

12.1 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. However, protein sequencing is currently still a very difficult problem. Instead, current research focuses on peptide sequencing, the identification of the amino acid sequence of a short fragment of a protein.

The problem of peptide sequencing may be formulated as follows: given some experimentally determined characteristics of a peptide, determine the peptide sequence. In addition, the peptide sequence may either be constrained to be within some database (peptide identification), or unconstrained (*de novo* peptide sequencing). An obvious means of identifying the peptide sequence is to first derive a model of the experimentally determined characteristics of a peptide and then find some sequence whose predicted characteristics best match the experimental characteristics.

In this lecture, the mass spectrum of a peptide's fragments is used as the experimentally determined characteristic. This is the dominant characteristic used because mass spectrometry (MS) is fast and sensitive.

12.1.1 Mass spectrometry

Mass spectrometry (MS) is used to separate the components of a sample according to their mass¹, commonly expressed in Daltons (*Da*).

Figure 12.1 shows a simple hypothetical MS spectrum derived from a sample with three components (assuming the absence noise). Each component produces a peak in the MS spectrum at its mass value. The height of a peak indicates the relative abundance of its component. Depending on the equipment used, the accuracy of the mass measurement ranges from ± 0.01 to ± 0.5 Da.

¹In reality, the mass/charge ratio m/z is the quantity that the components are separated by, but a unit charge assumption is normally taken. In these notes, the term 'mass' will be exclusively used except in cases where the difference matters.

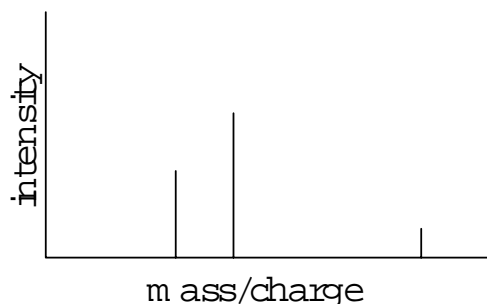


Figure 12.1: A simple mass spectrum

12.2 Obtaining the mass spectrum of a peptide

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 plays a major role in determining its 3D conformation and charge distribution, ultimately determining its function². Thus, knowing this sequence is pertinent to further investigation of a protein's functions and involvement in biological processes.

Because of the difficulties in directly sequencing a protein, small peptide fragments are sequenced instead. To do so, the target protein sample is first digested by enzymes into many different peptide fragments. Each peptide is relatively short (≈ 10 amino acids) compared to the original protein. The mixture of peptides after digestion is first separated via High Performance Liquid Chromatography (HPLC) followed by mass spectrometry. The peptide of a particular mass is selected and further fragmented via collision induced disassociation. The tandem mass spectrum³ (MS/MS) is then obtained. These steps are the biological steps for generating a peptide spectrum. The details of these steps are described in Sections 12.2.1 to 12.2.4.

Once the peptide spectrum is obtained, computational techniques are employed to derive the sequence of the peptide. This involves either searching a protein database for a match, or *de novo* sequencing by analyzing the spectrum alone. Table 12.1 summarizes the entire process just described.

²Other proteins may also play a role in determining a polypeptide's shape and function. For example, prions are infectious agents made only of proteins which are believed to refold normal proteins into the abnormally structured form of the prion.

³Thus termed because of the repeated application of mass spectrometry.

Table 12.1: Protein identification process using LC-MS/MS method

Input: A Protein Sample

A. Biology/Experimental 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 MS/MS spectrum of the selected peptide

B. Computational Part:

1. De novo sequencing, or
2. Protein database search

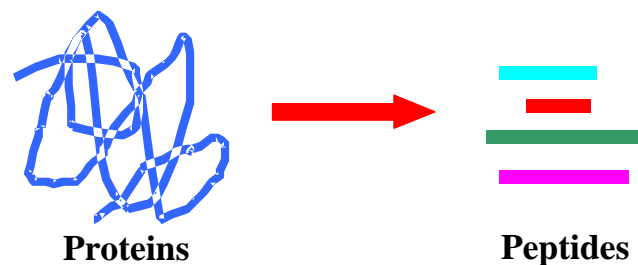
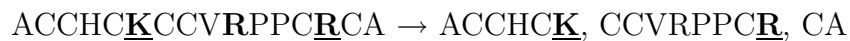


Figure 12.2: Digesting proteins into short peptides

12.2.1 Protein digestion

In the first step, a pure protein sample is digested into short peptides using a protease⁴ (see Figure 12.2).

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. For example, for



the protein is not cut at the first **R**, because the next residue is P.

⁴An enzyme which breaks peptide bonds between amino acids of proteins.

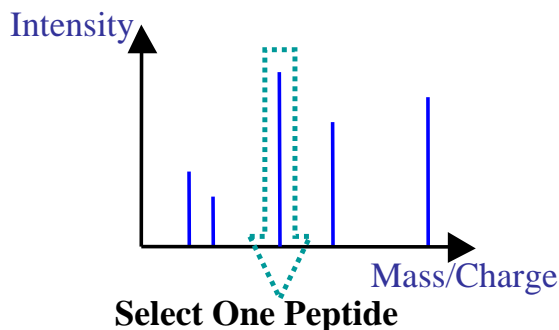


Figure 12.3: Spectrum of peaks after HPLC and MS – Each represents one peptide

12.2.2 Peptide separation and selection

Given the mixture of peptides, the next step is to separate them so that a distinct peptide can be extracted. HPLC is first used to separate the mixture into regions. MS is then performed on each region to separate the peptides in that region. Each peptide corresponding to a specific mass is then separately processed (see Figure 12.3).

12.2.3 Peptide fragmentation

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 containing a pressurized inert gas⁵ (e.g. argon). The high pressure causes the peptide ions to collide with the argon atoms, thus breaking the bonds along the peptide backbone.

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 12.4 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.

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 other ions are even less abundant.

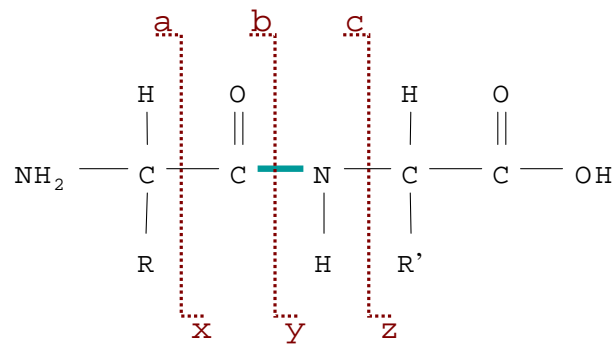


Figure 12.4: Ions resulting from a peptide with 2 amino acids

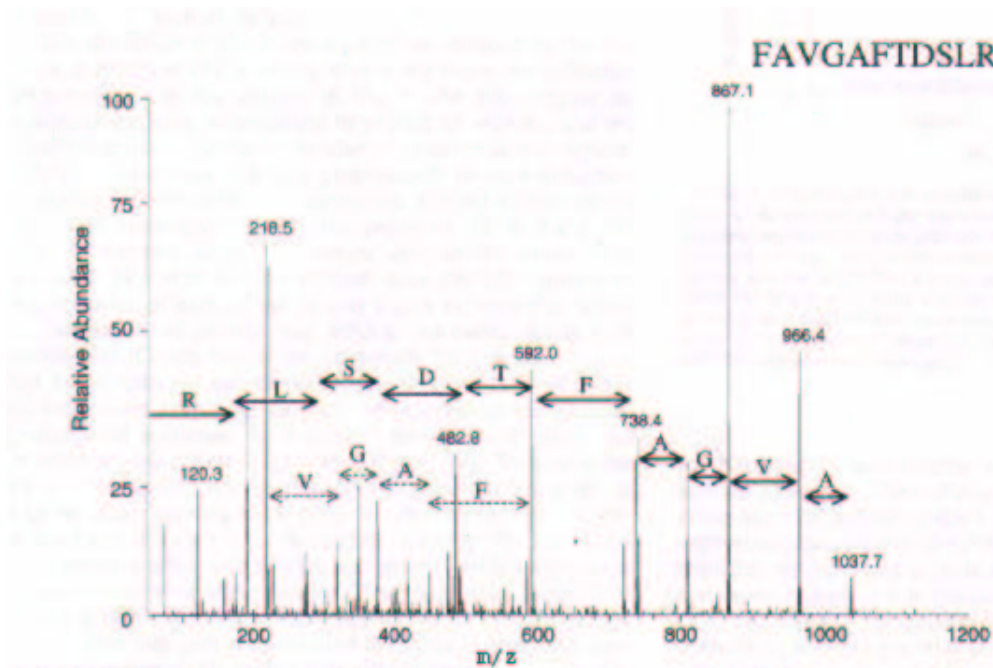


Figure 12.5: A tandem mass spectrum (MS/MS spectrum)

12.2.4 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 12.5 shows an example of such a spectrum. The fragments are separated according to their mass (represented horizontally) and their relative abundance is indicated vertically.

The MS may be preprocessed to compensate for multiply charged ions, or for noise removal, and is then converted to the following mathematical form:

$$M = \{(x_i, h_i) \mid 1 \leq i \leq k\} \quad (12.1)$$

where x_i is the mass of the i th peak and h_i is its relative abundance.

12.3 Modelling the MS of a fragmented peptide

In order to identify a peptide sequence from its MS, the MS of an arbitrary fragmented peptide sequence must first be modelled. The modelling problem is: given a peptide sequence, derive a prediction of its MS.

12.3.1 Amino acid residue mass

As peptides are composed of amino acid residues⁶, their mass must be known in order to derive a peptide's MS. Let A be the set of amino acids and $w(a)$ be the mass of some amino acid residue $a \in A$ (in Daltons). Table 12.2 shows the value of $w(a)$ for all 20 amino acids.

Note that $w(I) = w(L)$. Hence, these 2 amino acids are indistinguishable by any method based on mass and are treated as the same amino acid. The residue with the smallest mass is G (glycine) at 57.05 Da and the one with the largest mass is W (tryptophan) at 186.21 Da.

12.3.2 Fragment ion mass

As explained in Section 12.2.3, during the fragmentation of the peptide, ions are formed. The different peaks in the MS of the fragmented peptide thus correspond to the ions.

The most common fragmentation mode and its resultant ions (b- and y-ions) is shown in Figure 12.6. With the knowledge of how the fragmentation occurs, we can calculate the mass of resulting b-ions and y-ions.

Let $w(a_1 a_2 \dots a_k) = \sum_{i=1}^k w(a_i)$ be the total mass of an amino acid chain. Because a b-ion has an extra H ion, its mass is $w_b(C) = w(C) + 1$. Similarly, because a y-ion has 3 extra H ions and 1 extra O ion, its mass is $w_y(C) = w(C) + 19$.

⁵Thus avoiding any chemical reaction.

⁶What remains of an amino acid after the removal of a water molecule.

Table 12.2: Average mass $w(a)$ of amino acid residues

a	$w(a)$		a	$w(a)$
<i>A</i>	71.08		<i>M</i>	131.19
<i>C</i>	103.14		<i>N</i>	114.1
<i>D</i>	115.09		<i>P</i>	97.12
<i>E</i>	129.12		<i>Q</i>	128.13
<i>F</i>	147.18		<i>R</i>	156.19
<i>G</i>	57.05		<i>S</i>	87.08
<i>H</i>	137.14		<i>T</i>	101.1
<i>I</i>	113.16		<i>V</i>	99.13
<i>K</i>	128.17		<i>W</i>	186.21
<i>L</i>	113.16		<i>Y</i>	163.18

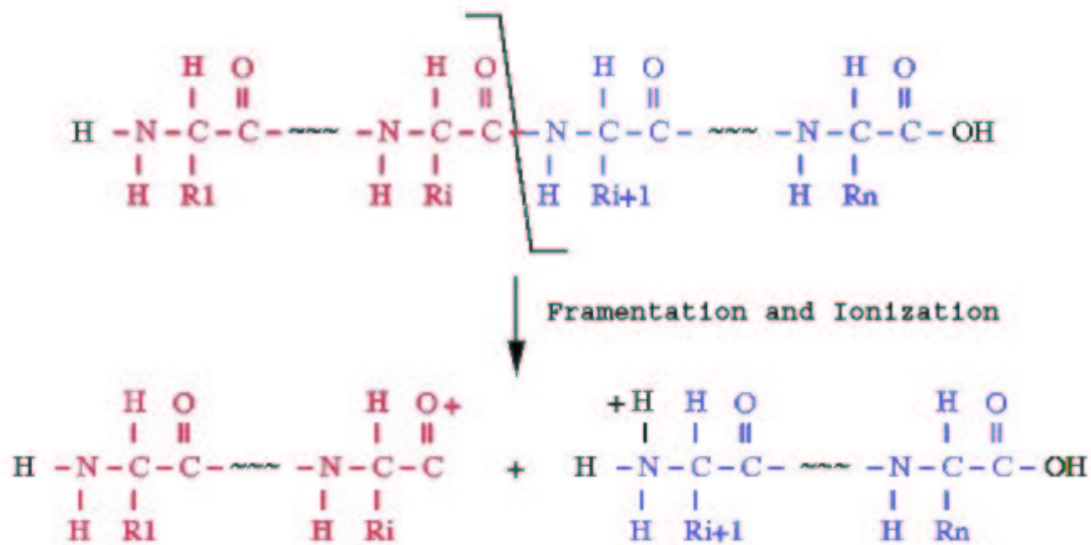


Figure 12.6: Fragmentation of a peptide into a b-ion (bottom left) and a y-ion (bottom right)

12.3.3 Aggregate mass spectrum (simplified)

It can be deduced that the mass spectrum of a fragmented peptide corresponds to the sum of the mass spectra of its fragments, weighted according to their frequency (probability) of occurrence.

However, for the sake of simplicity and illustration, the following assumptions are made for the rest of the discussion: (a) fragments only occur at a C-N bond (hence resulting in b- and y-ions), (b) each peptide molecule is fragmented at most one time, (c) each fragment has unit charge, and (d) the probability of fragmentation at each position on the peptide is uniform.

Consider a peptide $P = a_1a_2 \dots a_n$ with n amino acid residues. The mass of the peptide P is $w_p(P) = w(P) + 18$ due to the H- and -OH groups on either end of the peptide. The breaking of the peptide bond between the i th and $(i + 1)$ th residues results in a b-ion of length i and a y-ion of length $n - i$. The mass of the b-ion containing first i residues is $b_i = w_b(a_1 \dots a_i)$ and the mass of the y-ion containing the last j residues is $y_j = w_y(a_{n-j+1} \dots a_n)$.

For example, for a peptide $P = SAG$,

$$\begin{aligned} w(P) &= w(S) + w(A) + w(G) = 87.08 + 71.08 + 57.05 = 215.21 \\ w_p(P) &= w(P) + 18 = 233.21 \\ y_1 &= w_y(G) = w(G) + 19 = 76.05 \\ y_2 &= w_y(AG) = 147.13 \\ y_3 &= w_y(SAG) = 234.21 \\ b_1 &= w_b(S) = w(S) + 1 = 88.08 \\ b_2 &= w_b(SA) = 159.16 \\ b_3 &= w_b(SAG) = 216.21 \end{aligned}$$

Hence, we would expect a peak for each of the ions formed (a total of 6 peaks).

12.3.4 Aggregate mass spectrum (more detail)

The model in the previous section may be more developed by considering the following: (a) other ion types corresponding to different cleavage locations on the peptide backbone, (b) additional loss of ammonium or water molecules, (c) the presence of isotopic ions, (d) multiply charged ions, and (e) probability distribution of fragmentation over position and ion type. It can be seen that the first three 'complications' may be generalised as 'other ion types', where the mass of an 'ion type' is a constant added to the mass of a base ion type which is an amino acid residue. Multiply charged ions, on the other hand, correspond to a scaling of the mass/charge ratio, and may be compensated for by a process called deconvolution.

Other than these, multiple fragmentation of a single peptide molecule and noise also cause the actual mass spectrum to differ from theoretical spectrum.

12.4 Protein database search

A protein database search problem may be formulated as follows: given a database of proteins D , a raw MS/MS spectrum M , and the mass $w \pm \delta$ of the peptide used to produce M , find a protein from D containing a peptide likely to have a mass of w and a MS/MS spectrum similar to M .

An obvious and brute-force approach is to compute the theoretical spectrum of each mass $w \pm \delta$ peptide in the database, and compare the matching score of the theoretical and experimental spectra. Some scoring functions will be examined in the following section on *de novo* peptide sequencing.

12.5 *De novo* peptide sequencing

A *de novo* peptide sequence problem may be formulated as follows: given a raw MS/MS spectrum M , and the mass $w \pm \delta$ of the peptide used to produce M , find the peptide sequence.

This problem may be solved in a fashion similar to a database search, with a database containing all the possible peptides.

Thus, the output of the algorithm would be the peptide sequence whose theoretical spectrum is closest to M (where closeness is defined by the scoring function) and whose mass is $w \pm \delta$.

12.5.1 Scoring by considering y-ions

Because fragmentation produces y-ions with relative abundance, we expect the MS/MS spectrum to contain peaks at masses corresponding to the mass of all the y-ions which can be generated from the peptide under test. Then, a simple score function would be:

$$\text{score}(M, P) = \sum \{h | (x, h) \in M, |x - y_i| \leq \delta \text{ for } i = 1, 2, \dots, k\} \quad (12.2)$$

Essentially, this is a correlation of M with a theoretical spectrum consisting of unit peaks at each y-ion mass.

For example, consider the peptide $P = \text{SAG}$, and the spectrum shown in Figure 12.7. The calculation of the score for P would be $210 + 405 + 0 = 615$.

12.5.2 A Brute Force Solution

A straightforward way of solving the problem is to iterate through every possible peptide P with $|w_p(P) - w| \leq \delta$, and report $\arg \max_P \text{score}(M, P)$. However, this is an exponential time algorithm.

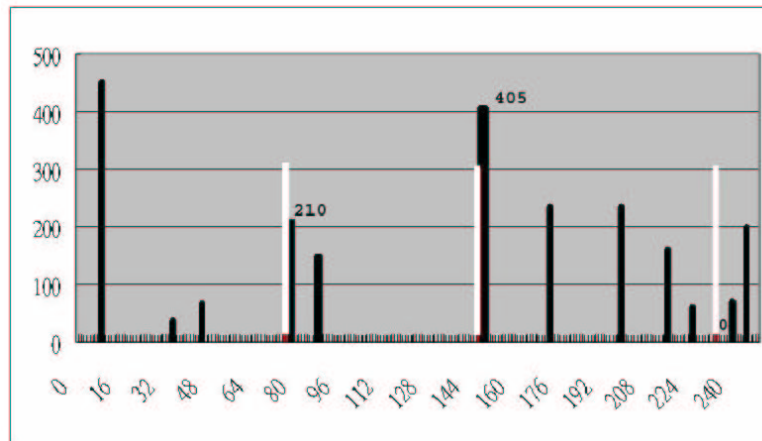


Figure 12.7: Example spectrum M (real peaks are in black), with ideal y-ion peaks (in white) superimposed onto it

12.5.3 A Simple Dynamic Programming Solution

We use the assumption that each peptide molecule is cleaved at only one position, and the fact that with the simple scoring scheme, all residues are represented by suffices of target sequence. The idea in this dynamic programming solution is to identify the amino acid residues one by one from rightmost to leftmost amino acid (from y_1 to y_k).

Let $V(r) = \max_{w_y(P)=r} \{score(M, P)\}$. Our aim is to find $\max_{|r-(w+1)| \leq \delta} \{V(r)\}$ ($w + 1$ to convert from the peptide mass to the y-ion mass). Once this is found, the peptide sequence can be obtained by backtracking.

Let

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

be the sum of all peaks whose mass is near r . Then,

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

The value of $V(r)$ is computed for r ranging from 0 to $w + 1 + \delta$.

Table 12.5.3 shows the computation for a peptide of mass $w = 233.21$.

12.5.3.1 Time complexity

Each entry in the $V(r)$ must be filled, in time proportional to $O(|A|)$. There are $O(w)$ entries with constant step size. Time complexity is hence $O(|A|w)$.

12.5.4 Scoring by considering y- and b-ions

The simple scoring scheme and dynamic programming (DP) algorithm described earlier takes only y-ions into account. However, fragmentation of a peptide pro-

Table 12.3: Example run for the simple dynamic programming algorithm

when $r = 0.00$,	$V(0.00) = 0$
when $r = 0.01$,	$V(0.01) = \dots$
\vdots	\vdots
when $r = 76.05$,	$V(76.05) = \max\{V(r - w(A)), \dots, V(r - w(G)), \dots\}$ $+ f_M(76.05)$ $= V(19) + 210 \text{ (we take } V(r) = 0, \forall r < \text{lightest y-ion})}$ $= 0 + 210 = 210$ <p style="margin-left: 2em;">(G was the residue that gave this maximum)</p>
\vdots	\vdots
when $r = 147.13$,	$V(147.13) = \max\{V(r - w(A)), \dots, V(r - w(Y))\}$ $+ f_M(147.13)$ $= \max\{V(147.13 - 71.08), \dots, V(147.13 - 163.18)\}$ $+ f_M(147.13)$ $= V(76.05) + 405 = 210 + 405 = 615$ <p style="margin-left: 2em;">(A was the residue that gave this maximum)</p>
\vdots	\vdots
when $r = 234.21$,	$V(234.21) = \max\{V(r - w(A)), \dots, V(r - w(S)), \dots\}$ $+ f_M(234.21)$ $= \max\{V(234.21 - 71.08), \dots, V(234.21 - 87.08)\}$ $+ f_M(234.21)$ $= V(147.13) + 0 = 615 + 0 = 615$ <p style="margin-left: 2em;">(S was the residue that gave this maximum)</p>

duces pairs of ions, i.e. y- and b-ions. Considering both these ion types will improve the model of the theoretical MS spectrum.

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 with masses $y_{1, \dots, k}$ and b-ions with masses $b_{1, \dots, k}$ of P . Hence the score function is redefined as

$$\text{score}(M, P) = \sum_i \{h | (x, h) \in M, |x - y_i| \leq \delta \text{ or } |x - b_i| \leq \delta\} \quad (12.3)$$

Figure 12.8 shows the theoretical spectrum considered for the peptide sequence SAG . The b-ions are drawn lower than the y-ion peaks to show their relative scarcity (although their weighting in the score function is equal).

As shown in Figure 12.9, matching the hypothetical peaks in Figure 12.8 to the real mass spectrum M , will produce a similarity score, $\text{score}(M, P) = f_M(y_1) + f_M(y_2) + f_M(b_1) + f_M(b_3) = 210 + 405 + 150 + 160 = 925$. If this similarity score is the highest among all possible peptides, we deduce that $P = SAG$ is the

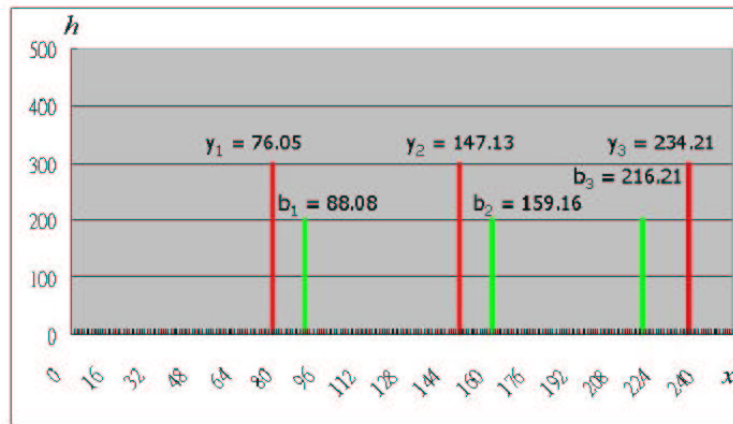


Figure 12.8: Theoretical (ideal) spectrum of a hypothetical peptide, showing peaks for y-ions (in red) and b-ions (in green)

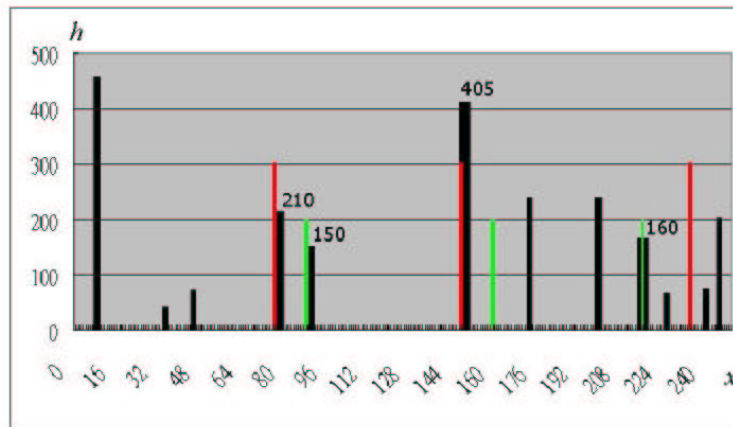


Figure 12.9: Real spectrum M with ideal peaks of y-ions and b-ions superimposed onto it (Real peaks are in black and numbers are the intensities of the black peaks.)

optimal peptide sequence for M .

12.5.5 Why previous DP algorithm cannot be used

The previous DP algorithm cannot be used here 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 avoids double counting [MZL03]. This way, each peak will serve as only one piece of supporting evidence.

For example, in Figure 12.10, a peak may be matched by two ions — a b-ion

with mass b_i and a y-ion with mass y_j . Such a peak should only be summed at most once (either in $f_M(b_i)$ or in $f_M(y_j)$).

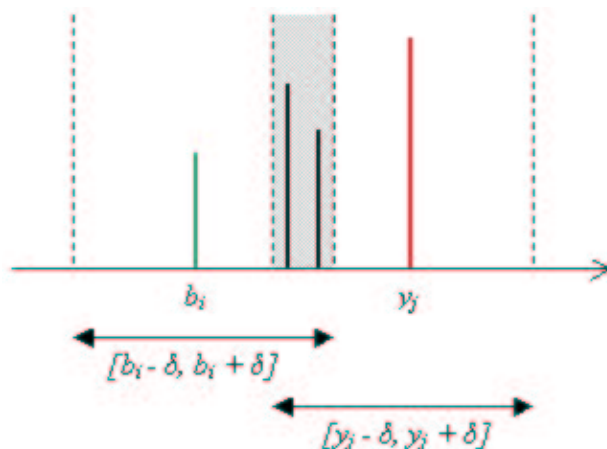


Figure 12.10: Peaks that are matched by two ions with masses b_i and y_j should be summed only once.

12.5.6 Observation

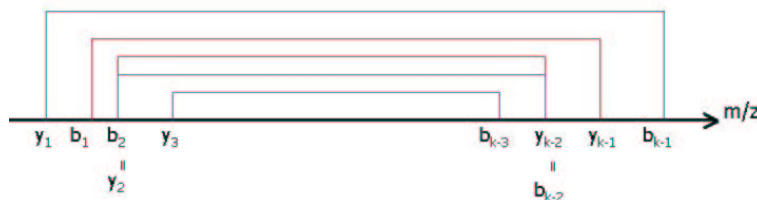


Figure 12.11: Complementary y-ions and b-ions form a set of nested regions.

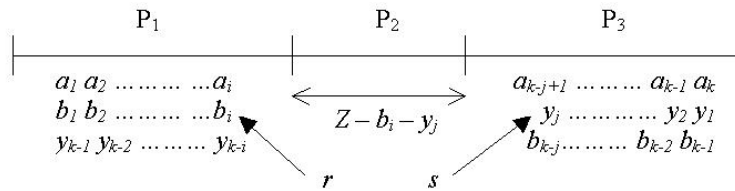
The b-ion and y-ion due to the same breakpoint have complementary masses, i.e. they sum to $wt + 20$. Figure 12.11 shows the nested regions with same mid-point formed by such pairs. Therefore it suffices to consider

1. b-ions to the left of (or at) the mid-point, and
2. y-ions to the right of the mid-point

We use this in DP solution below.

12.5.7 A Modified Dynamic Programming Solution

For simplicity, we assume that b-ions and y-ions did not attract additional atoms (i.e. ion masses are complementary if they sum to $Z = wt$, not $wt + 20$).

Figure 12.12: Peptide $P = P_1P_2P_3$

Call an ion *generated* by P_i (P_i is a substring of P), if the breakpoint is within or on the border of P_i .

Let's parameterize scoring function $score(M, P)$, so that $score'(M, P_1, P_2, P_3)$, where $P = P_1P_2P_3$, now measures $score(M, P)$ using only ions generated by P_2 . That is, ignoring the ions due to breakpoints within P_1, P_3 . Denote empty string as ϵ . Then $score'(M, \epsilon, P, \epsilon)$ is just $score(M, P)$.

Define $V(r, s)$ to be the maximum $score'(M, P_1, P_2, P_3)$, among all P_1, P_2, P_3 satisfying $w(P_1) = r, w(P_3) = s, w(P_1P_2P_3) = Z$. We need to find the minimum of all $V((Z - a)/2, (Z - a)/2)$ with $a \in A$, that is, over all possible "middle" amino acids. V satisfies the following recursive relation:

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

$f_M(x, y) =$ sum of intensities near x and $Z - x$, but not near y and $Z - y$

with base case

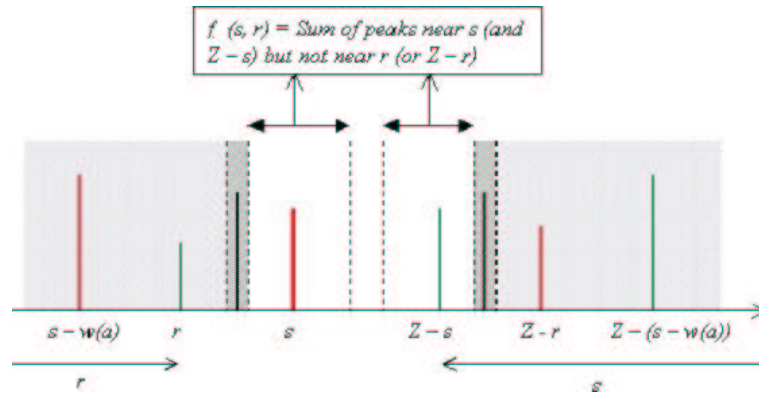
$$V(r, s) = 0 (r \leq 0, s \leq 0)$$

(Intuitively $f_M(x, y)$ gives y higher priority: if a peak is near both, x will not get it)

12.5.7.1 An informal proof of correctness

Let's consider the case $r < s$. Let P_1, P_2, P_3 maximize $score'$ for $V(r, s)$. Let R be the b-ion due to breakpoint between P_1, P_2 , and S be the y-ion due to breakpoint between P_2, P_3 . The next y-ion to the right of S must have one of these masses: $s - w(a) : a \in A$. Also, $V(r, s - w(a))$ is the optimal score measured using only ions generated by P_2a . For any possible (P_1, P_2, P_3) , the difference between $score'(M, P_1, P_2, P_3)$ and $score'(M, P_1, P_2a, \vec{P}_3)$ ($a\vec{P}_3 = P_3$), for any $a \in A$, is indeed caused by S .

Say (r, s) overlaps (u, v) if the peaks counted by $f_M(r, s)$ and $f_M(u, v)$ are not disjoint. To show each peak is counted at most once is to show that the

Figure 12.13: Calculation of $V(r, s)$ when $r > s$

difference caused by S is indeed $f_M(s, r)$. In other words, we want to show that (s, r) does not overlap $(\max\{u, v\}, \min\{u, v\})$ for all $V(u, v)$ whose computation involves $V(s, r)$ — that is, when $u \geq r, v \geq s$ and not both $u = r, v = s$. Note that $\min\{a : a \in A\} = 57.05$. There are two cases:

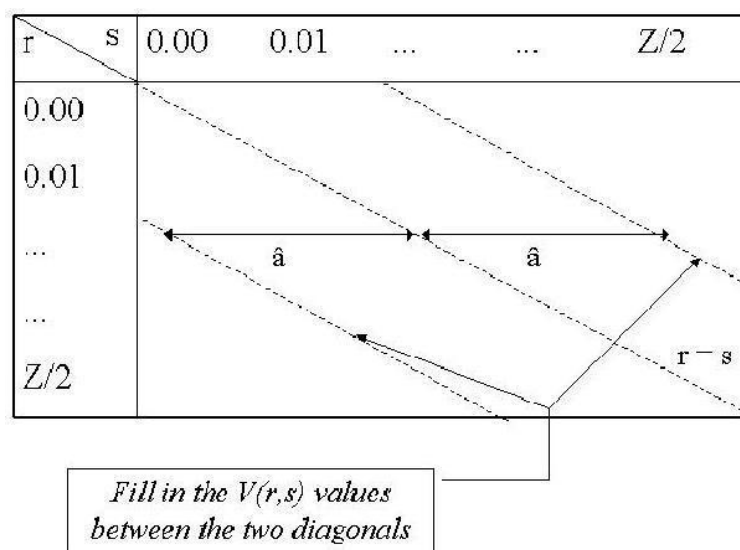
1. $v = s$. In this case $f_M(u, s)$ will be called (by $V(u, s)$), and (u, s) of course does not overlap (s, r) for all r, u by definition of f .
2. $v \geq s + 57.05$. If $u = r < s < v$, $f_M(v, r)$ will be called, and (s, r) does not overlap (v, r) since $v \geq s + 57.05$. If $u > r$, either $f_M(u, v)$ or $f_M(v, u)$ may be called:
 - (a) $u < v$ and $f_M(v, u)$ is called. No overlap since $v \geq s + 57.05$.
 - (b) $u > v$ and $f_M(u, v)$ is called. No overlap since $u > v \geq s + 57.05$.
 - (c) $u = v$ and it goes to one of the above cases

Finally, since $f_M(s, r)$ is independent of choice of a , maximizing $\text{score}'(M, P_1, P_2, P_3)$ may only be achieved by maximizing $\text{score}'(M, P_1, P_2, \vec{P}_3)$ for the best a .

A similar argument of correctness applies to the case $r > s$. With $r = s$, the optimal score is the larger of the scores of two choices. This case introduces no double-counting as it does not invoke f itself.

12.5.7.2 Time complexity

By opting to always decrease the larger of r, s , we must have $|r - s| \leq$ the maximal decrement, i.e. $\hat{a} = \max\{a\}$. Thus we can limit domain of V to $\{(r, s) : |r - s| \leq \hat{a}\}$, i.e. we need to compute $\hat{a}w$ entries. Each entry can be computed in $O(|A|)$ (constant) time. The time complexity of this algorithm is $O(|A|\hat{a}w)$.

Figure 12.14: Table of $V(r, s)$ values

12.5.7.3 An Example for the Modified DP Algorithm

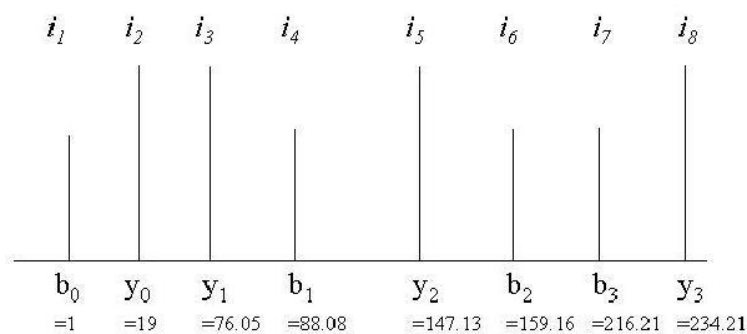
Figure 12.15: Peaks of hypothetical b-ions and y-ions of peptide $P = SAG$. i_1 to i_8 are the intensities (heights) of these peaks in M .

Table 12.4 illustrates how the $V(r, s)$ values are calculated starting from $V(0, 0)$ until $V(Z/2, Z/2)$. 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 all the 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 $P_1aP_3 = SAG$.

$$V(0, 0) = 0 \text{ (base case)}$$

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

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

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

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

Table 12.4: Example run of the modified DP algorithm

12.5.8 Spectrum graph methods [D99]

Another method of viewing the *de novo* sequencing problem is to reformulate it as a graph problem.

12.5.8.1 Spectrum graph

A spectrum graph $G_\Delta(S)$ of an MS/MS spectrum S with ion-types $\Delta = \{\delta_i\}$ (S contains modifications of fragmented peptides P' with mass $m(P') - \delta_i$) is a directed acyclic graph.

The vertices of $G_\Delta(S)$ are numbers representing the potential fragment masses. Every peak (x, h) in S generates vertices $V(x) = \{x + \delta_1, x + \delta_2, \dots\}$. The set of vertices in the spectrum graph are then $\{v_{\text{start}} = 0\} \cup \{v_{\text{end}} = m\} \cup V(x_1) \cup V(x_2) \cup \dots$. Two vertices u, v are connected by an edge from u to v if $v - u$ is the mass of some amino acid $a \in A$ and the edge is labelled by a .

If the mass spectrometer is infinitely accurate and the fragmentation process produces fragment ions corresponding to every possible fragmentation position, then there will exist some path from v_{start} to v_{end} which corresponds to the entire peptide. However, in reality, neither of the 2 assumptions hold. To account for the mass spectrometer's accuracy δ , vertices with masses closer than δ may be merged (the new mass being the weighted average), and an edge from u to v may be added if $|v - u - m(a)| < \delta$. To account for incomplete fragmentation, gap

edges that model di- and tripeptides spanning the unrepresented fragments may be added.

12.5.8.2 Path scoring

Because there will be multiple paths from v_{start} to v_{end} in the spectrum graph, it is necessary to assign a score to each path in order to select one path (and hence, one corresponding peptide sequence) which accounts for the spectrum well. The ideal scoring function will be one which reports $Pr(S|P)$, the conditional probability that the spectrum S is produced by the peptide P , which is another way of describing the model of the mass spectrometer operation as described in Section 12.3.

A simple scoring function is given below; more elaborate and accurate models will naturally result in better results.

In the simple scoring method, each ion type δ_i is assumed to be produced independently with probability $p(\delta_i)$. Additionally, the spectrometer is assumed to produce noise uniformly randomly with probability q_R . Therefore, $Pr(S|P) = \prod_t Pr(s_t|P)$, where $Pr(s_t|P)$ is the probability that the presence or absence of a peak in S at position t results if the peptide is P .

For a position representing an ion-type δ_i of some partial peptide of P ,

$$Pr(s_t|P) = \begin{cases} q_j & \text{if } s_t = 1 \text{ (there is a peak at position } t\text{),} \\ 1 - q_j & \text{otherwise} \end{cases}$$

For any other position,

$$Pr(s_t|P) = \begin{cases} q_R & \text{if } s_t = 1, \\ 1 - q_R & \text{otherwise} \end{cases}$$

More accurate scoring methods would use $Pr(h_t|P)$, accounting for the intensity of the MS at position t , and would probably model the ion production probability more intricately (not a simple independent assumption).

12.5.8.3 Sequencing

With the scoring function assigned the sequencing problem can be cast as a longest path problem in a weighted directed acyclic graph. Standard methods can then be used to solve this problem.

However, due to the presence of both C- and N-terminal ions in the spectrum graph, each C-terminal ion is accompanied by a corresponding 'fake' N-terminal ion which should not be included in the path. This observation transforms the problem to an antisymmetric longest path problem in a proper graph.

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.
- [D99] DANČÍK, V., ADDONA, T.A., CLAUSER, K.R., VATH, J.E. and Pevzner, P.A., “*De Novo* Peptide Sequencing via Tandem Mass Spectrometry.”, *Journal of Computational Biology*, 6(3,4): 327–342, 1999.