CS5238 Combinatorial methods in bioinformatics 2004/2005 Semester 1

Lecture 4: Peptide Sequencing - September 3, 2004

Lecturer: Wing-Kin Sung *Scribe: Lee Terk Shuen, Lee Ping Alison*

4.1 Introduction

Proteomics is a field that is growing in importance for molecular biology research. It is defined as the systematic analysis of proteins in a cell or a tissue sample, which generally involves steps like separation, identification, and characterization of proteins. In order to identify a particular protein, its amino acid sequence must be determined. Thus, peptide sequencing is an important task in proteomics. Mass-spectrometry (MS) is the dominant technique used for this purpose, because it is fast and sensitive. In MS, a peptide is broken into fragments and their masses, together with each fragment's abundance, are measured to give a peptide spectrum. Given this spectrum, the sequence of a peptide may be derived by either searching within a database of known sequences, or by using computational methods that match mass calculations of candidate sequences with the spectrum.

The determination of protein sequences without the help of a protein database is known as *de novo* sequencing. This method is particularly important for sequencing a novel protein, whose sequence cannot be found by database searching. In this lecture, we shall focus on this method of peptide sequencing.

A mass-spectrometer is able to measure and separate peptide fragments with different mass-to-charge ratio. The sequencing problem now is to derive the amino acid sequence of the peptide given this MS result. If the fragmentation process of the peptide had been ideal (such that each peptide is cleaved between every two consecutive amino acids) and if the mass-spectrometer yields ideal results, then the sequencing problem would be simple. The separations (or peaks) found in the MS spectrum would represent every possible combination of amino acids resulting from the ideal peptide fragmentation, thus it would be straightforward to construct the sequence by reading each peak. However, in practice, the fragmentation process is not ideal, because

the peptide chain can be broken at points other than in-between amino acids. The challenge of *de novo* sequencing is thus to efficiently find the candidate sequence, amongst all possible combinations, that best fits the MS data.

This document is generally divided into two parts. The first part describes the biological aspect of peptide sequencing, which includes the process of obtaining the MS spectrum from proteins to be analyzed. The second part describes the computational aspects of determining the sequence based on the MS spectrum.

4.2 Obtaining Peptide Spectrum using MS

A protein consists of a long chain of amino acids linked by peptide bonds. The chain of amino acids is also known as a polypeptide. The sequence of amino acids in each protein is unique and determines its 3-D shape, which ultimately determines its function. Thus knowing this sequence is pertinent to further investigation of a protein's functions and involvement in a biological process.

The current best way of differentiating different peptide fragments is by measuring and comparing their masses. Mass Spectrometry (MS) is used to measure and separate different fragments according to their mass-to-charge ratio (m/e or m/z). Figure 4.1 shows a simple hypothetical MS spectrum derived from three different fragments. Each fragment gives rise to a peak at the specific mass/charge value. The height generally indicates the relative abundance. The mass is expressed in terms of the atomic mass unit *Dalton* (Da). For example, the mass of a Hydrogen atom is 1Da. In practice, the spectrum is very much noisier due to errors, presence of isotopes, and random fragmentation. Depending on the equipment used, the accuracy of MS can range from $\pm$0.01Da to $\pm$0.5Da.

Table 4.1 gives the average residue masses of the 20 amino acids. An amino acid residue is the amino acid after loosing a water molecule, and this is the exact composition of one unit within a peptide chain. It can be observed that I (*isoleucine*) and L (*leucine*) have the same mass, whilst all the others have different masses. The residue with the smallest mass is G at 57.05Da and the one with the largest mass is W at 186.21Da.

Since all except two amino acid residues have the same mass, it would be highly likely that different peptides (fragments) will have different masses. Thus MS is able to seperate most of the peptides and it would be possible to derive correctly the unique peptide sequences most of the time, based on

Figure 4.1: A simple MS spectrum

the information about mass. However, since I and L have identical masses, they are indistinguishable using usual MS based sequencing. Additional steps may have to be taken to differentiate between the two. We shall follow the common practice, which is to ignore this problem for now and simply treat I and L as the same amino acid.

4.2.1 Protein Identification Process using LC-MS/MS

A common and powerful MS method used in protein identification is the *LC-MS/MS* method. In such an experiment, the target protein sample is first digested by enzymes into many different peptides. In this case, a peptide refers to one fragment resulting from the digestion. Each peptide is relatively short compared to the original protein. The mixture of peptides after digestion is then separated using HPLC and MS. Next, each separated peptide is individually selected and fragmented. Again, by the mass spectrometer, the tandem mass (MS/MS) spectrum of the selected peptide is obtained. These steps are the biological steps for generating a peptide spectrum. The details of these steps are described in Sections 4.2.2 to 4.2.6.

Once the peptide spectrum is obtained, computational techniques are employed to derive the sequence of the peptide. This involves either searching the protein database for a match, or *de novo* sequencing by just analyzing the spectrum. We will only discuss the latter method, in later sections. Table 4.2 summarizes the whole protein identification process just described.

Table 4.1: Average masses of amino acid residues

Residue	**Mass in *Da***		**Residue**	**Mass in *Da***
A	71.08		M	131.19
C	103.14		N	114.1
D	115.09		P	97.12
E	129.12		Q	128.13
F	147.18		R	156.19
G	57.05		S	87.08
H	137.14		T	101.1
I	113.16		V	99.13
K	128.17		W	186.21
L	113.16		Y	163.18

4.2.2 Digesting a Protein into Peptides

In the first step, the pure protein sample to be targeted is digested into short peptides using a protease (enzyme). The objective of this is to break down the protein sequencing problem into several smaller problems of short peptides (see Figure 4.2).

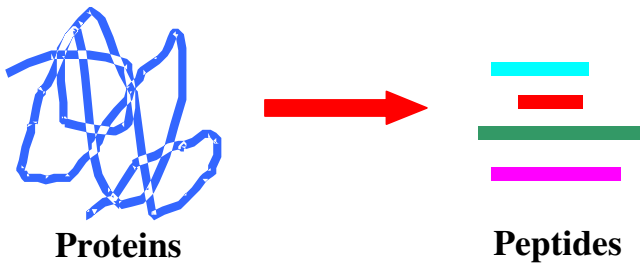

Figure 4.2: Digesting proteins into short peptides

For example, if the protease *trypsin* is used, the protein will be cut at K or R provided they are not followed by P. After digestion, we will get a set of peptides mostly ending with K or R:

Eg: ACCHC**K**CCV**R**PPC**R**CA → ACCHC**K**, CCVRPPC**R**, CA

Table 4.2: Protein identification process using LC-MS/MS method

Input: A Protein Sample
A. Biology Part:
1. Digest the protein into a set of peptides
2. By HPLC+Mass Spectrometer, separate the peptides
3. Select a particular peptide
4. Fragment the selected peptide
5. Obtain the tandem mass (MS/MS) spectrum of selected peptide
B. Computing Part:
• De Novo Sequencing
• Protein Database Search

In the above example, the protein is not cut at the first **R** (in bold), because the subsequent residue is a P.

4.2.3 Selecting a Particular Peptide

Given the mixture of peptides, the next step is to separate them so that every distinct peptide can be extracted from the mixture. High Performance Liquid Chromatography (HPLC) is first used to separate the mixture into regions. Then MS is performed on each region to obtain a spectrum containing peaks at different mass/charge values. Each peak represents a unique peptide ion with corresponding mass/charge value. The peptide corresponding to each peak is then selected, extracted and analyzed separately (see Figure 4.3).

4.2.4 Fragmentation of Peptide

Fragmentation at this stage involves breaking the selected peptide at random positions along the peptide backbone. Usually, fragmentation is performed by Collision Induced Dissociation (CID). The peptide is passed into a collision cell which has been pressurized with argon. Since argon is an inert gas, no chemical reaction takes place. The high pressure causes the peptide ions to collide with the argon atoms, thus breaking the bonds along the peptide backbone.

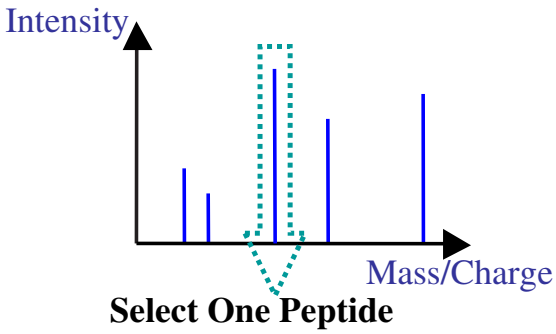

Figure 4.3: Spectrum of peaks after HPLC and MS – Each represents one peptide

Since most of the bonds are broken along the peptide backbone, the types of bonds broken are the C–C, C–N, N–C bonds. The resulting ions are termed a-ions, b-ions, c-ions, x-ions, y-ions and z-ions. Figure 4.4 shows how a peptide with two amino acids is fragmented. The resulting fragment ions with the N-terminus are the a-, b-, and c-ions, while the ions with the C-terminus are the x-, y-, and z-ions.

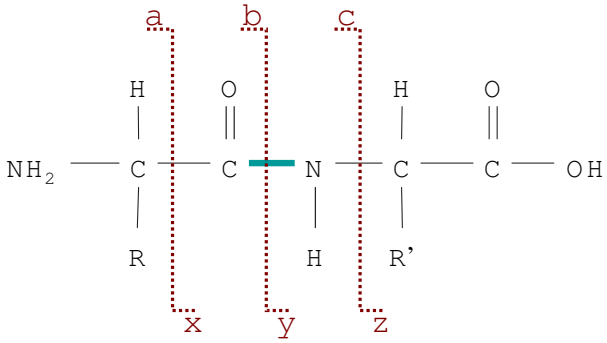

Figure 4.4: Ions resulting from a peptide with 2 amino acids

Based on experimental results, the relative abundance of y-ions is greater than that of b-ions. Since the peptide C–N bonds are more likely to be broken during fragmentation, the abundance of other ions are even smaller.

4.2.5 Calculation of the Mass of Ions

Consider a peptide with n amino acid residues. As shown in Figure 4.5, the breaking of one peptide bond between the ith and $(i+1)$th residues results in a b-ion of length–i and a y-ion of length–$(n-i)$. The fragmentation causes ionization such that the y-ion gains 2 more H atoms. The residues in the figure are represented in blue and red, while the letters in black are the extra H and O atoms in both ions after considering all residues.

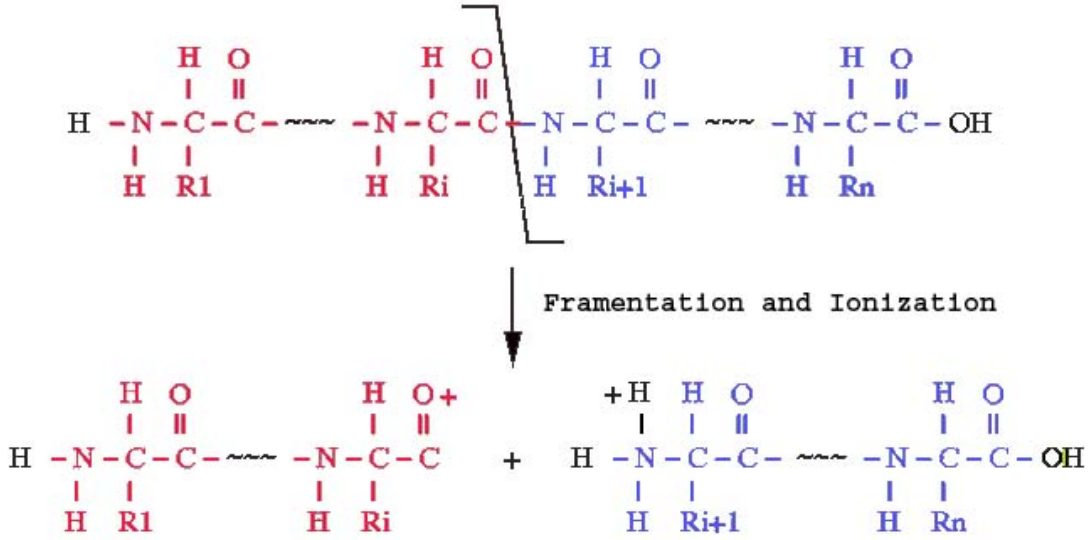

Figure 4.5: Fragmentation of a peptide into a b-ion (bottom left) and a y-ion (bottom right)

With the knowledge of how the fragmentation occurs, we can calculate the mass of resulting b-ions and y-ions, so that we know what we should be expecting from the MS spectrum after the fragmentation process.

Here we shall define some notations. Let A be the set of amino acid residues. For every residue $a \in A$, the mass of this residue is $w(a)$. Let $P = a_1 a_2 \ldots a_k$ represent a peptide that we are sequencing. We define $w(P) = \sum_{1 \le j \le k} w(a_j)$. The actual mass of the peptide sequence P is actually $w(P) + 18$, since it is the sum of all residues plus and an extra H_2O (See Figure 4.5. Note that the mass of O is 16 and that of H is 1).

After fragmentation, the mass of the b-ion with the first i amino acid

residues is:

$$b_i = 1 + w(a_1 a_2 \dots a_i) \tag{4.1}$$

since it has 1 extra H after considering all the residues. The mass of the y-ion with the last i amino acids is:

$$y_i = 19 + w(a_{k-i+1} a_{k-i+2} \dots a_k) \tag{4.2}$$

since it has 3 extra H and 1 O after considering all the residues. Therefore, the sum of the masses of both ions is equal to the mass of the peptide plus the masses of 4 H and 1 O:

$$b_i + y_{k-i} = 20 + w(P) \tag{4.3}$$

For example, consider the peptide P=SAG,

$w(P) = w(S) + w(A) + w(G) = 87.08 + 71.08 + 57.05 = 215.21$
Actual mass of $P = w(P) + 18 = 233.21$
$y_0 = 19$
$y_1 = w(G) + 19 = 76.05$
$y_2 = w(AG) + 19 = 147.13$
$y_3 = w(SAG) + 19 = 234.21$
$b_1 = w(S) + 1 = 88.08$
$b_2 = w(SA) + 1 = 159.16$
$b_3 = w(SAG) + 1 = 216.21$

4.2.6 Tandem Mass Spectrum (MS/MS Spectrum)

After fragmentation in the collision chamber, MS is performed again to obtain the MS/MS spectrum of the peptide. Figure 4.6 shows an example of such a spectrum. The fragments are separated according to their mass/charge values and their relative abundance is represented by the corresponding intensity values (height of peaks). Most of the significant peaks are due to the b-ions and y-ions. The other peaks may be due to ions resulting from the fragmentation at bonds other than the peptide bonds joining each amino acid residue. The figure also shows the contribution of each amino acid of a particular peptide to the spectrum.

Mathematically, an MS/MS spectrum is represented as:

$$M = \{(x_i, h_i) \mid 1 \leq i \leq k\} \tag{4.4}$$

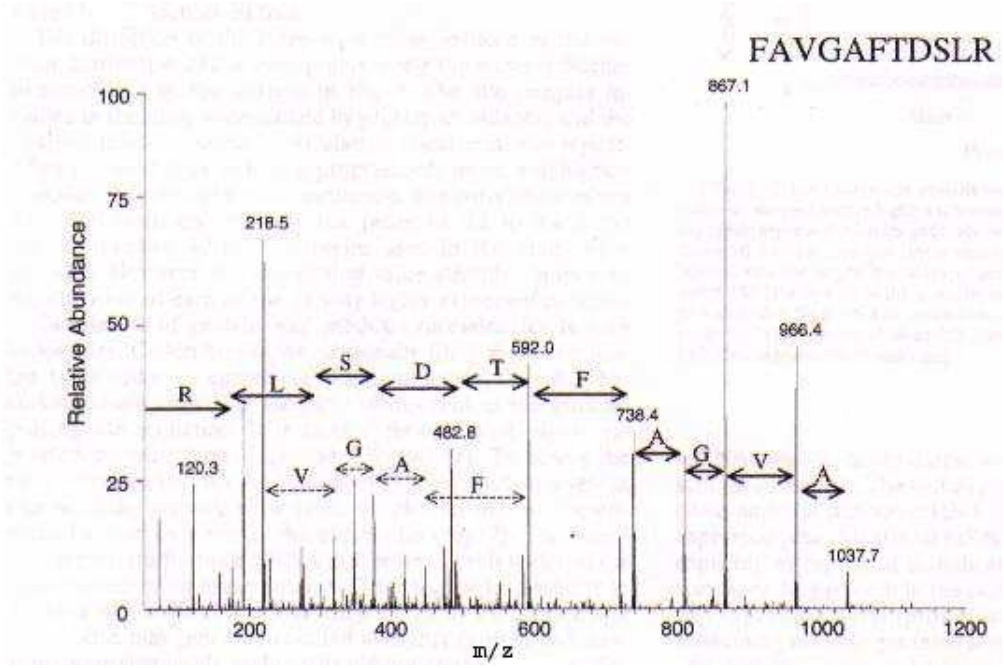

Figure 4.6: A tandem mass spectrum (MS/MS spectrum)

where x_i is the mass/charge (m/z) value for the ith peak and h_i is its intensity (or relative abundance).

For all the values x_i, we assume the ions are singly charged. In actual experiments, however, an ion may have charges greater than one. Fortunately, if a spectrum has peaks corresponding to multiply charged ions, there exists standard methods to convert those peaks to their singly charged equivalents. Thus in our subsequent calculations, we shall continue to consider only the singly charged case.

4.3 *De Novo* Peptide Sequencing

After performing the MS/MS experiment, we have obtained two important pieces of information about the peptide to be sequenced. First, we have the MS/MS spectrum which gives the m/z values of resulting ions from the fragmented peptide. Second, we know the mass of the peptide because this was known from the first MS used to separate peptides. Thus the computational problem for *de novo* peptide sequencing is as such:

Input:
- A MS/MS spectrum M
- The total mass wt of the peptide
- An error bound δ (default $\delta = 0.5$)

Output:
- The peptide sequence P

Given the spectrum, the mass of the peptide, and the error bound, the objective is to compute and output the sequence (with higher accuracy and lesser complexity). The error bound is to accommodate for the possibilities of error due to the limited accuracy of MS (which is commonly ± 0.5Da).

4.3.1 A Simple Scoring Scheme that considers only y-ions

A scoring scheme needs to be defined in order to lay down a criteria for assessing the suitability (or correctness) of a predicted sequence. Consider a peptide $P = a_1 a_2 \dots a_k$. Recall that y-ions are expected to have the highest intensities in the spectrum. If M is a spectrum for P, we expect to find peaks at m/z values y_i for $i = 1, 2, \dots, k$. Thus, we define the score function to be the sum of the intensities of all peaks, within the distance of the error bound, corresponding to each y-ion's m/z value:

$$\text{score}(M, P) = \sum \{h \mid (x, h) \in M, |x - y_i| \leq \delta \text{ for } i = 1, 2, \dots, k\} \quad (4.5)$$

For example, consider the peptide P=SAG, and the spectrum shown in Figure 4.7. The calculation of the score for P would be:

$y_1 = 57.05 + 19 = 76.05$
$y_2 = 57.05 + 71.08 + 19 = 147.13$
$y_3 = 57.05 + 71.08 + 87.08 + 19 = 234.21$
Therefore, $\text{score}(M, P) = 210 + 405 + 0 = 615$

First, the theoretical m/z values of the y-ions for P are calculated using Equation 4.2. At the m/z values of around 76 and 147, there are peaks of intensities 210 and 405 respectively. Thus using Equation 4.5, the score for P on M would be the sum of these intensities. Note that there is no peak at m/z value of around 234, thus the intensity is 0.

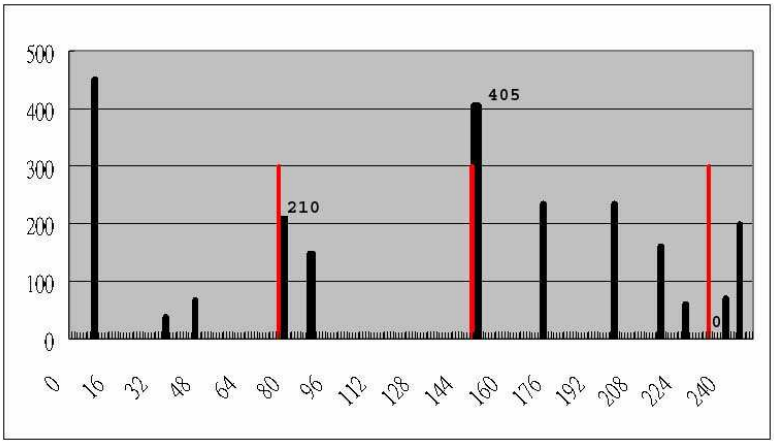

Figure 4.7: Example spectrum M (real peaks are in black), with ideal y-ion peaks (in red) superimposed onto it

Intuitively, this scoring scheme is equivalent to comparing the theoretical spectrum of y-ions (shown in Figure 4.8) with the actual spectrum. The theoretical spectrum consists of peaks for the ideal case where only the peptide bonds are broken and that they all have equal chance of breaking. Thus the peaks will have the same intensity and at exactly the m/z values that we calculated for the y-ions. The objective hence becomes finding the theoretical spectrum that best matches the real spectrum M.

Having defined the simple scoring scheme, we can now write a more refined definition for the sequencing problem:

Input:
- A MS/MS spectrum M
- The total mass wt of the peptide
- An error bound δ

Output:
- A peptide P, such that $(wt - \delta) \leq w(P) \leq (wt + \delta)$, which maximizes score$(M, P)$

The objective is to find a peptide P that attains the maximum score while under the constraint that its mass must be within $\pm\delta$ of the actual peptide mass, wt.

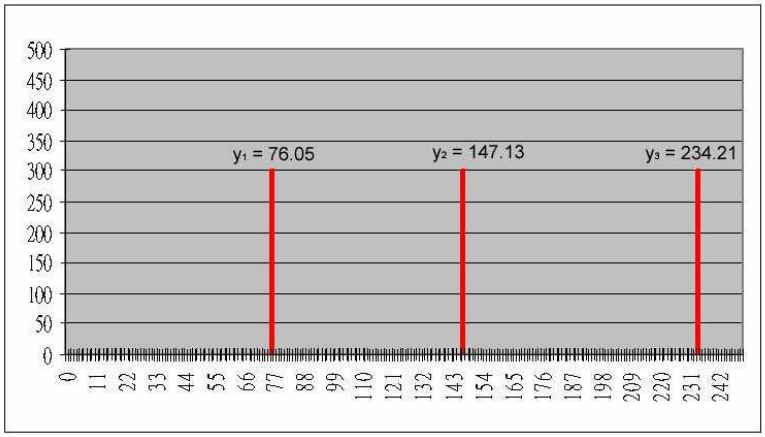

Figure 4.8: Theoretical (ideal) spectrum of y-ions

4.3.2 A Brute Force Solution

A straightforward way of solving the problem is to iterate through every possible peptide P with mass $w(P)$ within $\pm\delta$ of wt, i.e. $|w(P) - wt| \leq \delta$. During every iteration, score(M, P) is computed and the combination that has the highest score is reported, giving the derived sequence. However, this method runs in exponential time as there are exponential possible peptides. It also makes many redundant calculations of the score.

4.3.3 A Simple Dynamic Programming Solution

The idea in this dynamic programming solution is to identify the amino acid residues one by one from right to left (along the m/z axis). When we have identified the rightmost k residues, we would have discovered $y_1, y_2, \ldots, y_k$.

Let $V(r)$ be the maximum score, score(M, P), among all possible P such that $w(P) + 19 = r$. In other words, suppose we are told $y_k = r$ (recall that $y_k = w(P) + 19$), then $V(r)$ shall be the score(M, P) computed using Equation 4.5.

Our aim is to find $V(r)$ such that $|r - wt| \leq \delta$. In other words, we want to find the $V(r)$ for the r value that is close to the total mass of our sample peptide. Once we have found this, we can recover the peptide sequence by backtracking.

Let

$$f_M(r) = \sum\{h \mid (x,h) \in M \text{ and } |x-r| \leq \delta\}. \tag{4.6}$$

$f_M(r)$ is the sum of all peaks in M whose mass is close to r (within $\pm\delta$ of r). Thus the maximum score at r is

$$V(r) = \max_{a \in A}\{V(r - w(a)) + f_M(r)\} \tag{4.7}$$

where A represents the set of all possible amino acid residues (19 of them if we consider I and L as one residue). $V(r - w(a))$ gives the maximum score if we take a particular residue to be the next in the sequence.

The algorithm iterates through all values of r between 0 and wt, and applies the above recursive formula during each iteration. The step size is usually taken to be 0.01. The calculation of $V(r - w(a))$ requires the algorithm to look back at previously computed $V(r)$ values. The sequence is known once the final $V(r)$ value is found, by noting the amino acid residues used in this final recursion.

For example, given the spectrum M and the peptide mass $wt = 234.21$ (we take it that the peptide has an extra H, so $wt = 19 + \sum\{allresidues\}$), we compute $V(r)$ for values of r from 0 to 234.21. In this example, we take the step size to be 0.01. After each iteration, the $V(r)$ is stored for potential use in calculating $V(r)$ of larger r values in subsequent iterations. During each iteration, the algorithm finds the score for every amino acid type, by using the values $V(r - w(a))$ that has been computed in earlier iterations. Only the maximum score is taken as $V(r)$. When r reaches wt, the sequence is derived as we perform the recursion and backtrack to $V(0)$. The computation process is shown in Table 4.3.3, where the calculations at certain iterations are elaborated.

4.3.4 Time Analysis of Simple Dynamic Programming Solution

We need to fill-in the $V(r)$ table with a number of entries that is proportional to wt (if step size is 0.01, then we need to fill $wt \times 100$ entries). Each entry can be computed in O($|A|$). So, total time complexity is O($|A|wt$). Since $|A|$ is constant and fixed, the time complexity is just linear to the mass of peptide, which roughly depends on the peptide length.

Table 4.3: Example run for the simple dynamic programming algorithm

when $r = 0.00$,	$V(0.00) = 0$
when $r = 0.01$,	$V(0.01) = \cdots$
$\vdots$	$\vdots$
when $r = 76.05$,	$V(76.05) = \max\{V(r - w(A)), \ldots, V(r - w(G)), \ldots\}$ $+f_M(76.05)$ $= V(19) + 210$ (we take $V(r) = 0$, $\forall r <$lightest y-ion) $= 0 + 210 = 210$ (G was the residue that gave this maximum)
$\vdots$	$\vdots$
when $r = 147.13$,	$V(147.13) = \max\{V(r - w(A)), \ldots, V(r - w(Y))\}$ $+f_M(147.13)$ $= \max\{V(147.13 - 71.08), \ldots, V(147.13 - 163.18)\}$ $+f_M(147.13)$ $= V(76.05) + 405 = 210 + 405 = 615$ (A was the residue that gave this maximum)
$\vdots$	$\vdots$
when $r = 234.21$,	$V(234.21) = \max\{V(r - w(A)), \ldots, V(r - w(S)), \ldots\}$ $+f_M(234.21)$ $= \max\{V(234.21 - 71.08), \ldots, V(234.21 - 87.08)\}$ $+f_M(234.21)$ $= V(147.13) + 0 = 615 + 0 = 615$ (S was the residue that gave this maximum)

4.4 Using Both y-ions and b-ions to Determine Peptide Sequence

The simple scoring scheme and dynamic programming (DP) algorithm described earlier takes into account only the y-ions. However, fragmentation of a peptide produces pairs of ions, i.e. y-ions and b-ions. In order to obtain a better estimate of the similarity score of a mass spectrum M and a hypothetical peptide sequence P, we need to make use of the information provided by both y-ions and b-ions.

4.4.1 An Improved Scoring Scheme

Consider a peptide $P = a_1 a_2 \dots a_k$. If M is a mass spectrum for peptide P, M should contain peaks for the y-ions $y_{1,\dots,k}$ and b-ions $b_{1,\dots,k}$ of P. Hence the score function is redefined as:

$$\begin{aligned} score(M,P) \;\; = \;\; & \sum_{\forall i=1,\dots,k} \{h | (x,h) \in M, \\ & |x - y_i| \;\leq\; \delta \text{ or } |x - b_i| \;\leq\; \delta \;\} \end{aligned} \tag{4.8}$$

Equation 4.8 measures the sum of all peaks in M whose mass is close to any of the y-ions or b-ions of P. δ is the maximum error of the mass spectrometer and its value is typically between 0 and 0.5.

For example, given a mass spectrum M, we consider a possible peptide $P = SAG$ and calculate the masses of various y-ions and b-ions that can be obtained after fragmenting SAG. The following calculations determine the positions of the y-ion and b-ion peaks as shown in Figure 4.9. The y-ions peaks are drawn higher than the b-ion peaks to show that they are the most often detected ions in the mass spectrum.

$y_1 = w(G) + 19 = 76.05$
$y_2 = w(AG) + 19 = 147.13$
$y_3 = w(SAG) + 19 = 234.21$
$b_1 = w(S) + 1 = 88.08$
$b_2 = w(SA) + 1 = 159.16$
$b_3 = w(SAG) + 1 = 216.21$

As shown in Figure 4.10, matching the hypothetical peaks in Figure 4.9 to the real mass spectrum M, will produce a similarity score, $score(M,P) = f_M(y_1) + f_M(y_2) + f_M(b_1) + f_M(b_3) = 210 + 405 + 150 + 160 = 925$. If this similarity score is the highest among all possible peptides, we deduce that $P = SAG$ is the optimal peptide sequence for M.

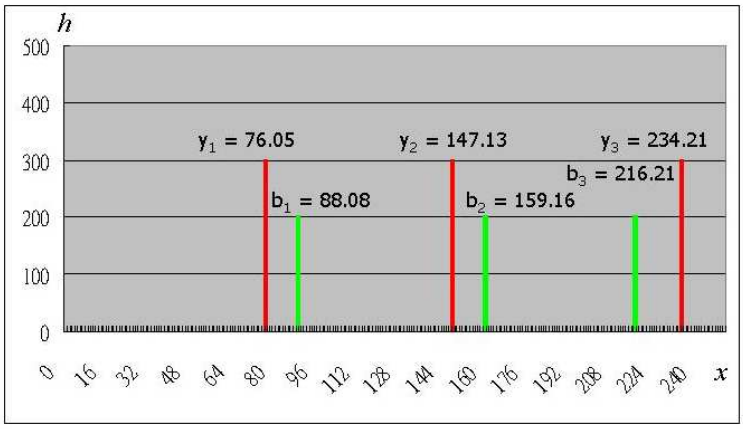

Figure 4.9: Theoretical (ideal) spectrum of a hypothetical peptide, showing peaks for y-ions (in red) and b-ions (in green)

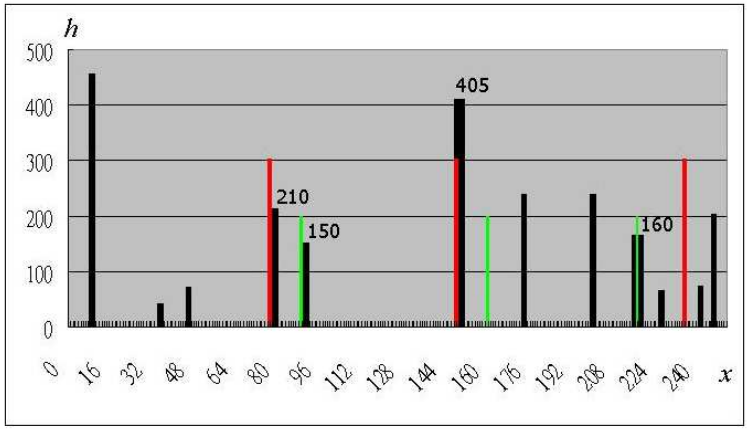

Figure 4.10: Real spectrum M with ideal peaks of y-ions and b-ions superimposed onto it (Real peaks are in black and numbers are the intensities of the black peaks.)

4.4.2 Why the Previous DP Algorithm Cannot Be Used in This Scenario

The previous DP algorithm cannot be used in this scenario because whenever a peak in M is matched by two ions (e.g. a b-ion and a y-ion of approximately equal mass), the height of this peak is counted twice. Such an algorithm will

tend to match the highest peaks more than once, rather than match more peaks. Hence, we need a modified DP algorithm that determines whether a peak has been matched earlier, before it evaluates the match between the peak and a new ion [MZL03]. A peak that has been matched before will not be counted again. Therefore, each peak will serve as only one piece of supporting evidence.

For example, in Figure 4.11, a peak may be matched by two ions - a b-ion b_i and a y-ion y_j. Such a peak should only be summed either once in $f_M(b_i)$ or in $f_M(y_j)$ but not twice.

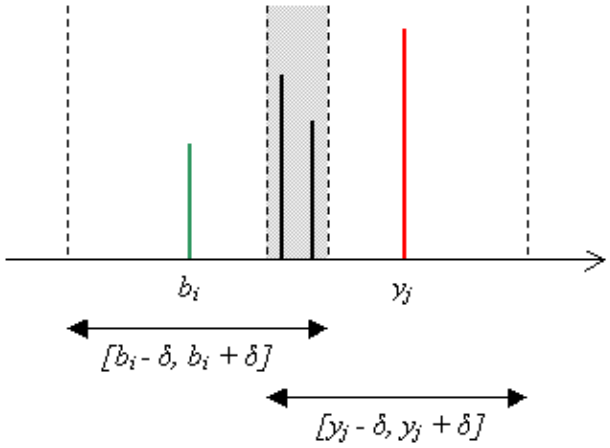

Figure 4.11: Peaks that are matched by two ions b_i and y_j should be summed only once.

4.4.3 Observations

Given a peptide $P = a_1 a_2 \dots a_k$, there are two important observations:

1. The pairs of complementary b-ions and y-ions, (b_i, y_{k-i}) $\forall i = 1, \dots, k$, form a set of nested regions, as shown in Figure 4.12. Every b-ion in the first half of the mass spectrum has a complementary y-ion in the second half of the mass spectrum. The same applies for every y-ion and its complementary b-ion.

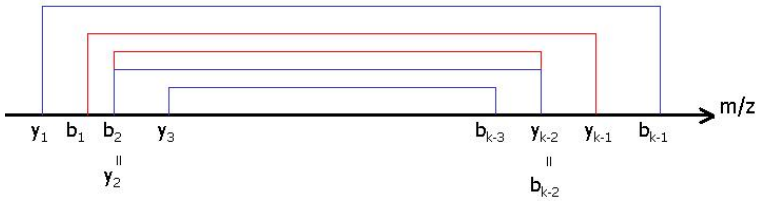

Figure 4.12: Complementary y-ions and b-ions form a set of nested regions.

This is a useful observation because the modified DP algorithm can proceed from left to right of the first half of the mass spectrum, while summing up the intensities of every b-ion (and its complementary y-ion) and every y-ion (and its complementary b-ion). This means that the intensities are summed up, beginning from the outermost ion-pair to the innermost ion-pair in Figure 4.12.

2. Both b_i and y_j are strictly increasing. This means that $b_{i-1} < b_i$ and $y_{j-1} < y_j$ $\forall i, j = 1, \ldots, k$. This is because appending a residue to the front or the end of an ion increases the ion's mass. In particular,

$$b_i - b_{i-1} \geq min_{a \in A} w(a) = 57.05 \tag{4.9}$$

$$y_j - y_{j-1} \geq min_{a \in A} w(a) = 57.05 \tag{4.10}$$

This observation implies that the peaks of masses b_{i-1} and b_i, y_{j-1} and y_j, $\forall i, j$ will not overlap. Hence, the modified DP algorithm only needs to detect possible overlapping of peaks between b_i and y_j (of approximately equal mass) and prevent double-counting of their peaks.

4.4.4 A Modified Dynamic Programming Solution

Consider a peptide $P = P_1P_2P_3$ (Figure 4.13) that is partitioned into three regions P_1, P_2 and P_3.

Let r be the mass of the b-ion b_i formed by cleavage of the peptide backbone to the right of residue a_i, and let s be the mass of the y-ion y_j formed by cleavage of the peptide backbone to the left of residue a_{k-j+1}.

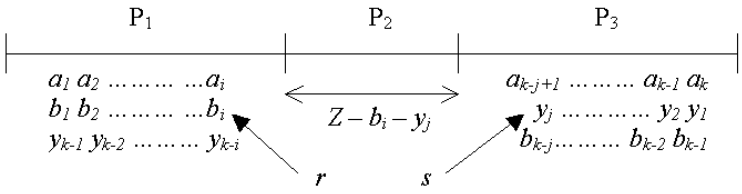

Figure 4.13: Peptide $P = P_1P_2P_3$

The terms used in the modified DP algorithm are defined in Table 4.4.

wt	Total mass of the peptide studied
Z	$wt + 20$
$\hat{a}$	$max_{a \in A} w(a) = w(W) = 186.21$
r	Mass of the b-ion consisting of peptide P_1 (See Figure 4.13)
s	Mass of the y-ion consisting of peptide P_3 (See Figure 4.13)
$V(r, s)$	The maximum $score(M, P)$ among all possible peptide sequences $P = P_1P_2P_3$
$f_M(u, v)$	The sum of all peaks in M whose mass is close to u (and $Z - u$) but is not close to v (and $Z - v$).

Table 4.4: Definition of terms used in the modified DP algorithm

The recurrence formula for calculating $V(r, s)$ for all (r, s) such that $|r - s| \leq \hat{a}$, is given below:

$$V(r,s) = \begin{cases} max_{a \in A}\{V(r, s - w(a)) + f_M(s, r)\} & \text{if } r < s \\ max_{a \in A}\{V(r - w(a), s) + f_M(r, s)\} & \text{if } r > s \\ max_{a \in A}\{V(r - w(a), s), V(r, s - w(a))\} & \text{if } r = s \end{cases}$$

Let us consider the recurrence formula in the case when $r > s$. In Figure 4.14, when $r > s$, $V(r, s)$ is computed as the addition of the highest sum of all previously matched peaks as denoted by $V(r, s - w(a))$, and the sum of the currently matched peaks as denoted by $f_M(s, r)$. $f_M(s, r)$ prevents double-counting of peaks by selectively adding the peaks whose mass is close

to s (and close to the complementary mass $Z - s$) but not close to r (and not close to the complementary mass $Z - r$). Each time $V(r, s)$ is computed, because the amino acid is still unknown at that point in time, scores for all $a \in A$ have to be computed. The amino acid that gives the maximum score is marked for backtracking purposes at the end of the algorithm. During backtracking, each marked amino acid is appended either to the C-terminus of P_1 or the N-terminus of P_3. In Figure 4.14, the marked amino acid will be appended to the N-terminus of P_3 during backtracking at the end of the algorithm. P_1 and P_3 will continuously be extended rightwards and leftwards respectively, until there remains a single amino acid between them in the region P_2.

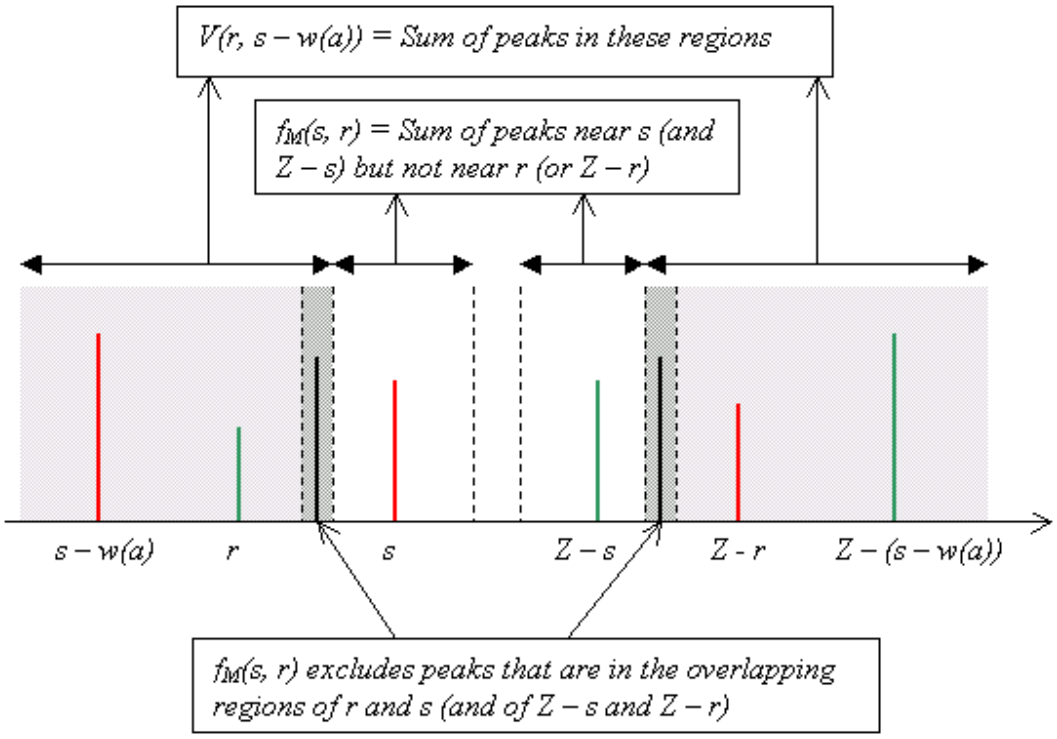

Figure 4.14: Calculation of $V(r, s)$ when $r > s$

The modified DP algorithm is shown in Table 4.5. It uses the recurrence formula shown earlier.

$V(0,0) = 0; V(u,v) = -\infty$ for $(u,v) \neq (0,0)$;
for r from 1 to $Z/2 + \hat{a}$ step 0.01 do
  for s from $(r - \hat{a})$ to $(r + \hat{a})$ step 0.01 do
    for $a \in A$ do
      if $r < s$ then
        $V(r,s) = V(r, s - w(a)) + f_M(s,r)$
      else if $r > s$ then
        $V(r,s) = V(r - w(a), s) + f_M(r,s)$
      else
        $V(r,s) = max\{V(r - w(a), s), V(r, s - w(a))\}$
find r, s, a such that $r + s + w(a) = Z$ and $V(r,s)$ is maximized.
find QaR by backtracking.

Table 4.5: The modified DP algorithm (adapted from [MZL03])

Suppose that (Q,R) is the pair (P_1, P_3) that maximizes $V(r,s)$ such that $r + s + w(a) = Z$. In other words, QaR is the optimal peptide sequence and the aim is to find Q, R and $a \in A$. Based on the algorithm in Table 4.5, we construct a table of $V(r,s)$ values as shown in Figure 4.15. Only $V(r,s)$ values for all (r,s) such that $|r - s| \leq \hat{a}$ need to be calculated. Assuming a step of 0.01, the table is populated starting from $V(r = 0.00, s = 0.00)$ and proceeding row-by-row until $V(r = Z/2 + \hat{a}, s = Z/2 + \hat{a})$. At the end of calculating all $V(r,s)$ values, backtracking is carried out to determine the optimal sequence QaR.

4.4.5 An Example for the Modified DP Algorithm

Table 4.6 illustrates how the $V(r,s)$ values are calculated starting from $V(0,0)$ until $V(Z/2 + \hat{a}, Z/2 + \hat{a})$. The reason why the algorithm scans only the first half of the mass spectrum is because while the first half of the mass spectrum is scanned, the peaks of the second half are summed together with the peaks of the first half. Note that in the above calculations, the intermediate values of r and s where there are no real peaks in mass spectrum M, are not shown.

At the end of scanning the first half of the mass spectrum and after having compared the hypothetical peaks in Figure 4.16 to the real peaks in M, all the

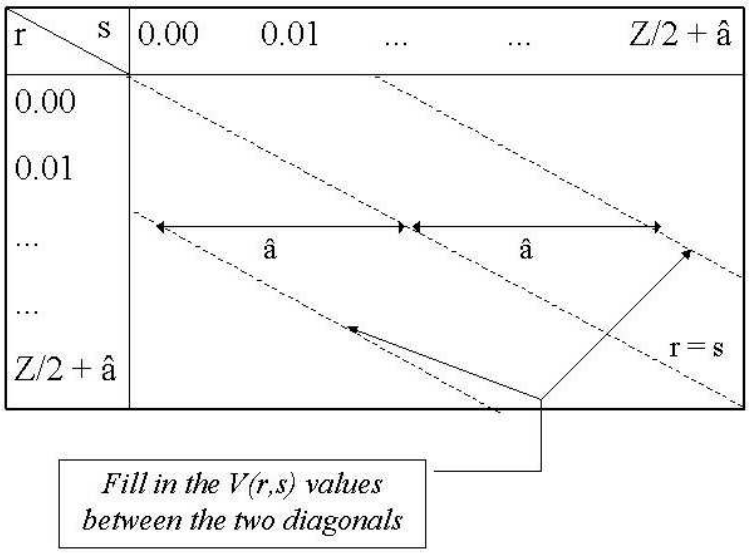

Figure 4.15: Table of $V(r, s)$ values

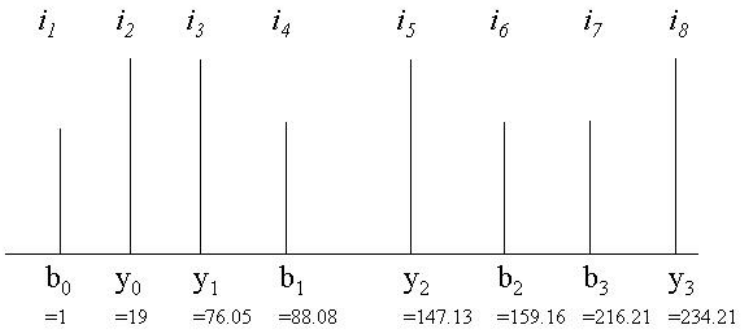

Figure 4.16: Peaks of hypothetical b-ions and y-ions of peptide $P = SAG$. i_1 to i_8 are the intensities (heights) of these peaks in M.

hypothetical peaks have been matched. With the constraint $r+s+w(a) = Z$, we find that if $a =$ 'A' and $r = 88.08, s = 76.05$, then $V(88.08, 76.05)$ gives the highest similarity score. This shows that $P_1 a P_3 = SAG$.

$V(0,0) = 0$ (base case)

$$\begin{aligned} V(1,0) &= V(0,0) + i_1 + i_8 \\ &= i_1 + i_8 \end{aligned}$$

$$\begin{aligned} V(1,19) &= V(1,0) + i_2 + i_7 \\ &= i_1 + i_2 + i_7 + i_8 \end{aligned}$$

$$\begin{aligned} V(1,76.05) &= max_{a \in A}\{V(1, 76.05 - w(a)) + f_M(76.05, 1)\} \\ &= V(1,19) + f_M(76.05,1) \text{ (due to } a = \text{'G')} \\ &= V(1,19) + i_3 + i_6 \\ &= i_1 + i_2 + i_3 + i_6 + i_7 + i_8 \end{aligned}$$

$$\begin{aligned} V(88.08,76.05) &= max_{a \in A}\{V(88.08 - w(a), 76.05) + f_M(88.08, 76.05)\} \\ &= V(1,76.05) + f_M(88.08,76.05) \text{ (due to } a = \text{'S')} \\ &= V(1,76.05) + i_4 + i_5 \\ &= i_1 + \ldots + i_8 \end{aligned}$$

Table 4.6: Example run of the modified DP algorithm

4.4.6 Time Analysis of the Modified DP Algorithm

$V(r,s)$ must be computed for all r and s such that $|r - s| \leq \hat{a}$. Hence, we need to compute $wt \cdot \hat{a}$ entries and each entry can be computed in $O(|A|)$ time i.e. constant time. Therefore, the time complexity of this second DP algorithm is $O(wt \cdot \hat{a} \cdot |A|)$ time i.e. linear with respect to the total mass of the peptide wt.

Comparing this with the time complexity of the previous DP algorithm, shows that this second algorithm will run at a factor $\hat{a}$ slower.

References

[MZL03] Ma, B., Zhang, K. and Liang, C., "An effective algorithm for the peptide de novo sequencing from MS/MS spectrum.", *CPM 2003*, LNCS 2676, 266-277, 2003.

[P00] Pevzner, P.A., "Computational Molecular Biology: An algorithmic approach.", *MIT Press*, 2000.