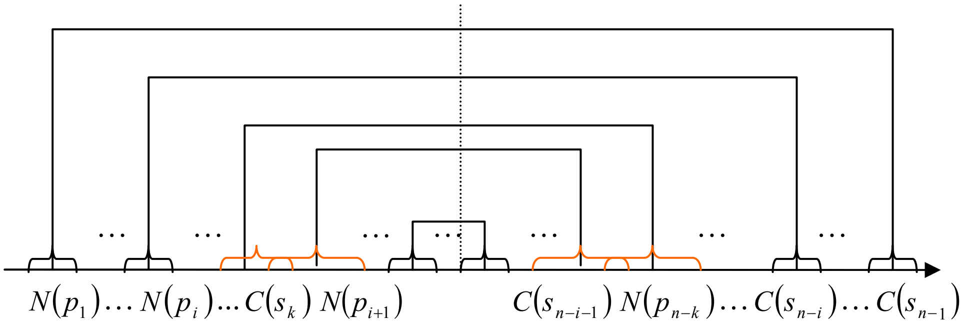
such that (1) every a'_i is the residue formed after a_i is modified by some m_i, (2) $| wt(H^m) + 18.01 - W^m | \leq \delta$ and (3) $G\left(\overline{\overline{L\left(H^m\right)}}\right)$ is maximized.

For example, consider histone H4, its unmodified mass W is 11229.34 Dalton. Moreover, after modification, experiment shows that its mass W^m is 11243.36 Dalton. Hence, by calculation, the total modification mass (W^m - W)=14.02 Dalton. Then the modified histone samples are fragmented by ECD and produce the FTMS spectrum. Given the FTMS spectrum, the algorithm described below found that the fifth amino acid (which is K) of H4 is methylated (the mass of methylation is 14.016 Dalton), which matches the ECD/FTMS best. Our founding matches with the known biology.

2.3 Algorithm

2.3.1 Dynamic Algorithm

The purpose of our algorithm is to choose the best combination of modifications for the protein so that the number and the intensity of matched peaks of this modified protein are maximized. The difficulty is that we do not know the corresponding ion types of the peaks in the experimental spectrum and one peak could be matched by more than one fragment ions generated from the protein. We need to identify whether the peak has already been matched or not.


$N(p_1)\ldots N(p_i)\ldots C(s_k)\ N(p_{i+1})$
$C(s_{n-i-1})\ N(p_{n-k})\ldots C(s_{n-i})\ldots C(s_{n-1})$

$p_j = wt(a_1 a_2 ... a_j)$ $s_j = wt(a_{n-j+1} a_{n-j+2} ... a_n)$.

$N(p_j)$ and $C(s_{n-j})$ is a complementary pair

Fig 2.1 Complementary Pair

Fortunately, the overlapping occurs only between the N-terminal ions of one prefix and the C-terminal ions of another suffix. All the N-terminal ion sets of prefixes do not overlap (the distance between two N-terminal ion sets is larger or equal to 57.02-44.02). All the C-terminal ion sets of suffixes also do not overlap (the distance between two C-terminal ion sets is larger or equal to 57.02-16.02). Figure 2.1 shows the distribution of all the ions in a spectrum. The figure shows that the overlapping could occur between $C(s_k)$ and $N(p_{i+1})$ or between $N(p_{n-k})$ and $C(s_{n-i-1})$. Thus we can solve the overlapping problem by calculating the complement pairs from the outside to the middle gradually. That is to construct optimal prefixes and suffixes step by step together[29].

Let $\overline{\overline{N(x)}}$ and $\overline{\overline{C(x)}}$ denote the peaks matched by sets $N(x)$ and $C(x)$ in the spectrum. Besides we define

$$score(x, y) = G\left(\left(\overline{\overline{N(x)}} \cup \overline{\overline{C(W^m - 18.01 - x)}}\right) \setminus \left(\overline{\overline{C(y)}} \cup \overline{\overline{N(W^m - 18.01 - y)}}\right)\right) \quad (2.4)$$

$score(x,y)$ is a simplified scoring function which sums up the intensities of all the peaks matched by the complement ion pairs of an *x* Dalton prefix excluding the peaks which are matched by another *y* Dalton suffix.

Now let *P* be one prefix sequence and *S* be one suffix sequence in a modified protein. And $|wt(S)+18.01-wt(P)| \le 186.08+m_{\max}$. $\overline{\overline{L(P,S)}}$ denote the list of peaks which can be matched by ions corresponding to any prefix sequences of *P* (including *P*) and any suffix sequences of *S*(including *S*),

When $wt(P) < wt(S)+18.01$:

$$G\left(\overline{\overline{L(Pa',S)}}\right) = G\left(\overline{\overline{L(P,S)}}\right) + score\left(wt(Pa'),wt(S)\right) \tag{2.5}$$

When $wt(P) \ge wt(S)+18.01$:

$$G\left(\overline{\overline{L(P,a'S)}}\right) = G\left(\overline{\overline{L(P,S)}}\right) + score\left(W^m-18.01-wt(a'S),W^m-18.01-wt(P)\right) \tag{2.6}$$

Detail Algorithm:

Based on the pervious part, it is obvious that by construction the modified sequence from both prefix and suffix we can solve the overlapping problem. In the following part we will describe the algorithm in detail.

The protein sequence tested is $a_1a_2a_3\cdots a_{len}$ and $W^m - W = D$. $score(x,y)$ is the scoring function. As defined before, Σ is the set of all the possible masses of modification types include 0 and A is the set of the different masses of the 20 types of amino acids.

Let $T[i,q_1,j,q_2]$ be the total score for the first i amino acids and the last j amino acids given that the total modification mass of the first i and last j amino acids equal to q_1 and q_2 respectively. ($0 \le i,j \le len-1$ and $0 \le q_1,q_2 \le D$)

$T[i,q_1,j,q_2]$ where $\left|p_i + q_1 - s_j - q_2\right| \le 186.08 + m_{\max}$ *satisfies the following equations.*

Basis:

$T[0,0,0,0] = 0$

Recurrence:

For $i > 0, 0 \le j \le len - i - 1, q_1 \ge 0, 0 \le q2 \le D - q_1$ we have the following recursive function:

$$T[i,q_1,j,q_2] = \max\left(m \in \sum \max \begin{cases} T[i-1,q_1 - wt(m),j,q_2] + score\left(p_i + q_1, s_j + q_2\right) \\ \quad if\ p_{i-1} + q_1 - wt(m) < s_j + q_2 + 18.01 \\ T[i,q_1,j-1,q_2 - wt(m)] + score\left(\overline{s_j + q_2}, \overline{p_i + q_1}\right) \\ \quad if\ p_i + q_1 \ge s_{j-1} + q_2 - wt(m) + 18.01 \end{cases} \right)$$

where $\bar{v} = W^m - v$; $p_i = wt(a_1 a_2 ... a_i)$; $s_j = wt(a_{n-j+1} a_{n-j+2} ... a_n) + 18.01$. The *Pseudo Code* is show in Figure 2.2.

Input: Total tested modification mass $D = W^m - W$;
A peak list of the spectrum;
Modification list Σ ;
Sequence of the tested protein;
Mass of unmodified protein W by calculation;
Calibration of the spectrum Δ ;
Error bound δ of the spectrum;

Output: the maximum scored modification allocations of modification masses *D'* such that $|D' - D| \le \delta$.

```
Initialize all $T[i,j,k,l] = -\infty$ ; Let $T[0,0,0,0] = 0$
for $i$ from 0 to $len$-1 step 1 do
   for $j$ from 0 to $D$ step $\Delta$ do
      for $k$ from 0 to $len$-i-1 step 1 do
         for $l$ from 0 to $D - j$ if $|p_i + j - s_k - l| \le 186.08 + m_{\max}$ step $\Delta$ do
            if $p_i + j < s_k + l$
                 for $m \in \Sigma$ such that $wt(m) + j + l \le D$
                    $T[i+1, wt(m)+j, k, l] = \max\begin{cases} T[i+1, wt(m)+j, k, l] \\ T[i,j,k,l] + score(p_{i+1} + wt(m) + j, s_k + l) \end{cases}$
            else   for $m \in \Sigma$ such that $wt(m) + j + l \le D$
                    $T[i, j, k+1, l+wt(m)] = \max\begin{cases} T[i, j, k+1, l+wt(m)] \\ T[i,j,k,l] + score(\overline{s_{k+1} + l + wt(m)}, \overline{p_i + j}) \end{cases}$
Compute the best $T[i,j,k,l]$ for all i, j, k, l and the $m \in \Sigma$ satisfying
   $i = len - k - 1$ and $|j + l + wt(m) - D| \le \delta$
Use backtracking to construct the best modification allocation
```

Figure 2.2 PTMs Algorithm

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.

Proof. For any *i,j,k,l* such that $T[i,j,k,l]$>0 and it is an optimal value, there is a prefix-suffix pair (P,S) such that $p_i + j = wt(P)$ and $s_k + l = wt(S) + 18.01$. Without loss of generality, assume that *P'a* = *P* and *wt(P')* < *wt(S)*+18.01. Based on the above algorithm, there is some *u* such that $T[i-1,u,k,l]$ corresponds to the pair (P',S). Line 8 shows that $T[i-1,u,k,l]$ must also be an optimal value if $T[i,j,k,l]$ is an optimal one. Thus $T[i,j,k,l]$ can be calculated from$T[i-1,u,k,l]$. The best modification allocation then can be got straightforwardly.

Line 5 shows that only when $|p_i + j - s_k - l| \le 186.08 + m_{\max}$, the following part will be executed. Thus for the fixed *i, j* and *l*, there are at most $\frac{(186.08 + m_{\max}) \times 2}{57.02}$ possible values for *k*. Thus there are $O\left(len \times \min(len, \frac{(186.08 + m_{\max}) \times 2}{57.02}) \times \left(\frac{D}{\delta} \right)^2 \right)$ elements in *T* need to be considered.

Since there are at most $O\left(\frac{\delta}{\Delta} \right)$ peaks in the spectrum can be explained by one mass value. The time complexity of$score(x,y)$ is$O\left(\frac{\delta}{\Delta} \right)$. Thus the time complexity of the algorithm is $O\left(len \times \min(len, \frac{(186.08 + m_{\max}) \times 2}{57.02}) \times \left(\frac{D}{\delta} \right)^2 \times \frac{\delta}{\Delta} \right)$.

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.

2.3.2 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[43]. 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 $H = a_1 a_2 a_3 \cdots a_i \cdots a_j \cdots a_k \cdots a_n$ 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 that there is a modification between a_i and a_k. This is realized as follows:

Consider $H = a_1 a_2 \cdots a_n$ and a spectrum *M* of the modified *H*. Let $p_i = wt(a_1 a_2 ... a_i)$ and $s_j = wt(a_{n-j+1} a_{n-j+2} ... a_n) + 18.01$. Let $\overline{\overline{L^p(i,q)}} = \overline{\overline{L(N(p_i + q) \cup C(W^m - 18.01 - p_i - q))}}$ and $\overline{\overline{L^S(j,q)}} = \overline{\overline{L(N(W^m - s_j - q) \cup C(s_j + q - 18.01))}}$

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

$$T[i+1, q_1 + wt(m), j, q_2]'s\ parent = \begin{cases} T[i, q_1, j, q_2] & \overline{\overline{L^p(i,q_1)}} \neq \phi \\ T[i, q_1, j, q_2]'s\ parent & \overline{\overline{L^p(i,q_1)}} = \phi \end{cases}$$

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

$$T[i, q_1, j+1, q_2 + wt(m)]'s\ parent = \begin{cases} T[i, q_1, j, q_2] & \overline{\overline{L^S(j,q_2)}} \neq \phi \\ T[i, q_1, k, q_2]'s\ parent & \overline{\overline{L^S(j,q_2)}} = \phi \end{cases}$$

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.

2.3.3 Change the Modification Mass Storing Method in the Table Element

In the above algorithm, we only constraint that $0 \leq q_1 + q_2 \leq D$ 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 $0 \leq q_1 + q_2 \leq D$.

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 $E[D\text{-}m_i]$=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 $0 \le q_1, q_2 \le 3$.

2.3.4 Scoring Function

The scoring function we used is similar to which was stated in Bin Ma's paper[26]. The difference is that in FTMS the most frequently appearing types of ions are c-ion and z-ion. Other types of ions are a, b and y.

The main idea of the scoring function is that the more and higher peaks the ions of the sequence matches in the spectrum the higher score it will get. We choose c-ion and z-ion as the main ions and other types of ions as the supporting ions.

N-terminal ions: c, a and b, the main ion is c

$$scon(u) = r(h_1 / h) \times r(h_2 / h) \times \exp\left(-\left((w' - w)/\delta\right)^2\right) \times \log h \tag{2.7}$$

C-terminal ions: z and y, the main ion is z

$$scoc(u) = r(h_1 / h) \times \exp\left(-\left((w' - w)/\delta\right)^2\right) \times \log h \tag{2.8}$$

In the above formulas *w* is the theoretical mass of the main ion and *w'* is the mass of the observed peak explained by *w*. In formula (2.7), *u* is a mass of prefix, *w* is the theoretical mass of *u's* corresponding c-ion and *w'* is the mass of the observed peak. While in formula (2.8) *u* is a mass of suffix, *w* and *w'* are the theoretical mass and the observed mass of *u's* corresponding z-ion respectively.

In the formulas, *h* is the relative intensity of the peak corresponding to the main ion and h_1 h_2 are the relative intensities of the peaks corresponding to the supporting ions. In formula (2.7), *h* is the relative intensity of c-ion and h_1 h_2 are the relative intensities of a-ion and b-ion respectively. In formula (2.8), *h* is the relative intensity of z-ion and h_1 is the relative intensity of the peak corresponding to y-ion. In the case when the main ion of the formula can not match any peak in the spectrum, we will give the formula a constant negative score.

In both formulas, $r(x)$ is a function whose value is always larger than one. It reflects the relationship between the main ions and supporting ions. When the ratio is in the reasonable range, the value of this function is big. On the other hand, when the ratio is too large or too small, the value of this function is small.

In total, we get the scoring function as follows:

$$score(x,v) = scon(x) + scoc(W - x) \tag{2.9}$$

Where x is a mass of prefix while v is a mass of suffix. $score(x,v)$ only considers the peaks which are matched by x's corresponding ions but can not be explained by any corresponding ions of v. This insures that the peaks in the spectrum will only be used once.

2.4 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 zero there are 9 elements in $\sum$.

To compare with the database searching method [36], we first construct an artificial data set which is similar to the experimental data stated in[36] 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.

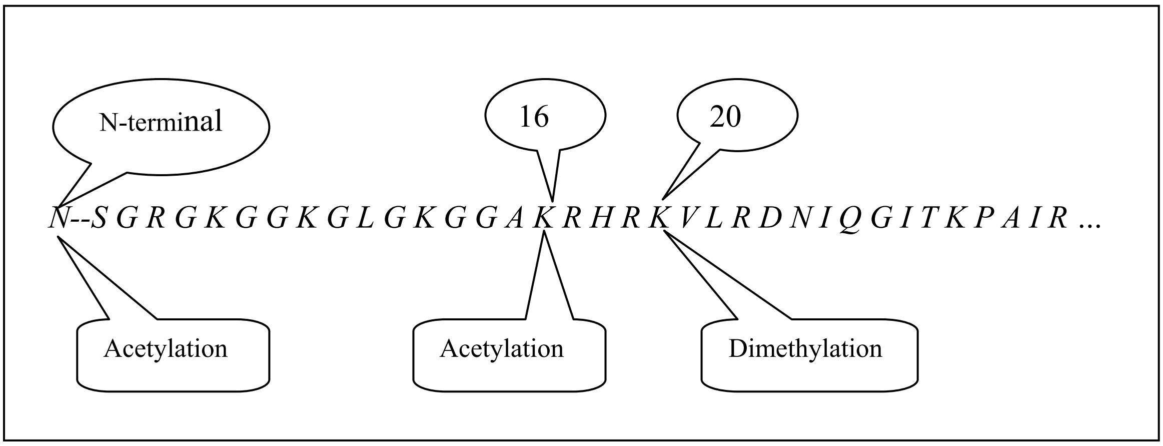

N-terminal
16
20
N--S G R G K G G K G L G K G G A K R H R K V L R D N I Q G I T K P A I R ...
Acetylation
Acetylation
Dimethylation

Figure 2.3 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 [36], 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.

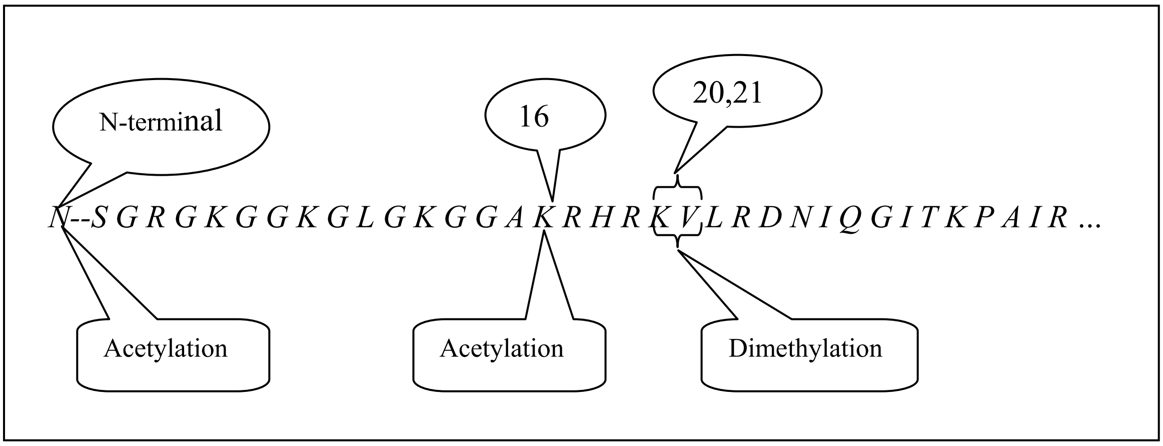

N-terminal
16
20,21
N--S G R G K G G K G L G K G G A K R H R K V L R D N I Q G I T K P A I R ...
Acetylation
Acetylation
Dimethylation

Figure 2.4 The PTMs Result Got by Our Algorithm

To test the robustness of the algorithm, we gradually delete some matched peaks from the original spectrum. Table 2.1 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 2.1. 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 2.1 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 2.4, if we have the prior knowledge that *V* could not be modified by methylation, we can conclude that the position 19 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.

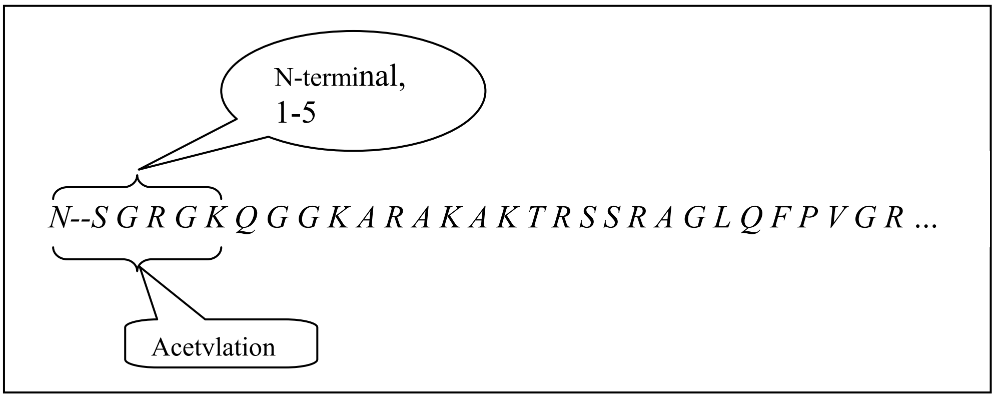

N-terminal,
1-5
N--S G R G K Q G G K A R A K A K T R S S R A G L Q F P V G R ...
Acetvlation

Figure 2.5 The Modification Allocation of H2A

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.

However, there are several limitations of our method. First, the algorithm 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 algorithm 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.

Chapter 3

A DECISION - TREE PROBABILITY MODEL FOR DE NOVO PEPTIDE SEQUENCING

3.1 Related work

There are two classes of algorithms nowadays for solving the peptide sequencing problem. The first class is database searching method[1, 10, 35]. This approach is very popular and it can successfully identify some already-known proteins. The core of this approach contains three modules: (a) Interpret the tandem mass spectrum; (b) By using the interpreted spectrum and a protein database, some candidate peptides are identified; (c) Rank the candidate-peptides by a score function and output those high ranked peptides. Widely used algorithms such as Sequest[10] and Mascot[35] apply this approach. Although database search is a powerful tool for peptide identification, there are still some problems. A protein database is indispensable in this method and the peptides found by this method must already exist in the database. However, due to alternatively spliced genes, many peptides may not exist in the database [25]. Besides, because of the dynamic nature of peptides, database searching method may fail due to mutation and modification in the peptides.

Because of the disadvantages of database searching method, a lot of researches have been focused on another class of algorithms, de novo peptide sequencing methods[2, 4, 6, 13, 26-29, 32, 47]. De novo peptide sequencing problem is to derive the peptide sequence directly from the mass spectrum. Most popular algorithms use a *spectrum* *graph*[4, 6, 26, 27] to solve the problem. A spectrum graph is formed by transforming the peaks in the tandem mass spectrum into an acyclic graph. Each peak in the spectrum is transformed to several vertices in the graph by assuming the peak is of different types of ions. Each edge in the graph links two vertices which are different by the mass of an amino acid. The de novo peptide sequencing problem is equivalent to finding the longest path in the spectrum graph. However, since every peak can be interpreted into several vertices, when a peak has a high intensity, there is a tendency that the longest path will include more than one vertex corresponding to the same peak. Although forbidding the simultaneous occurrences of pairs of nodes corresponding to the same peak can avoid the problem, when there are really different nodes corresponding to the same peak, this method will fail. Thus another algorithm Peaks[28, 29] is proposed which performs de novo peptide sequencing without using the spectrum graph. Peaks uses a dynamic programming to pick out the highest scored peptide from all possible peptides whose masses are equal or close to the experimental mass value. Basically, the algorithm gradually constructs optimal pairs of prefixes and suffixes in a carefully designated way, until the prefix and the suffix becomes long enough to form the optimal solution.

For both de novo sequencing algorithms and database searching algorithms, the scoring function is critical to determine the accuracy of the methods. In general, there are two popular scoring functions. The first one is to correlate the experimental spectrum with the theoretical spectrum produced by candidate peptide[10]. Algorithms such as Sequest use this kind of scoring function. Another kind of scoring function uses probability value to evaluate the peaks in the experimental spectrum[1, 6, 9, 15, 18]. Banfa and Edwards proposed a probabilistic model for database searching method which considers the factors such as fragment ion probabilities and instrument measurement errors. Danick *et al.* designed a probability based scoring function for the de novo peptide sequencing algorithm, Sherenga[6]. However this scoring function does not fully exploit the factors that influence the intensities of the peaks in the spectrum. Since intensities are reproducible, some researches have focused on studying the chemical and physical properties of the peptides that will influence the intensity. Elias *et al.* used a probabilistic decision tree to model the probability of observing certain intensity for a given peak with certain particular chemical and physical properties. Then, they applied their intensity-based scoring function in database searching. Frank and Pevzner[15] proposed a scoring function using a probabilistic network which reflects the chemical and physical rules in peptide fragmentation. Then, they applied the score function in de novo peptide sequencing.

In this thesis we proposes another way to use the probabilistic decision tree to model the peak's intensity based on the chemical and physical properties of the fragment ions. Using our probabilistic decision tree model, we give an algorithm DTSeq that accurately solves the de novo peptide sequencing problem. Experimental results show that DTSeq has high accuracy. The rest of this chapter is organized as follows: Section 2 introduces some terminologies; Section 3 gives the scoring function and algorithm and Section 4 presents the experiment results.

3.2 Preliminary

In this section, we will describe some basic terminologies and concepts.

3.2.1 Amino acid property

Amino acids are small biomolecules which are the principal building blocks of proteins. There are 20 common amino acids and we use A to denote the alphabet of the 20 amino acids. In this chapter, each amino acid residue is characterized by four attributes: mass, gas-phase basicity, hydrophobicity[III] and helicity[IV]. (Table 3.1) The monoisotopic mass of each amino acid $a \in A$ is denoted as $wt(a)$. Note that 57.02 Dalton ≤ $wt(a)$ ≤ 186.08 Dalton. The gas-phase basicity[19] of an amino acid a is denoted as *gb(a)*. It measures the tendency of a molecule to accept a proton in the reaction. Thus it is highly related to the proton affinity, which partially determines the site of proton attachment. A lot of evidence shows that the site of proton attachment influences the fragmentation reactions. Note that $202.7 \leq gb(a) \leq 237.0$. Hydrophobicityand and helicity[7] of an amino acid a are denoted as *hyd(a)* and *hlx(a)*, respectively. Hydrophobicity is an important factor to determine the protein stability while helicity is found to influence the folding of the nascent polypeptide chain. We have $-5.00 \leq hyd(a) \leq 5.00$ and $0.57 \leq hlx(a) \leq 1.29$.

[III] Scaled from high-pressure liquid chromatography (HPLC) retention times
[IV] Scaled from circular dichroism measurements of peptides in n-butanol

Amino Acid	Mass	Basicity	Hydrophobicity	Helicity
A	71.0	206.4	0.16	1.24
C	103.0	206.2	2.50	0.79
D	115.0	208.6	-2.49	0.89
E	129.0	215.6	-1.50	0.85
F	147.1	212.1	5.00	1.26
G	57.0	202.7	-3.31	1.15
H	137.1	223.7	-4.63	0.97
I	113.1	210.8	4.41	1.29
K	128.1	221.8	-5.00	0.88
L	113.1	209.6	4.76	1.28
M	131.0	213.3	3.23	1.22
N	114.0	212.8	-3.79	0.94
P	97.1	214.4	-4.92	0.57
Q	128.1	214.2	-2.76	0.96
R	156.1	237.0	-2.77	0.95
S	87.0	207.6	-2.85	1.00
T	101.0	211.7	-1.08	1.09
V	99.1	208.7	3.02	1.27
W	186.1	216.1	4.88	1.07
Y	163.1	213.1	2.00	1.11

Table 3.1 Amino Acid Properties

3.2.2 Fragment Ions

Consider a peptide sequence constructed by *n* amino acids $P = a_1a_2a_3\ldots a_n$. We denote $wt(P) = \sum_{1\le i\le n} wt(a_i)$. Because of the extra H_2O, the actual mass of *P* is $wt(P)+18$ and we denote it as *Ma(P)*. In the mass spectrometry experiment, the original whole peptide is charged by some H^+ ions which is called the precursor ion. In this chapter we only consider the doubly charged peptide. Thus the peptide precursor mass is *Ma(P)*+2.

As mentioned in Chapter 1, peptides are then fragmented into pieces during the Collision Induced Dissociation (CID) process. For instance, suppose a peptide $P = a_1a_2\cdots a_ia_{i+1}\cdots a_n$ is fragmented into two parts by the cleavage between a_i and

a_{i+1}. Then, a_i is called the N-terminal amino acid to the cleavage site while a_{i+1} is called the C-terminal amino acid to the cleavage site. The fragments are charged and only charged pieces can be detected by mass spectrometer. The charged fragments are called fragment ions. After a single cleavage along the peptide backbone, there are six possible types of fragment ions (Figure 1.3). Among them, the b-ion and y-ion are the most frequently appeared ions in the spectrum. Besides, by some neutral losses of chemical group such as H_2O and NH_3, $b-H_2O$, $b-NH_3$, $y-H_2O$ and $y-NH_3$ are produced. These types of ions are much less observed than the b-ion and y-ion.

Based on the discussion above, the ions can be classified into two groups: the N-terminal group and the C-terminal group. The N-terminal group contains b-ion, a-ion, c-ion, $b-H_2O$ and $b-NH_3$ while the C-terminal group contains y-ion, x-ion, z-ion, $y-H_2O$ and $y-NH_3$. Consider an amino acid sequence $H = a_1a_2 \cdots a_k$, we define $B(H) = wt(H)+1$ while $Y(H) = wt(H)+19$, be the masses of H when H are b-ion and a y-ion, respectively.

3.2.3 Spectrum of a peptide

The LC/MS/MS spectrum of a peptide consists of many peaks. Each peak is generated by a large amount of copies of some fragment ion of the peptide. As we have mentioned, the mass position of the peak in the spectrum represents the mass over charge ratio of the corresponding fragment ion, while the height of the peak indicates the intensity of the fragment ion.

However, mass spectrum usually contains many other peaks which are not produced by any fragment ions of the peptide. They could be the results of chemical contaminants and machine error. All these peaks are treated as noise. Besides, in our experiment, we only consider b-ion and y-ion in the spectrum since they are the most abundant ions. Because the limitation of our model, peaks corresponding to other types of ions are also treated as noise. The appearance of noise peaks adds difficulty to the de novo sequencing problem, since they may be considered as real peaks produced by false fragment ions.

Given a peptide *P*, after fragmentation, let $S = \{(x_i, y_i) \mid 1 \le i \le num\}$ be the corresponding spectrum which shows the fragmentations results of *P*. *num* is the number of peaks in the spectrum, x_i is the position (i.e. the mass over charge ratio) of the *i*th peak in the spectrum and y_i is the intensity value of the *i*th peak. Since the experimental data is not accurate, the positions of the peaks may be shifted by a little bit. We denote $\delta > 0$ be the measurement error of the experiments, which is assumed to be 0.5 in this chapter. To simply the discussion, when we say that there exists a peak in S at position *w*, we refers it to be the peak (x_{ii}, y_{ii}) where $ii = \max \arg_i \{y_i \mid (x_i, y_i) \in S, | w - x_i | \le \delta\}$.

3.2.4 Factors which affect the abundance of a peak

The intensity of a fragment ion depends on many factors during the low-energy collision induced dissociation process. To develop a robust intensity-based scoring method, it is important to understand the factors influencing the gas-phase fragmentation of peptides.

It is known that, in general, the peak of a y-ion has higher intensity than that of a b-ion [20, 44]. Although y ions are only slightly more often appearing in the spectrum than b ions, their peaks are usually much more intense. Moreover, there are other factors. First, an abundant y-ion usually has an abundant complementary b-ion. Second, fragmentation near the N termini or C termini of the peptide causes low intensity peaks while fragmentation in the middle of the peptide causes much higher intensity peaks[44]. Tabb et al. have shown that the peaks of y-ion and b-ion are most intense around ~60% and ~45% of the precursor mass, respectively. Third, the intensity of a fragment ion is also influenced by its mass since the mass spectrum has certain observed scan range. Fourth, it is widely known that the intensity of a fragment ion depends on the type of the amino acids. For example, the fragmentation at the N-terminal side of proline produces low intensity peaks while the fragmentation at the C-terminal side of proline produces high intensity peaks[3]. Besides, based on the 'mobile protone' hypothesis[V][41, 45], some other information such as peptide length, precursor charge state and the presence of basic residues also influence the fragment intensity[21, 41].

3.2.5 Normalization and discretization

The intensity of every peak in a spectrum may change due to different experimental environment. It is necessary to normalize the intensities of the peaks before we use it. In our case, we transform the intensities of the peaks into 4 discrete levels as follows. First, for every spectrum, we transformed the raw intensity (I_r) of each peak into normalized intensity (I_n) by the following formula.

[V] 'mobile protone' hypothesis: the cleavage in a peptide is generally thought to be initiated by migration of the charge from the initial site of protonation to an amide carbonyl oxygen along the peptide backbone.

$$I_n = \frac{I_r}{I_i} \tag{3.1}$$

where I_i is the average intensity of the one third of peaks in the spectrum which have the lowest intensities. Then, we discretize the normalized intensities of the peaks into 4 levels. The peaks with $I_n < 1$ are included in *Level 0*. These peaks are treated as unobserved. For the remaining peaks, each of them is assigned to *Level 1, 2, and 3*, respectively, if $1 \le I_n < 6$, $6 \le I_n < 16$ and $I_n \ge 16$.

Based on such normalization and discretization, for peaks corresponding to b-ion fragment or y-ion fragments, 15.5% of them are assigned to *Level 1*, 25.1% of them are assigned to *Level 2* and 59.4% of them are assigned to *Level 3*.

3.3 Score Function

This section proposes an intensity-based scoring function that can be used to improve the accuracy of de novo peptide sequencing. The new scoring function is learnt from a training dataset of spectra and the corresponding peptide sequences. It is based on a probabilistic decision tree model which estimates the likelihood of observing certain intensity for a peak corresponding to a certain fragment ion. Below, we will first present the scoring function which is based on local peptide and fragment attributes. Then, given a training dataset of spectra and the corresponding peptide sequences, we describe how learn the probabilistic decision tree model from the training dataset. Finally, we will present the de novo peptide-sequencing algorithm based on such scoring function. Our algorithm is dynamic programming in nature and is similar to the one used in Peaks.

3.3.1 The likely scoring function

Consider some spectrum *S* of peptide *P*. Let *F* be an fragment ion of the peptide *P* whose mass equals *w*. For *I*=0,1,2,3, the likely scoring function evaluates the likelihood of observing an intensity *I* for the peak at mass position *w*. Depending on whether *F* is b-ion or y-ion, the likely scoring functions $sco_B()$ or $sco_Y()$ for b-ion or y-ion, respectively, are defined as follows.

$$sco_B(I \mid Info(F), S) = \ln \frac{p_{real}^{B}(I \mid Info(F), S)}{p_{random}(I \mid Info(F), S)} \tag{3.2}$$

$$sco_Y(I \mid Info(F), S) = \ln \frac{p_{real}^{Y}(I \mid Info(F), S)}{p_{random}(I \mid Info(F), S)} \tag{3.3}$$

where *Info*(*F*) is the local information related to the fragment *F* (defined below), $p_{real}^{B}(I \mid Info(F), S)$ and $p_{real}^{Y}(I \mid Info(F), S)$ are the probabilities of observing an intensity *I* at mass position *w* given that *F* are b-ion and y-ion fragments respectively; $p_{random}(I \mid Info(F), S)$ is the probability of observing an intensity *I* at mass position *w* by random. We estimate $p_{real}^{B}(I \mid Info(F), S)$ and $p_{real}^{Y}(I \mid Info(F), S)$ using a decision tree and the detail will be discussed below. $p_{random}(I \mid Info(F), S)$ estimates the probability that the peak is in fact some noise or is the peak of other fragment ion. It is computed based on the density estimation model in [15].

A positive score for $sco_B()$ or $sco_Y()$ means the intensity is more likely to be produced by the candidate fragment ion while a negative score means that the peak is randomly matched and is unlikely to be produced by the fragment ion. Since the scores of the fragment ions of the peptide are independent, for a spectrum *S* and a peptide $P=a_1a_2...a_n$, the score $score(P,S)$ for the peptide P can be computed by summing up the individual scores for all fragment ions of P as follows: Let I_i^B and I_i^Y be the observed intensities at mass positions $B(a_1...a_i)$ and $Y(a_{i+1}...a_n)$, respectively.

$$score(P,S) = \sum_{1 \le i \le n-1} score(a_1 \ldots a_i, a_{i+1} \ldots a_n, S); \text{ and} \tag{3.4}$$

$$sco(a_1 \ldots a_i, a_{i+1} \ldots a_n, S) = sco_B(I_i^B \mid Info(a_1 \ldots a_i), S) + sco_Y(I_i^Y \mid Info(a_{i+1} \ldots a_n), S) \tag{3.5}$$

Note that the bigger the score, the more likely that the spectrum S represents the peptide P. To complete the discussion, the remaining subsections will discuss how to compute the probabilities $p_{real}^B(I \mid Info(F), S)$, $p_{real}^Y(I \mid Info(F), S)$, and $p_{random}(I \mid Info(F), S)$.

3.3.2 Computing $p_{real}^B(I \mid Info(F), S)\, and\, p_{real}^Y(I \mid Info(F), S)$ using decision tree

Consider a fragment ion F of a peptide P. Suppose its mass is w. This section proposes to use probabilistic decision tree to learn how the properties of F affecting its intensity level.

Based on previous discussion, the intensity of a fragment ion F may be affected by many local attributes of F, including (1) fragment ion mass, (2) intensity of the complementary fragment ion, and (3) the gas-phase basicity, hydrophobicity and helicity of the terminal amino acid to the cleavage site. Table 3.2 summarizes the set of attributes for describing F when F is a y-ion. When F is a b-ion, it can be described by the same set of attributes in Table 3.2 except that the attribute “BionInt” is replaced by “YionInt” which represent the intensity of y-ion.

Attributes abbreviation	Attribute description
Pos[VI]	The position of the cleavage site along the peptide
BionInt	The intensity of b-ion
Pmas	Peptide precursor mass
Mc	Fragment mass/charge
Masd	Fragment mass minus peptide precursor mass
Mcd	Fragment mass/charge minus precursor mass/charge
Mdisn[VII]	Mass distance from cleavage site to N-terminus
Mdisc[VIII]	Mass distance from cleavage site to C-terminus
Gbn	Gas phase basicity of the N-terminal amino acid to fragmentation site
Gbc	Gas phase basicity of the C-terminal amino acid to fragmentation site
Hlxn	Helicity of the N-terminal amino acid to fragmentation site
Hlxc	Helicity of the C-terminal amino acid to fragmentation site
Hydn	Hydrophobicity of the N-terminal amino acid to fragmentation site
Hydc	Hydrophobicity of the C-terminal amino acid to fragmentation site
Resn	The N-terminal amino acid to fragmentation site
Resc	The C-terminal amino acid to fragmentation site

Table 3.2 Training Attributes For Decision Tree

Given the set of training peptides and their normalized spectra, then the probabilistic decision tree for y-ion can be generated as follows. First, from the training dataset, we generated all the y-ion fragment ions; for each y-ion fragment, a vector of its attributes and its intensity is generated to represent it. Then, the decision tree is trained using J4.8[27] based on the vectors. Furthermore, every leaf node of the decision trees is associated with a probability distribution of the 4 intensity levels.

Consider a peptide $P = a_1 a_2 \cdots a_i a_{i+1} \cdots a_n$, the cleavage site between a_i and a_{i+1} can creates the corresponding y-ion and b-ion.

[VI] $Pos = \frac{wt(a_1 a_2 \cdots a_i)}{wt(P)} = \frac{wt(P) - wt(a_{i+1} a_{i+2} \cdots a_n)}{wt(P)}$

[VII] $Mdisn = wt(a_1 a_2 \cdots a_i) = wt(P) - wt(a_{i+1} a_{i+2} \cdots a_n)$

[VIII] $Mdisc = wt(a_{i+1} a_{i+2} \cdots a_n) = wt(P) - wt(a_1 a_2 \cdots a_i)$

The probability of having intensity level *I* is estimated to be the proportion of the training fragment ions corresponding to this leaf node having intensity level equals *I*. By using a similar approach, we can get the probabilistic decision tree for b-ion.

Figures 3.1 and 3.2 show the learned probabilistic decision trees for b-ion and y-ion fragments, respectively. Arrows pointing to the left indicate the fragments which satisfy the condition stated by the source node while arrows pointing to the right indicate fragments that do not. The histogram in every leaf node shows the intensity distribution of the peaks that assigned to it. The attribute in the root node shows the most important factor that will influence the fragment intensity. And the nodes closer to the root are more important. In the figures, the root node indicates that, in general, the fragmentation near the N-terminal ($\mathrm{Pos} \leq 0.14$) produces low intensity peaks. This rule agrees with the known knowledge we have mentioned before. The decision tree also discovers other rules. For instance, the attributes of the N-terminal amino acid to the cleavage site have more influence on the fragment intensity and they appear in the decision tree while the attributes of the C-terminal amino acid to the cleavage site do not appear. Those identified rules prove the credibility of our method. In addition, it also increases our confident of the validity of the unknown rules discovered by the decision tree.

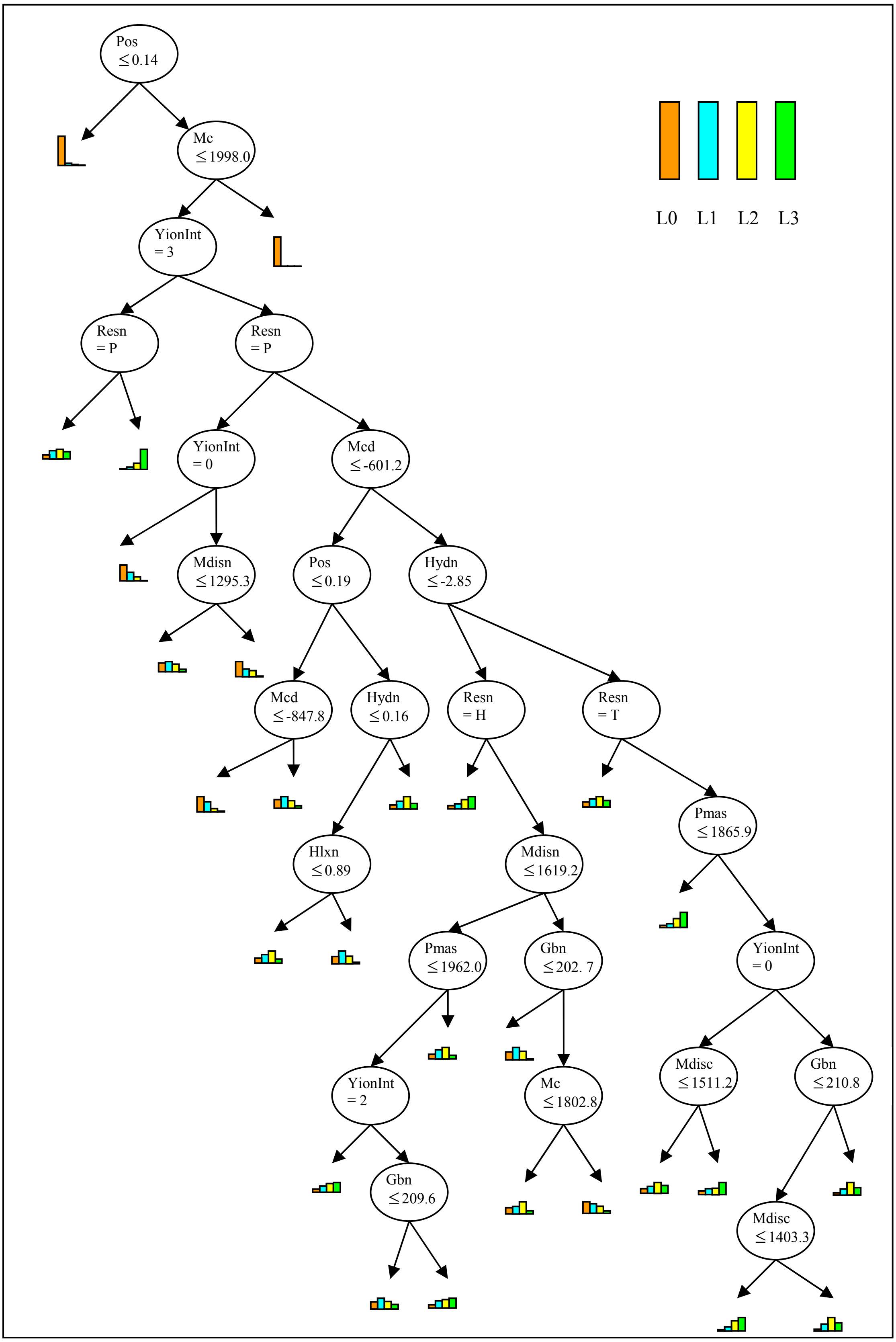


Figure 3.1 Decision Tree for b-ion

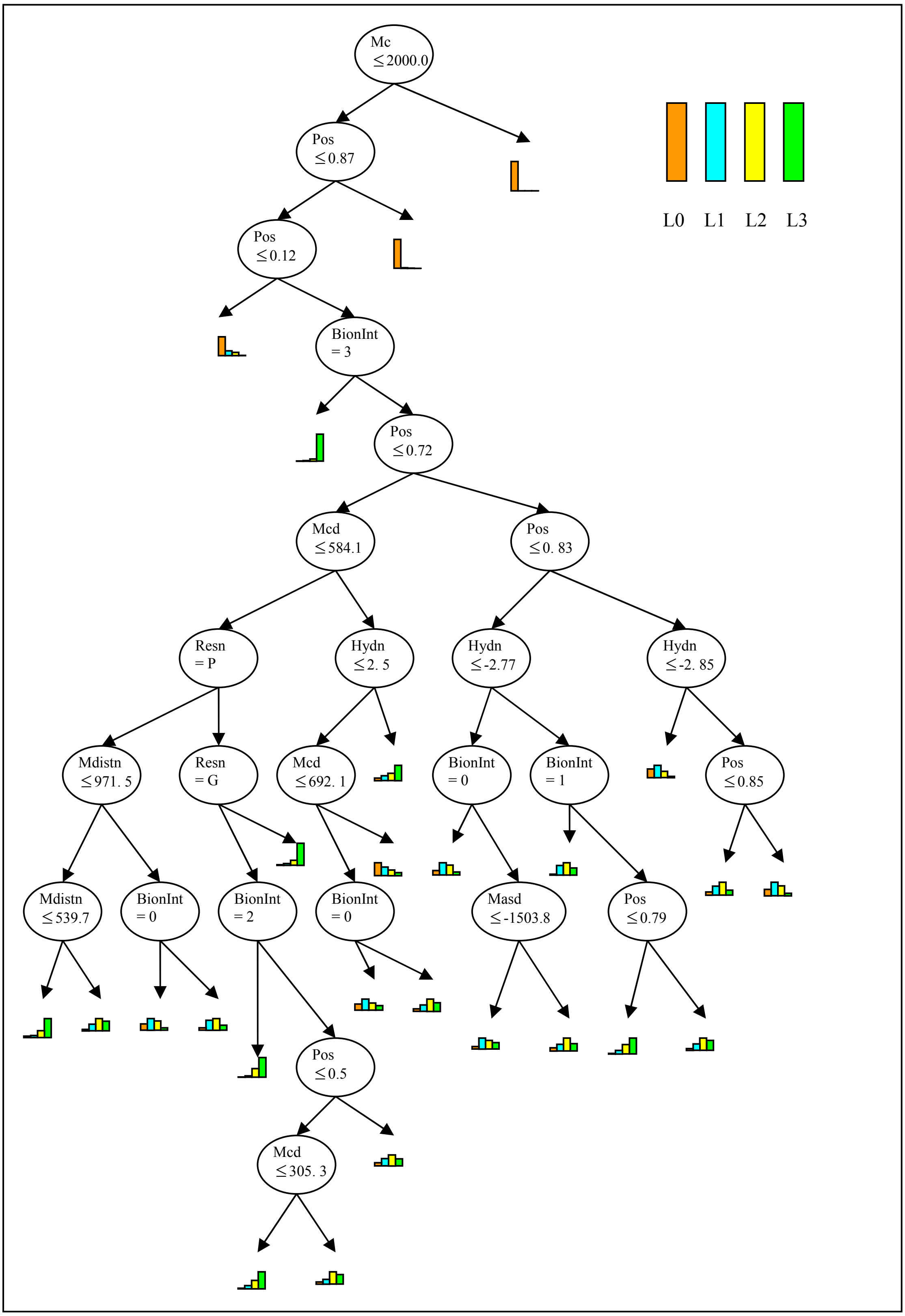


Figure 3.2 Decision Tree for y-ion

As a matter of fact, though the decision trees for the b-ion and the y-ion (see Figure 3.1 and 3.2) depend on many attributes, the values of those attributes can be computed based on the fragment mass of *F* (*w*=*B*(*F*) or *Y(F)*) and the N-terminal amino acid to the cleavage site (a_i) only. In other word, *Info*(*F*) = {*w*, a_i} is sufficient to compute $p^B_{real}(I \mid Info(F), S)$ and $p^Y_{real}(I \mid Info(F), S)$.

Given the decision tree in Figure 3.1 and a b-ion fragment *F*, $p^B_{real}(I \mid Info(F), S)$ is computed as follows. First, we search the decision tree for b-ion and find a leaf node corresponding to *F*. Such leaf node is associated with a probability distribution. Then, $p^B_{real}(I \mid Info(F), S)$ equals the probability for intensity *I* of such distribution. For a y-ion fragment *F*, by applying the same procedure on the decision tree for y-ion (Figure 3.2), $p^Y_{real}(I \mid Info(F), S)$ can be computed similarly.

3.3.3 Computing Random Probability $p_{random}(I \mid Info(F), S)$

Consider a fragment ion *F* of mass *w* and a spectrum *S*. For *I*=0,1,2,3, this section would like to build a random probability model to compute $p_{random}(I \mid Info(F), S)$, that is, the random chance that the peak in *S* at position *w* has an intensity *I* (that is, *I* is the highest intensity among the intensities of all the peaks in S within the range $[w-\delta..w+\delta]$).

As the intensities of the peaks in the middle of the spectrum are much higher then the intensities in the two ends of the spectrum, we cannot assume the intensity of a noise peak follows a uniform distribution. Instead, we use the local density estimation model proposed by Frank and Pevzner. For completeness, we present their solution in this section. Consider a window of size *u* around *w* (the range from *w*-*u*/2 to *w*+*u*/2).

For i=1,2,3, Let d_i be the number of peaks in S whose intensity level is i within the size-u window. For a randomly chosen peak within the size-u window, the probability that the peak falls outside the range $[w-\delta\,..\,w+\delta]$ can be estimated as $\gamma = 1-\frac{2\delta}{u}$. Thus, the probability that there is no peak within the range $[w-\delta\,..\,w+\delta]$ is as follows:

$$p_{random}(I = 0 \mid Info(F), S) = \gamma^{\sum_{i=1}^{3} d_i} \tag{3.6}$$

The probability that the highest intensity level among all peaks within the range $[w-\delta\,..\,w+\delta]$ is as follows:

$$p_{random}(I = l \mid Info(F), S) = (1-\gamma^{d_l}) \cdot \gamma^{\sum_{i=l+1}^{3} d_i} \tag{3.7}$$

The first factor $(1-\gamma^{d_l})$ in the above equation calculates the probability there is at least one level-l peak within the range $[w-\delta\,..\,w+\delta]$ while the other factor calculates the probability that all peaks that are in level higher than l fall outside the range $[w-\delta\,..\,w+\delta]$. Thus the product of them is the probability that the highest peak is of level l.

The above two equations imply that, in a dense region in S (region with many peaks), the probability $p_{random}(I > 0 \mid Info(F), S)$ is higher. This is reasonable since, in a dense region, it is more likely that the peak is matched by chance. Finally, note that $Info(F)=\{w\}$ is sufficient for the computation of $p_{random}(I \mid Info(F), S)$.

3.3.4 Algorithm

Given a spectrum *S* and an observed peptide mass *M*, this section describes a dynamic programming algorithm to compute a peptide *P*, where $|M\text{-}18\text{-}wt(P)|\leq\delta$, which maximizes *score*(*P*, *S*).

Our dynamic programming is based on DT[b, y, a], which is defined as

$$\max\left\{ \sum_{1\leq k\leq i \text{ or } j-1\leq k\leq n-1} score(x_1\ldots x_k, x_{k+1}\ldots x_n, S) \middle| b = wt(x_1\ldots x_i), y = wt(x_j\ldots x_n), a = x_{j-1}, M-18 = wt(x_1\ldots x_n) \right\} \tag{3.8}$$

Below lemma shows the usefulness of the table DT.

Lemma: Suppose $P=a_1a_2\ldots a_n$ maximizes *score*(*P,S*). Then, for any $1\leq i\leq n$, *score*(*P, S*) = DT[$wt(a_1a_2\ldots a_{i-1})$, $wt(a_{i+1}a_{i+2}\ldots a_n)$, a_i]. In particular, there exists *i* such that $|wt\ (a_1a_2\ldots a_{i-1})\ \text{-}\ wt(a_{i+1}a_{i+2}\ldots a_n)\text{-}18| \leq \max_{a\in\Sigma_a} wt(a)$ (*which is* 186.1) .

Proof: Note that *score*(*P,S*)= $\sum_{1\leq k\leq n} sco(a_1\ldots a_k, a_{k+1}\ldots a_n, S)$. By definition of DT, *score*(*P, S*) = DT[$wt(a_1a_2\ldots a_{i-1})$, $wt(a_{i+1}a_{i+2}\ldots a_n)$, a_i] for any $1\leq i\leq n$. Since the weight of any amino acid is smaller than 186.1, there should exist *i* such that $|wt(a_1a_2\ldots a_{i-1})$ - $wt(a_{i+1}a_{i+2}\ldots a_n)| \leq \max_{a\in\Sigma_a} wt(a)$ (*which is* 186.1) . □

Consider a b-ion fragment *F* of mass v. Suppose $\overline{F}$ is the complementary y-ion fragment of *F* and the rightmost amino acid of *F* is a. Note that *Info*(*F*)={*v,a*} and *Info*($\overline{F}$)={*M-18-v,a*}. Let I_v^B and I_v^Y be the intensities of the peaks in the spectrum S at mass *B*(*v*) and *Y*(*M-18-v*), respectively. Consider another y-ion fragment of mass v'. We define

$$score(v, v', a) = sco_B(I_v^B \mid Info(F), S) + sco_Y(I_v^Y \mid Info(\overline{F}), S) \qquad (3.9)$$

if I_v^B and I_v^Y are not peaks at mass *Y*(*v'*) and B(*M* – *v'* – 18), respectively. Otherwise, we set $score(v, v', a) = 0$.

In the above formula, $score(v, v', a)$ equals $score(F, \overline{F})$ if *F* and $\overline{F}$ but cannot be explained by *Y*(*v'*) and B(*M* – *v'* – 18). This insures that the peaks in the spectrum will only be used once.

Below recursive formula allows us to compute all entries DT[b,y,a], where $|b - y| \leq 186.1$ and $b+y+wt(a) \leq M-18+\delta$ using dynamic programming.

Lemma: For $|b - y| \leq 186.1$ and $a \in A$,

$$DT[b, y, a] = \max_{a' \in \Sigma_a} \begin{cases} DT[b, y - wt(a'), a'] + score(M - 18 - y, M - 18 - b, a) \; if \; y - wt(a') \leq b & (1) \\ DT[b - wt(a'), y, a] + score(b, y, a') \; if \; b - wt(a') < y & (2) \end{cases}$$

Basis: *DT*[*0,0, a*] = 0;

Proof: Without loss of generality, we just prove case (1). DT[*b*, *y-wt(a')*, *a'*] corresponds to the score of a prefix-suffix pair (F, F') such that $F = a_1 a_2 \cdots a_i$, $b = wt(F)$ and $F' = a_{j+1} a_{j+2} \cdots a_n$, $y - wt(a') = wt(F')$. Since $|b - (y - wt(a'))| \leq 186.1$ and $y - wt(a') \leq b$, we have $|b - y| \leq 186.1$. Suppose $F'' = a'F'$, thus $y = wt(F'')$. From formula (1) and the definition of DT, the score for a new prefix-suffix pair (F, F'') can be got, which is a candidate value for *DT*[*b*, *y*, *a*]. Because DT[*b*, *y-wt(a')*, *a'*] and DT[*b-wt(a')*, *y*, *a*] are both optimal values, DT[b, y, a] thus must also be the optimal value for certain prefix-suffix pair. □

Note that $score(P,S) = \max_{|b+y+wt(a)+18-M|\le\delta} DT[b,y,a]$. Hence, the target peptide *P* can be found as follows: First, we evaluate all entries DT[*b,y,a*], where $|b-y| \le 186.1$ and |*b*+*y*+*wt*(*a*)+18-*M*|≤*δ*, based on the above recursive formula; Then, among all entries DT[*b,y,a*] such that |*b*+*y*+*wt*(*a*)+18-*M*|≤*δ*, we find the entry DT[*b*, *y*, *a*] with maximum value; Finally, by backtracking, we can recover the peptide *P*. The *Pseudo Code* is shown in Figure 3.3.

```
Input: Observed peptide mass $M$;
       A peak list of the spectrum $S$;
       Error bound $\delta$ of the spectrum;
       A calibration $\Delta$;
       Window size $u$ for estimating the random hit probability.
Output: A peptide such that its score is maximized and $|wt(P)+18-M| \le \delta$
  Initialize all $DT[i,j,a] = -\infty$; Let $DT[0,0,a] = 0$ for all $a \in A$
  for $i$ from 1 to $M/2 + \max_{a\in\Sigma_a} wt(a)$ step $\Delta$ do
  for $j$ from $i - \max_{a\in\Sigma_a} wt(a)$ to $\min(i + \max_{a\in\Sigma_a} wt(a), M-18-i)$ step $\Delta$ do
    for $a \in A$ do
      if $i < j$
        for $a' \in A$ such that $i + j + wt(a') < M - 18$ do
    $DT[i+wt(a'),j,a] = \max\begin{cases} DT[i+wt(a'),j,a] \\ DT[i,j,a] + score(i+wt(a'),j,a') \end{cases}$
      else
        if $i + j + wt(a) < M - 18$
          for $a' \in A$
    $DT[i,j+wt(a),a'] = \max\begin{cases} DT[i,j+wt(a),a'] \\ DT[i,j,a] + score(M-j-wt(a)-18, M-i-18, a) \end{cases}$
  Find the best $DT[i,j,a]$ for all $i, j, a$ satisfying $|i+j+wt(a)+18-M| \le \delta$
  Use backtracking to construct the peptide sequence
```

Figure 3.3 De Novo Algorithm

Lemma: In lines 7 and 11, the *score*() function can be computed in $O(\frac{u}{\Delta})$ time

Proof: The *score*() function is got by the sum of $sco_B(I_F \mid Info(F), S)$ and $sco_Y(I_{\overline{F}} \mid Info(\overline{F}), S)$. As we have mentioned before, $sco_B(I_F \mid Info(F), S)$ (or $sco_Y(I_{\overline{F}} \mid Info(\overline{F}), S)$) is composed of two parts. The first part is $p^B_{real}(I_F \mid Info(F), S)$ (or $p^Y_{real}(I_{\overline{F}} \mid Info(\overline{F}), S)$) and this part can be calculated by going through the decision tree. There are at most $O(\frac{\delta}{\Delta})$ peaks explained by a single fragment mass, thus this part can be computed in $O(\frac{\delta}{\Delta})$ time. The second part is $p_{random}(I_F \mid Info(F), S)$ (or $p_{random}(I_{\overline{F}} \mid Info(\overline{F}), S)$) and it is calculated by using a window *u* to calculate the local density. Because there are at most $O(\frac{u}{\Delta})$ peaks in the window, the time complexity of this part is $O(\frac{u}{\Delta})$. Thus, in total, *score*() can be computed in $O(\frac{u}{\Delta})$ time. □

Lemma: The algorithm can compute the optimal solution of the peptide sequencing problem in $O\left(\frac{M}{\Delta} \times \frac{u}{\Delta} \times \frac{\max_{a \in \mathrm{A}} wt(a)}{\Delta}\right)$ time.

Proof: Since the scoring function can be calculated in $O(\frac{u}{\Delta})$ time, besides, based on line 2 and line 3, we can proof that the algorithm can compute the optimal solution of the de novo peptide sequencing problem in $O\left(\frac{M}{\Delta} \times \frac{u}{\Delta} \times \frac{\max_{a \in \Sigma_a} wt(a)}{\Delta}\right)$ time. □

3.4 Experiment Result

3.4.1 Data Set

In Genome Institute of Singapore (GIS), we analyzed multiple Yeast Hormone protein sources using electrospray ion trap mass spectrometers and generated many MS/MS spectra. Then, a set of 1260 spectra of doubly charged tryptic peptides are selected, which were identified by Sequest with high score ($Xcorr \geq 2.0$ and $\Delta Cn \geq 0.10$). These 1260 spectra are used as training set. Note that doubly charged tryptic peptides are selected since this class of peptides is the most common in mass spectrometry experiments. Besides, the fragment ion considered in the experiments are single charged b-ion and y-ion. This is because these two kinds of ions are most frequently appeared in the spectrum.

For test set, we selected 400 spectra from Open Proteomics Database (OPD)[38]. These spectra were also identified by Sequest with high score (*Xcorr* > 2.5 and multiple hits). The peptides corresponding to these spectra contain 9 to 18 amino acids. The average length of these peptides is 13.7.

3.4.2 Result

Consider a predicted peptide from a particular de novo peptide sequencing algorithm. An amino acid of the predicted peptide is considered as correct if its mass position in the predicted sequence is within 1.5 Daltons from its expected mass position in the correct sequence. Then, the overall *accuracy* of the predicted peptide is defined as follows.

$$accuracy = \frac{number\ of\ correct\ predicted\ amino\ acids}{number\ of\ predicted\ amino\ acids} \qquad (3.10)$$

Besides, as the mass difference between amino acids Isoleucine (I) and Leucine (L) and between Lysine (K) and Glutamine (E) are smaller than 0.05 Daltons, we do not distinguish them in our accuracy measurement.

To test the performance of our algorithm, we compare our algorithm DTSeq with Peaks and another de novo peptide sequencing algorithm PepNovo based on the above measurement. (To the knowledge of the authors, PepNovo and Peaks are the most accurate de novo peptide sequencing algorithms in the literature.) The experiment is as follows. For all three algorithms, we supplied the 400 test spectra to them and computed the predicted peptides sequences. Then, the average accuracies of the three different algorithms are measured. Table 3.3 shows the results.

Algorithm	Average Accuracy	#Predicted Amino Acids
DTSeq	0.689	11.6
PepNovo	0.617	12.8
Peaks	0.550	13.7

Table 3.3 Average Accuracy of Three Algorithms

Since the cleavage sites in the center of the peptide produce much more stronger peaks, while the peaks of terminal parts are weak. Our algorithm can avoid predict the unconfident terminal amino acids to improve the accuracy.

From Table 3.3, the accuracy of our method for the test set is highest among all the three algorithms. Note that both our method and PepNovo use intensity-based scoring function. Thus intensity-based scoring function seems to be able to improve the accuracy.

As de novo sequencing algorithms are often used to predict partial, rather than complete peptides. The capability of the algorithms to reconstruct correct consecutive amino acids subsequences is very important. We also compared the maximal length of correct subsequence of each predicted peptide generated by all three algorithms.

Algorithm	Ratio of maximal correct subsequence length							
	≥ 3	≥ 4	≥ 5	≥ 6	≥ 7	≥ 8	≥ 9	≥ 10
DTSeq	0.94	0.87	0.75	0.63	0.51	0.42	0.32	0.21
PepNovo	0.92	0.83	0.72	0.60	0.51	0.41	0.30	0.21
Peaks	0.86	0.80	0.65	0.53	0.43	0.33	0.25	0.17

Table 3.4 Proportions of Subsequence Length longer than l ($3 \leq l \leq 10$)

Table 3.4 shows that the proportions of the predicted sequences which have a maximal correct subsequence length longer than l ($3 \leq l \leq 10$).The result implies that the predictions made by intensity-based scoring methods are consistently having longer correct subsequences.

Although the experiment shows that our method performed the best, there are still several limitations. First in our decision tree model for DTSeq, only b-ion and y-ion are considered. In the future, we may train more types of ions such as a-ion and some neutral losses ions. Second, our model could only be applied to double charged peptides, we may expand the model to include additional charge states. Third, our method is not fast enough to get the results, we will try to modify the algorithm and make it more efficient. Last but not the least, we plan to do more tests in the future to validate the robustness of our method.

Chapter 4

CONCLUSION

4.1 Conclusions

Protein sequencing is an important problem in the post-genome era. In this thesis, we studied two problems related to protein sequencing.

The first is the protein post translational modifications identification problem. We proposed a dynamic programming algorithm via a “top-down” mass spectrometry to solve this problem. There are many advantages of this new method. First, our method can work without a protein database. Second, there is no prior knowledge of the modification sites in the protein needed. Last but not the least, it can identify the modifications in polynomial time, which is very efficient compared to the widely used database searching method. The experiment shows that our algorithm can get the correct results while much more efficient.

The second is the de novo peptide sequencing problem. A lot of research has been done to solve the peptide sequencing problem. Generally there are two kinds of algorithms. One is the database searching method and another is de novo peptide sequencing. However, little work has been done to utilize the intensities of the peaks in the mass spectrum to improve the accuracy of the peptide sequencing. We proposed a decision tree probability model which fully explores the factors that influence the intensity pattern. The scoring function of this algorithm is based on two models. First we introduced a decision tree probability model which estimates the likelihood of certain observed intensity. Unlike Elias et al.[9] decision tree, our decision tree can model the dependence between y-ion and b-ion. Moreover, to avoid high computational complexity, our decision tree only utilizes the local chemical and physical attributes of the fragment. Besides, a random probability model is used to estimate the likelihood that a certain peak is a noise. In the experiment, we compared DTSeq with two de novo peptide sequencing algorithms: Peaks and PepNovo. The results showed that DTSeq performed better than the other two algorithms. It obtained the longest maximum subsequence of predicted peptide as well as the highest prediction accuracy.

4.2 Future Work

The results obtained for both problems in the thesis demonstrate the advantage of our new algorithms. There are still several possibilities for future work. In our PTMs - identification method, we would like to explore if it is possible to detect PTM sites without knowing the modification types in advance. In our decision tree model for DTSeq, only b-ion and y-ion are considered. In the future, we may train models for more types of ions such as a-ion and some neutral losses ions. Peptide of other charge states will also be included into the models in future.

REFERENCES

1. Bafna, V.; Edwards, N. 2001. SCOPE: a probabilistic model for scoring tandem mass spectra against a peptide database. *Bioinformatics* 17, 13-21.
2. Bartels, C. 1990. Fast algorithm for peptide sequencing by mass spectroscopy. *Biomedical and Environmental Mass Spectrometry* 19, 363-368.
3. Berci, L.A.; Tabb, D.L.; Yates J.R., III.; Wysocki, V.H. 2003. Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra. *Anal. Chem.* 75, 1963-1971.
4. Chen, T.; Kao, M.Y.; Tepel, M.; Rush, J.; Church, G.M. 2000. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comp. Biol.* 8, 325-327.
5. Creasy, D.M.; Cottrell, J.S. 2002. Error-tolerant searching of uninterpreted tandem mass spectrometry data. *Proteomics* 2, 1426-1434.
6. Dancik, V.; Addona, T.A.; Clauser, K.R.; Vath, J.E.; Pevzner, P.A. 1999. De novo peptide sequencing via tandem mass spectrometry. *J. Comp. Biol.* 6, 327-342.
7. Deber, C.M.; Wang, C.; Liu, L.P.; Prior, A.S.; Agrawal, S.; Muskat, B.L.; Cuticchia, A.J. 2001. TM Finder: A prediction program for gransmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci.* 10, 212-219.
8. Dunham, I. *et al.* 1999. The DNA sequence of human chromosome 22. *Nature 402*, 489-495.

9. Elias, J.E.; Gibbons, F.D. King, O.D.; Roth, F.P.; Gygi, S.P. 2004. Intensity-based protein identification by machine learning from a library of tandem mass spectra. *Nat. Biotech.* 22, 214-219.

10. Eng, J.K.; McCormack, A.L.; Yates, J.R., III. 1994. 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.

11. Fenn, J.B.; Mann, M; Meng, C.K.; Wong, S.F.; Whitehouse, C.M. 1989. Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science* 246, 64-71.

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

13. Fernandez-de-Cossio, J.; Gonzalez, J.; Besada, V. 1995. A computer program to aid the sequencing of peptides in collision –activated decomposition experiments. *Comput. Appl. Biosci.* 11, 427-434.

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

15. Frank, A.; Pevzner, P. 2005. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* 66, 946-973.

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

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

18. Havilio, M.: Haddad, Y.; Smilansky, Z. 2003. A intensity-based statistical scorer for tandem mass spectrometry. *Nal. Chem.* 75, 435-444.

19. Harrison, A.G. 1997. The gas-phase basicities and proton affinities of amino acids and peptides. *Mass Spectrom. Rev.* 16, 201-217.

20. Huang, Y.; Wysocki, V.H.; Tabb, D.L.; Yates J.R., III. 2002. The influence of histidine on cleavage C-terminal to acidic residues in doubly protonated tryptic peptides. *Int. J. Mass Spectrom.* 219, 233-244.

21. Kapp, E.A.; Schuz, F.; Reid, G.E.; Eddes, J.S.; Moritz, R.L.; O'Hair, R.A.J.; Speed, T.P.; Simpson, R.J. 2003. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. *Anal. Chem.* 75, 6251-6264.

22. Karas, M.; Hillenkamp, F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* 60, 2299-2301.

23. Krogan, N.J.; Kim, M.; Tong, A.; Golshani, A.; Cagney, G.; Canadien, V.; Richards, D.P.; Beattie, B.K.; Emili, A.; Boone, C.; Shilatifard, A.; Buratowski, S.; Greenblatt, J. 2003. Methylation of Histone H3 by Set2 in

Saccharomyces cerevisiae Is Linked to Transcriptional Elongation by RNA Polymerase II. *Mol. Cell. Biol.,* 23, 4207-4218.

24. Krogh, A. 1998. Guide to Human Genome Computing. *San Diego, CA: Academic*, 261-274.
25. Leipzig, J.; Pevzner, P.; Hever, S. 2004 The Alternative Splicing Gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res*. 32, 3977-3983.
26. Lu, B.; Chen, T. 2003. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *J. Comp. Biol.* 10, 1-12.
27. Lubeck, O.; Sewell, C.; Gu, S.; Chen, X.; Cai, D. M. 2002. New computational approaches for de novo peptide sequencing from MS/MS experiments. *Proc. IEEE* 90, 1868-1874.
28. Ma, B.; Zhang, K.; Lajoie, G.; Doherty-Kirby, A.; Hendrie, C.; Liang, C.; Li, M. 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spec.* 17, 2337-2342.
29. Ma, B.; Zhang, K.; Liang, C. 2003. An effective algorithm for the peptide de novo sequencing from MS/MS spectrum. *CPM'03*. 2676, 266-278.
30. MacCoss, M.J.; Wu, C.C.; Yates, J.R. III. 2002. Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.*, 74, 5593-5599.
31. Mann, M; Hendrickson, R.C.; Pandey, A. 2001. Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437-473.
32. Mann, M.; Wilm, M. Anal. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390-4399.

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

34. Pendey A.; Mann, M. 2000. Proteomics to study genes and genomes. *Nature*, 405, 823-826.

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

36. Pesavento, J.J.; Kim, Y.B.; Taylor, G.K.; Kelleher, N.L. 2004. 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.

37. Pevzner, P.A.; Dancik, V.; Tang, C.L. 2000. Mutation-tolrant protein identification by mass spectrometry. *J. Comp. Biol.*, 7, 777-787.

38. Prince, J.T.; Carlson, M.W.; Wang, R.; Lu, P.; Marcotte, E.M. 2004. The need for a public proteomics repository. *Nat. Biotech.* 22, 471-472.

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

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

41. Schutz, F.; Kapp, E.A.; Simpson R.J.; Speed, T.P. 2003. Deriving statistical models for predicting peptide tandem MS product ion intensities. *Biochemical Society* 1479-1483.

42. Shi, S.D.H.; Hemling, M.E.; Carr, S.A. 2001. Phosphopeptide/Phosphoprotein Mapping by Electron Capture Dissociation Mass Spectrometry. *Anal. Chem*., 73, 19-22.

43. Sze, S.K.; Ge, Y.; Oh, H.; McLafferty, F.W. 2002. 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.

44. Tabb, D.L.; Smith, L.L.; Breci, L.A.; Wysocki, V.H.; Lin D.; Yates J.R., III. 2003. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Anal. Chem.* 75, 1155-1163.

45. Tang, X.J.; Thibault, P.; Boyd, R.K. 1993. Fragmentation Reactions of Multiply-Protonated Peptides and Implications for Sequencing by Tandem Mass Spectrometry with Low-Energy Collision-Induced Dissociation*. Anal. Chem.* 65, 2824-2834.

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

47. Taylor, J.A.; Johnson, R.S. 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* 11, 1067-1075.

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

49. Wilkins, M.R.; Gasteiger, E.; Gooley, A.A.; Herbert, B.R.; Molloy, M.P.; Binz, P.A.; Ou, K.; Sanchez, J.C.; Bairoch, A.; Williams, K.L.; Hochstrasser,

D.F. 1999. High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.*, 289, 645-657.

50. Witten, I.H.; Frank, E. 1999. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.
51. Zubarev, R.A.; Kelleher, N.L.; McLafferty, F.W. 1998. Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *J. Am. Chem. Soc.*, 120, 3265-3266.

APPENDIX

Mass Spectrometry

Mass spectrometry in proteomics is used in three major areas [31]. First it is usually used for protein identification. Second, because mass spectrometry is able to measure the molecular weight of a protein, it is a tool for detection and characterization of post translational modifications (PTMs) in protein. Finally, mass spectrometry is a good technique for characterization and quality control of recombinant proteins and other macromolecules. In this thesis we discuss the first two usages of mass spectrometry.

A mass spectrometer has three components: a source of ions, a mass analyzer and a detector. The sample is first evaporated in a vacuum and exposed to a high voltage, converting the molecules into gas phase ions. The ions are then accelerated through a mass analyzer towards a detector. The mass analyzer separates the ions according to their mass/charge ratio. The detector records the impact of individual ions, producing peaks on a mass spectrum. The mass of a molecule can then be calculated from the mass/charge ratio of its derivative ions.

Matrix-assisted laser desorption ionization (MALDI) and Electrospray (ES) are the two important ionization techniques that should be credited most for the success of mass spectrometry in the life sciences. During the MALDI process[22], a matrix material is first coprecipitated with the analyte molecules. The resulting solid is then irradiated by nanosecond laser pulses. The amount of energy imparted to the biomolecules by the matrices during desorption and ionization are different, which causes the different degree of fragmentation. The precise nature of the ionization process in MALDI is still largely unknown and it is difficult to relate peptide peak height with the quantity of sample present unless an internal standard is used. Besides, the mass range below 500 Daltons is often obscured by matrix-related ions in MALDI.

During the ES process[11], liquid containing the analyte is pumped at low microliter-per-minute flow rates through a hypodermic needle at high voltage to electrostatically disperse, or electrospray, small, micrometer-sized droplets, which rapidly evaporate and which impart their charge onto the analyte molecules. There is no upper mass limit to the analysis by ES mass spectrometry. Because large mass ions are typically multiple charge. Thus they can be into the certain range of mass/charge ratio of the mass spectrometers. ES mass spectrometry can analyze very complex mixtures. But when the molecular weight and the number of molecules increases, the spectra become increasingly difficult to interpret. ES is generally performed in three situation: the infusion mode; the nanoelectrospray format and in combination with high-performance liquid chromatography (HPLC).

There are three different principles [31] applied to achieve mass separation: separation on the basis of time-of-flight (TOF MS); sepration by quadrupole electric fields generated by metal rods (quadrupole MS) or separation by selective ejection of ions from a three-dimensional trapping field (ion trap MS or Fourier transform MS). The same separation principle or different separation principles can be used twice to perform the two step mass spectrometry (MS/MS), which is used for structural analysis such as peptide sequencing. These three separation methods can be coupled to either MALDI or ES. However, regards the special attributes of MALDI and ES, MALDI is usually coupled with TOF MS while ES is usually coupled with quadrupole and ion-trapping MS.