Identification of Peptides

References:

- Perkins et al.: Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis, 1999, 20, 3551-3567

- MASCOT web page at www.matrixscience.com.

- R.G.Sadygov et al. Large Scale database searching using tandem mass spectra: Looking up the answer in the back of the book, Nature Methods,2004,1,3,pp 195-202

- Bafna, Edwards: SCOPE, a probabilistic model for scoring tandem mass spectra against a peptide database, Bioinformatics, 2001, 17, 1, 13-21

Identification of Peptides (2)

We will speak about two ways to identify (determine the sequence) of peptides using mass spectrometry:

1. Peptide mass fingerprinting (PMF) in MS spectra

2. Peptide identification using MS/MS spectra (also called MS^2)

In PMF the proteins are *digested* using an restriction enzyme like Trypsin. The digestion is so specific that it is often possible to identify the protein from the list of MS feature masses alone. But this is clearly *not feasible* for complex mixtures.

In MS/MS a peptide is *further fragmented* using for example CID (collision induced dissociation). If we are lucky, the peptide breaks once after each amino acid, so we can determine its sequence from the list of masses in the MS/MS spectrum.

Identification of Peptides (3)

Although the data derived from PMF and MS/MS experiments has slightly different characteristics, the general approach for using it is similar.

The experimental data are compared with calculated peptide mass or fragment ion mass values, obtained by applying appropriate cleavage rules in the sequence database. Corresponding mass values (and sometimes intensities) are counted or scored in a way that allows the peptide which matches the data best to be identified.

Identification of Peptides (4)

Algorithmically there are two interesting problems:

1. How do we score the data against the theoretical spectrum and how significant is the score?

2. How do we quickly generate theoretical candidate spectra from large protein or transcript databases?

We will give two different answers to the first question. First we describe the MOWSE score, used in the popular MASCOT package (www.matrixscience.com).

Furthermore, we will give a general description of MS/MS based identification and introduce the SCOPE algorithm, developed at Celera Genomics (see reference 3).

Peptide Mass Fingerprinting

The input is here a list of masses of the tryptic peptides. For example for human albumin the list of masses contains 49 masses:

```
2917.322 2593.242 2433.263 2404.170 2203.001 2045.095 1915.773
1853.910 1742.894 1623.787 1600.731 1511.842 1386.620 1384.535
1381.533 1342.634 1320.490 1311.741 1257.523 1191.574 1149.615
1024.455 1018.477 2917.322 2593.242 2433.263 2404.170 2203.001
2045.095 1915.773 1853.910 1742.894 1623.787 1600.731 1511.842
1386.620 1384.535 1381.533 1342.634 1320.490 1311.741 1257.523
1191.574 1149.615 1024.455 1018.477 2917.322 2593.242 2433.263
2404.170 2203.001 2045.095 1915.773 1853.910 1742.894 1623.787
1600.731 1511.842 1386.620 1384.535 1381.533 1342.634 1320.490
1311.741 1257.523 1191.574 1149.615 1024.455 1018.477 1017.536
1013.598 1013.424 1000.603  984.488  960.562  951.441  940.448
etc....
```

Peptide Mass Fingerprinting (2)

Of course proteolysis is not always complete. Steric hindrance, the local context of the cleavage site, or simply insufficient enzyme concentration might lead to one or more *miscleavages*, that means two peptides that should be digested are still together.

For example, if we allow for two miscleavages in the albumin example, the mass list contains 204 masses.

How would MASCOT score the peaks list? It first computes a MOWSE score and then the probability p that this MOWSE score was achieved by chance. Then the probability p is converted into the MASCOT score as $-10 \log p$. So the lower the probability, the higher the MASCOT score.

Peptide Mass Fingerprinting (3)

MOWSE compares the calculated peptide masses for each entry in the sequence database with the set of experimental data. Each calculated value which falls within a given mass tolerance of an experimental value counts as a match.

Rather than just counting the number of matching peptides, MOWSE uses empirically determined factors to assign a *statistical weight* to each individual peptide match. The matrix of weighting factors is calculated during the database build stage, as follows:

Peptide Mass Fingerprinting (4)

A *frequency factor matrix*, F, is created, in which each row represents an interval of 100 Da in peptide mass, and each column an interval of 10 kDa in intact protein mass. As each sequence entry is processed, the appropriate matrix elements $f_{i,j}$ are incremented so as to accumulate statistics on the size distribution of peptide masses as a function of protein mass. The elements of F are then normalized by dividing the elements of each 10 kDa column by the largest value in that column to give the MOWSE factor matrix (i: peptide, j: protein)

$$M = (m_{i,j}) = \left(\frac{f_{i,j}}{\max_i f_{i,j}} \right)$$

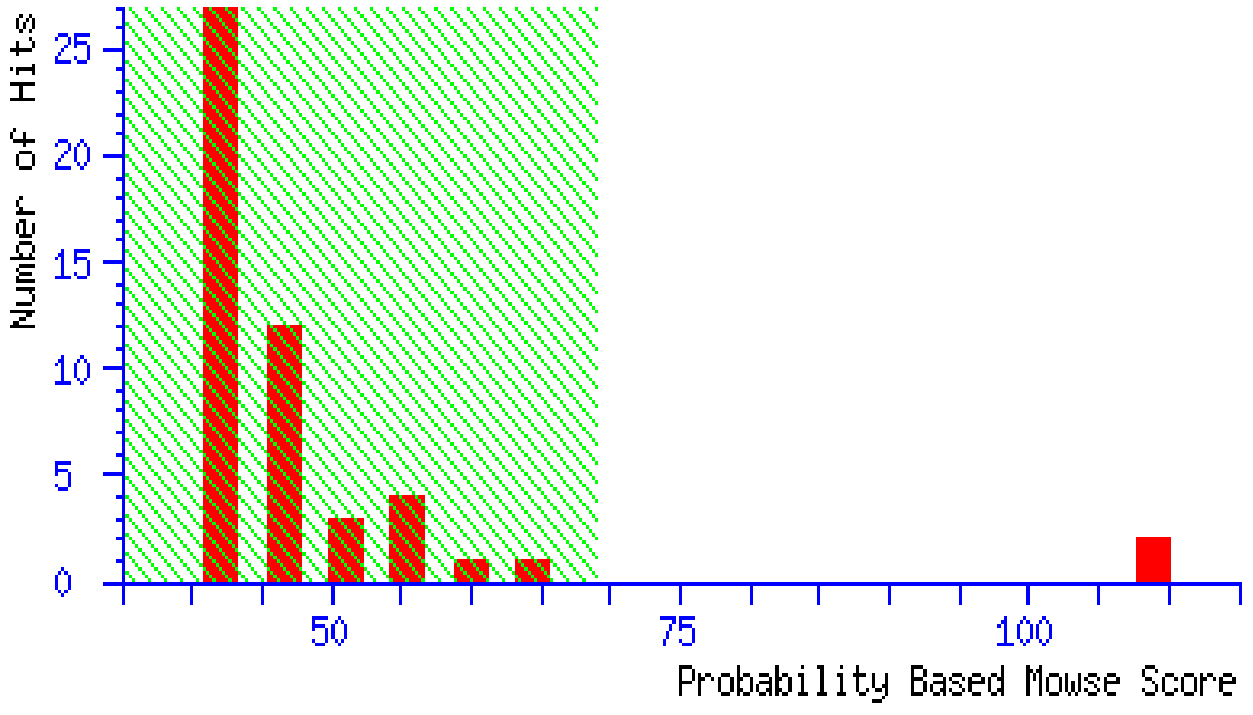
Peptide Mass Fingerprinting (5)

After scanning the experimental mass values against a calculated peptide mass database, the score for each database entry is calculated according to:

$$score = \frac{50000}{M_{prot} \prod_{\substack{(i,j) \\ \text{of match}}} m_{i,j}},$$

where M_{prot} is the molecular weight of the entry and the product term is calculated from the MOWSE factor elements for each of the n matches between the experimental data and peptide masses calculated from the entry. Finally this score is turned into a probability as mentioned before, and its significance is computed.

Idea: Matches of peptide masses that occur more frequently for a protein of size M_{prot} receive a higher weight.



Scores in the green, shaded region are not significant (at a level of p = 5%).

Peptide Mass Fingerprinting (6)

As mentioned before, the measured stick spectrum is compared against the theoretical spectrum derived from the sequence database.

Unfortunately is is usually not sufficient (both in PMF and MS/MS) only to consider the theoretical digest, even when considering miscleavages. What is generally worse is that the amino acids can occur in *modified form*.

Peptide Mass Fingerprinting (7)

There are the *natural* posttranslational modifications, such as phosphorylation and glycosylation. There are the *accidental* modifications which are artefacts of sample handling, such as oxidation. Finally, there are the modifications *deliberately introduced* during sample work-up, such as cysteine derivatisation.

Hence, it is usually not known beforehand which modifications occur. MASCOT models two modifications, *fixed* and *variable* modifications.

Peptide Mass Fingerprinting (8)

Fixed modifications come at no cost, since the molecular weight of an amino acid is just replaced by its modified weight.

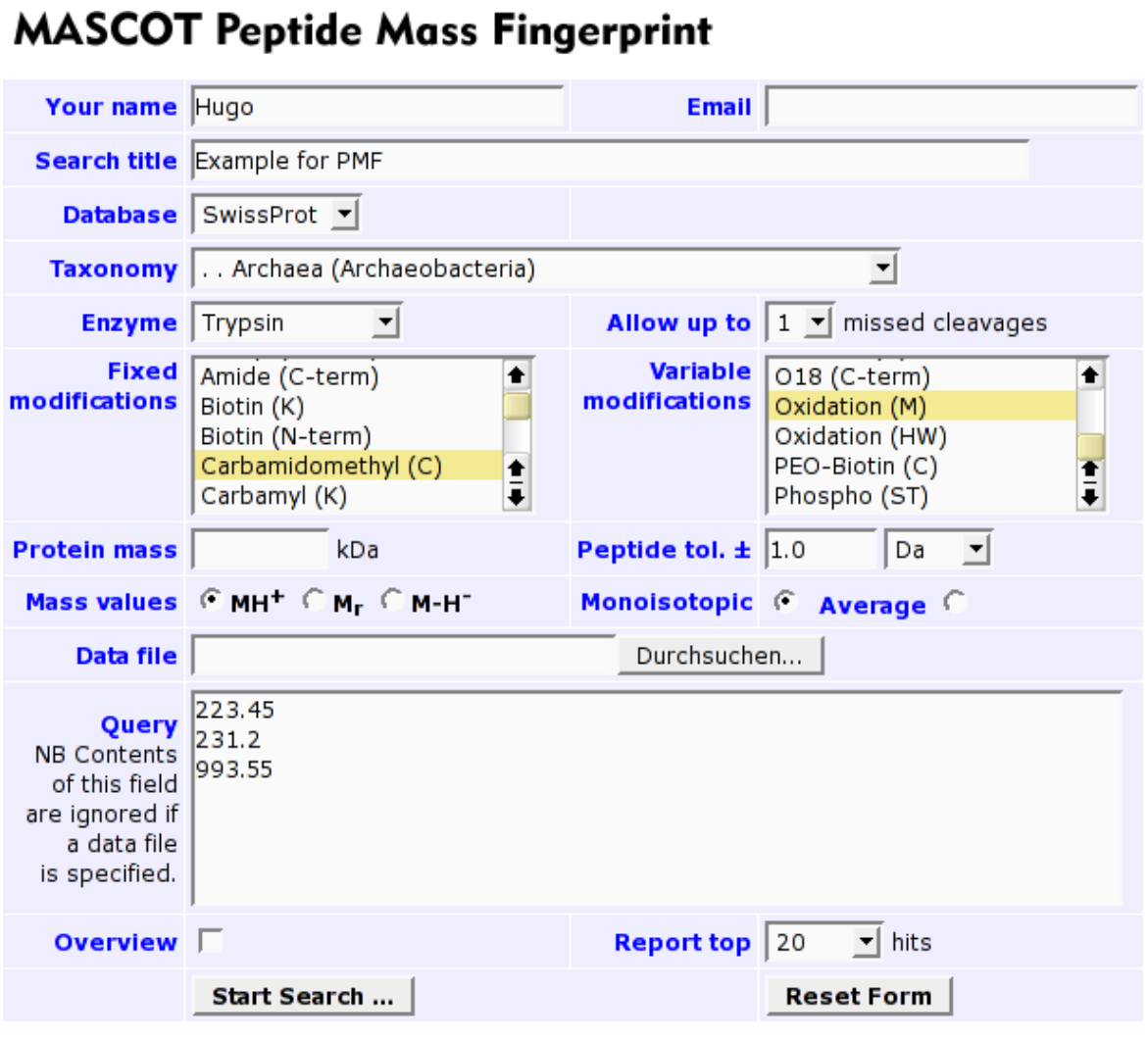
Variable modifications are those which may or may not be present. MASCOT tests all possible arrangements of variable modifications to find the best match. For example, if Oxidation (M) is selected, and a peptide contains 3 methionines, Mascot will test for a match with the experimental data for that peptide containing 0, 1, 2, or 3 oxidised methionine residues. This greatly increases the complexity of a search, resulting in longer search times and reduced specificity, so variable modifications should be used sparingly. (There are hundreds of modifications known. A database of such modifications is *Delta Mass*).

Peptide Mass Fingerprinting (9)

Finally, *noise peaks* coming from imperfect data processing and chemical noise (contaminants), make the identification difficult. The most common contaminants are keratin (hair, skin, dandruff).

MASCOT search form

Screenshot taken from http://www.matrixscience.com/



Peptide identification from MS/MS spectra

Peptide Mass Fingerprinting (PMF) is a straightforward procedure. But it is clearly unfeasible for complex mixtures. The workhorse for peptide sequencing is MS/MS based identification.

MASCOT can also handle MS/MS spectra and contains an algorithm to identify peptides from these spectra. But this algorithm is commercial and its details are not published.

We will therefore outline another algorithm, called *Sequest*, which is based on an academic algorithm and present a scoring function of MS/MS identification, called *SCOPE*.

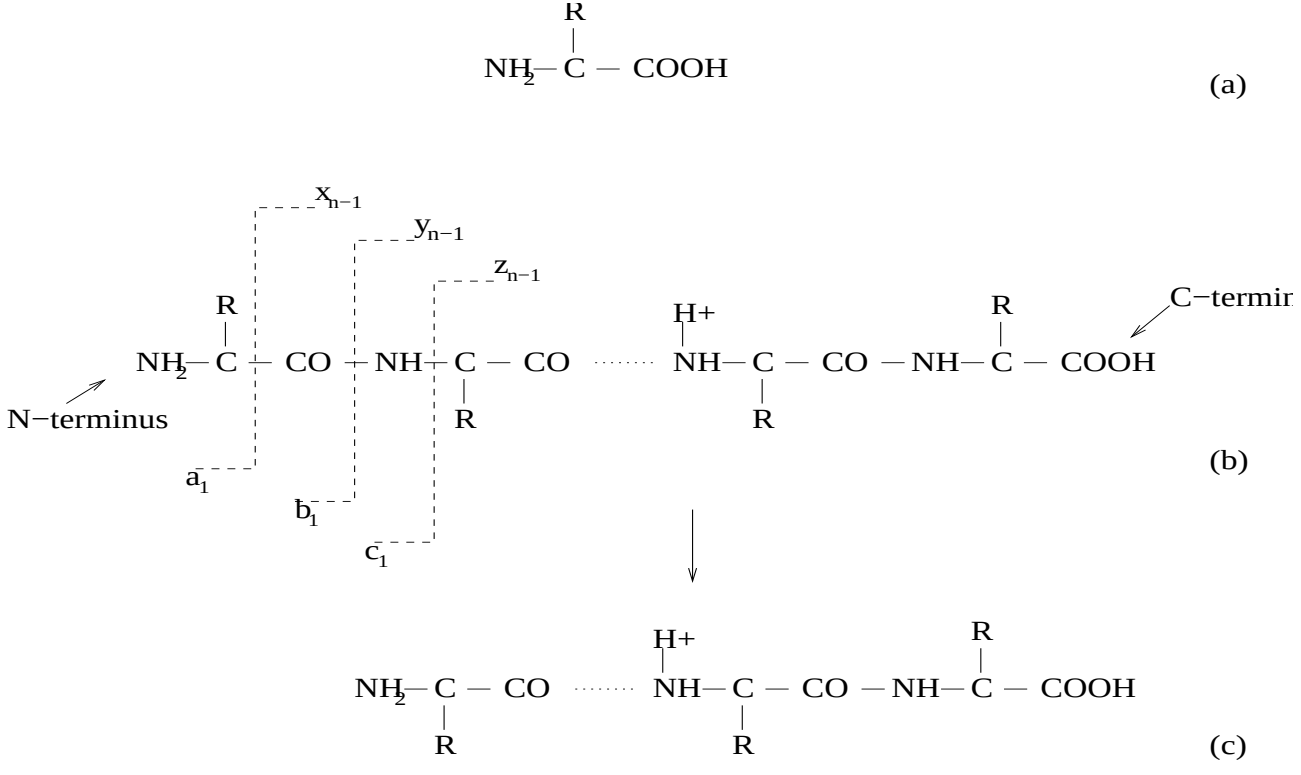
Tandem Mass Spectrometry

Recall again the structure of a peptide chain, consisting of different amino acids joint by *peptide bonds*. Amino-acids are distinguished from each other by the secondary structure of the side chain R.

In tandem mass spectrometry (MS/MS) ionized peptides are fragmented by *collision-induced dissociation* (CID). Fragments retaining the ionizing charge after CID have their mass-to-charge ratio measured. Since peptides typically break a peptide-bond when they fragment by CID, the resulting spectrum contains information about the constituent amino-acids of the peptide.

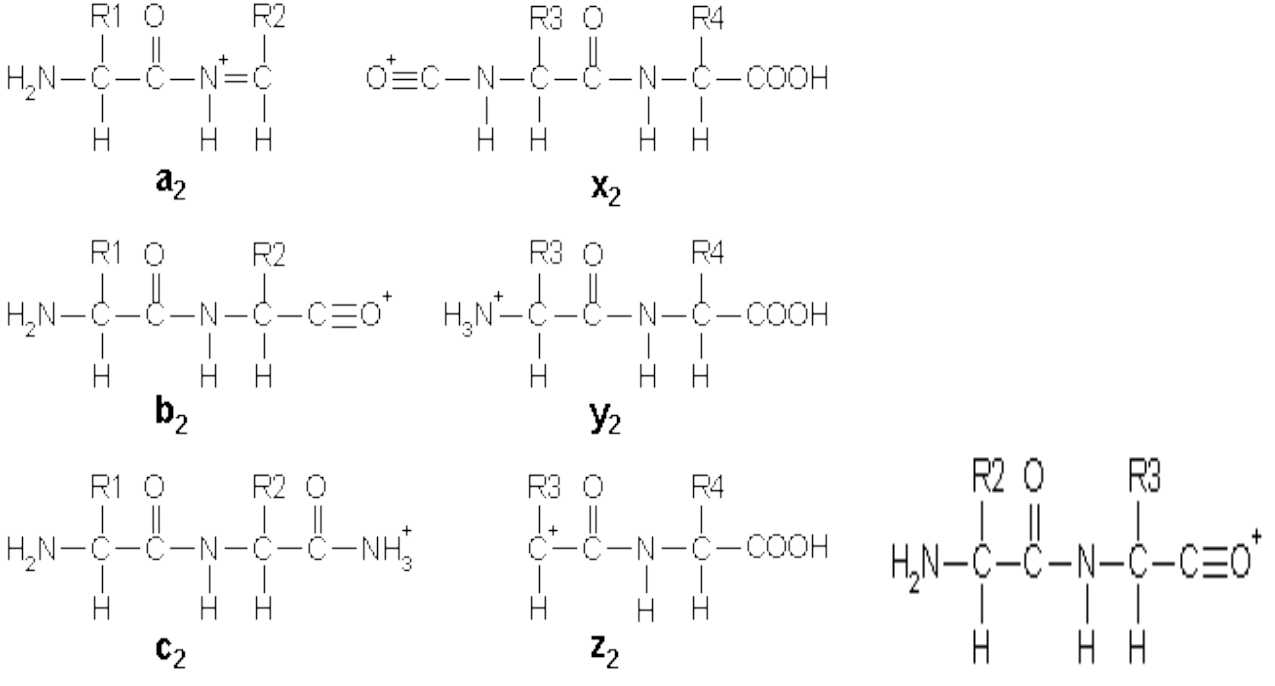
Ion types

The fragmentation of the peptide in CID is a stochastic process governed by the physiochemical properties of the peptide and the energy of collision. The charged fragment can be inferred by the position of the broken bond and the side retaining the charge. In the figure below, the N-terminal a_1, b_1, c_1 fragments, and the C-terminal x_{n-1}, y_{n-1}, and z_{n-1} fragments are shown.



Ion types (2)

While *a*, *b*, *x* and *y* represent the commonly occurring fragments, a high energy collision often results in other fragments, including *internal* fragments formed by breakage at two points, and fragments formed by breaks in side-chains.

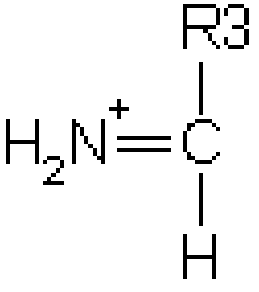


source: www.matrixscience.com

The above pictures show the *a*, *b*, *c*, *x*, *y*, *z* ions as well as an internal ion.

Ion types (3)

In high energy collision side chain cleavages can occur which helps distinguishing isomers like Leucin and Isoleucin. Another help can be the presence of *immonium* ions that represent a single amino acid (minus the loss of CO and the addition of H).

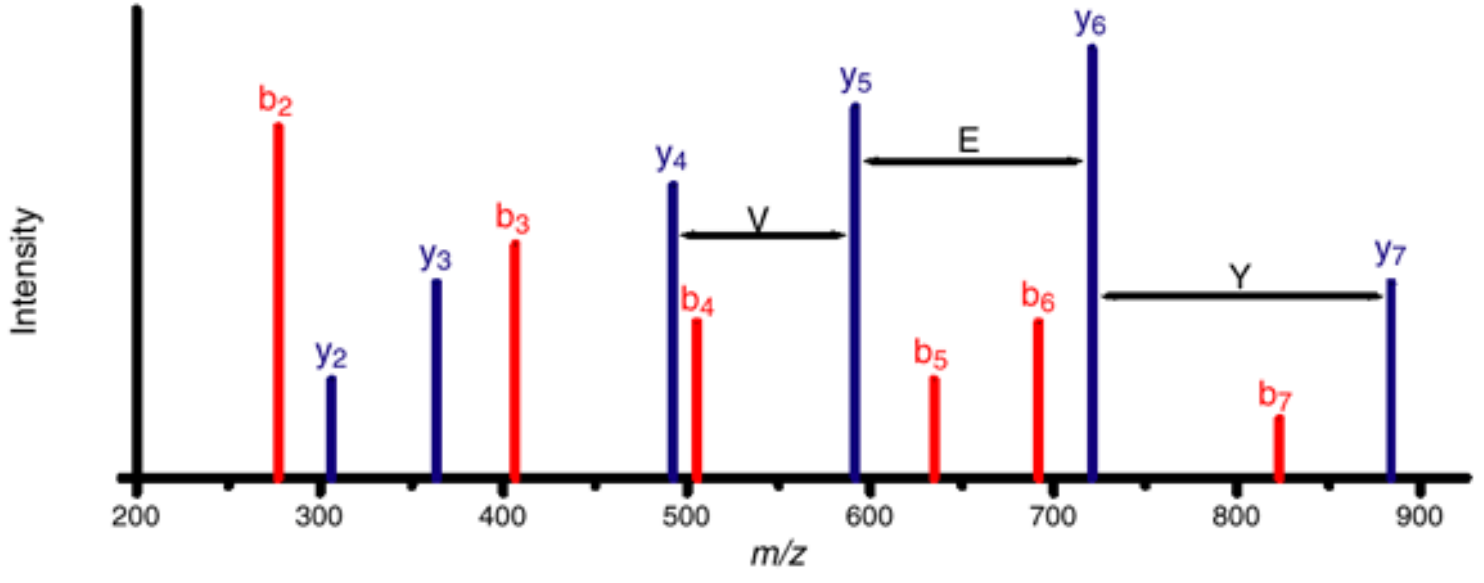


In any case, the *mass difference* between ions of the same type is the same and characteristic for amino acid.

MS/MS spectrum

Hence, if we measure the *MS/MS spectrum* of the peptide fragments and if we can identify the correct ions, we can read off the respective amino acids.

Below is a cartoon MS/MS spectrum for the peptide *IYEVEGMR*. The b-ion ladder is shown in red and the y-ion ladder in blue. Distances between peaks can be used to infer partial sequences of the peptide.



MS/MS spectrum (2)

A MS/MS spectrum is usually already pre-processed by the software of the MS instrument. We therefore deal only with *peak data* and not raw data.

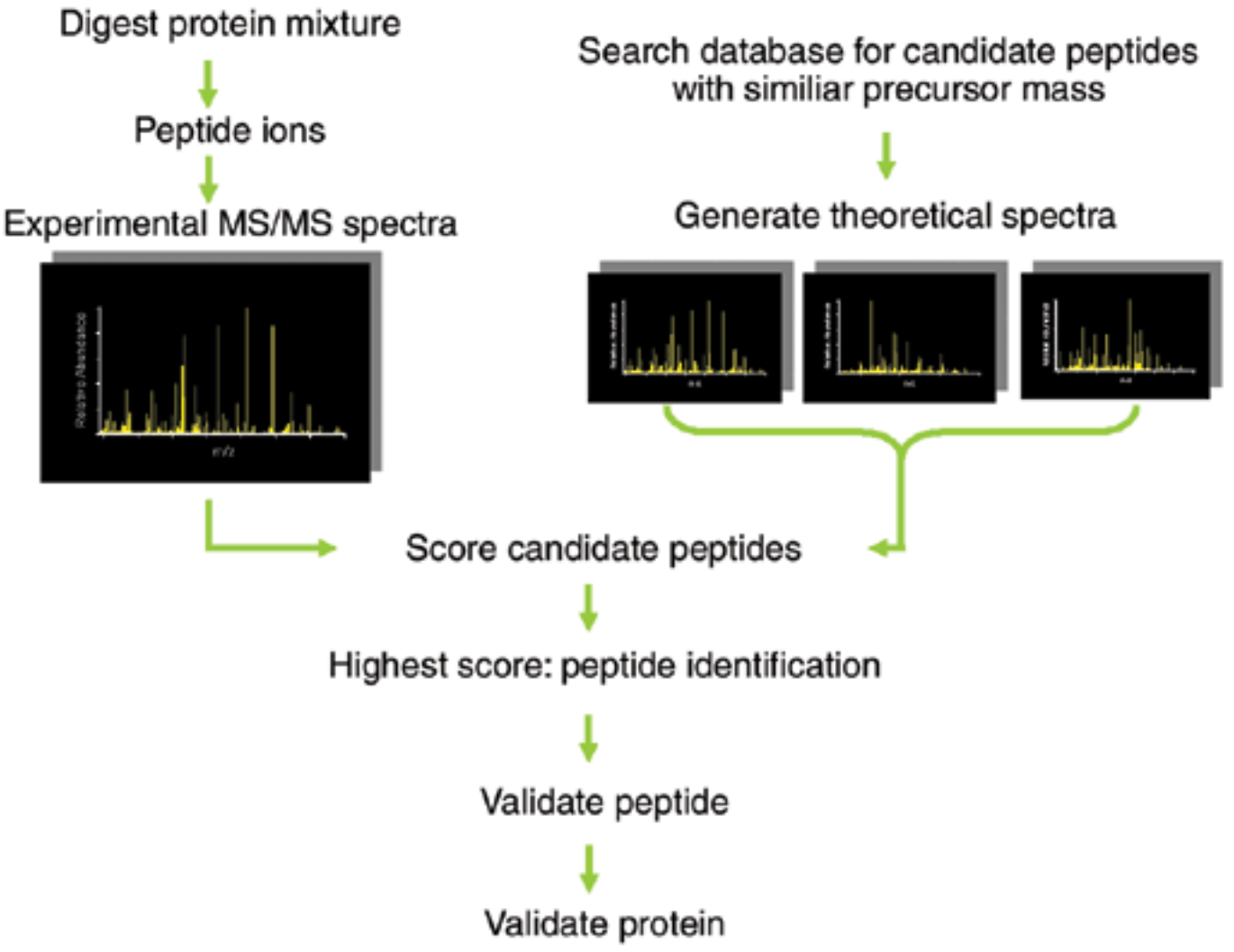
We can read off the peptide sequence from the *peak mass distances* in the MS/MS spectrum. So we are done?

MS/MS spectrum (3)

Not quite. Other possible fragments, the presence of multiply charged ions, the absence of some ions in a series, and finally noise and measurement error pose a real challenge. So we have to resort to a different approach.

Quiz: If we know the molecular weight of the peptide generating the MS/MS spectrum, we can dramatically reduce our search space. How could we estimate this *parent ion mass* from the MS/MS spectrum?

Overview of MS/MS-based identification Process



Modules of MS/MS algorithms

Most algorithms for analyzing MS/MS data address the following three modules:

Interpretation: The *input* is an *MS/MS spectrum*, the *output* is *interpreted-MS/MS-data*. Interpreted-MS/MS-data may include parent peptide mass, partial or complete sequence tags, and combinations of sequence tags and molecular masses.

Filtering: The *input* is *interpreted-MS/MS-data* and a peptide sequence database. The *output* is a list of *candidate-peptides* that might have generated the MS/MS spectrum.

Scoring: The *input* is a list of *candidate-peptides* and the *MS/MS spectrum*. The *output* is a ranking of the *candidate-peptides* along with a score and possibly a *p*-value (probability that the score was achieved by random chance).

A general MS/MS scoring schema

As an example, the first versions of the popular *Sequest* algorithm computed a *preliminary score* S_p :

$$S_p = \left(\sum_k I_k \right) m(1 + \beta)(1 + \rho)/L$$

where the first term is the sum of the intensities of all matched peaks, m is the number of matches, β is a reward for each *consecutive match* of an ion series, ρ is a reward for the presence of an *immonium ion* and L is the number of all theoretical ions of an amino acid sequence.

Only spectra with a S_p score passing a threshold a further examined.

Quiz: What is the disadvantage of this score? (Think of mutations or posttranslational modifications...)

A general MS/MS scoring schema

Spectra with a high S_p receive a second score, called *XCorr*. This is basically a modified version of the cross-correlation of the experimental MS/MS spectrum (E) and the theoretical spectrum (T):

$$Corr_\tau(E, T) = \sum_{i=0}^{N-1} x_i y_{i+\tau}$$

The autocorrelation is a measure of similarity and is computed for a range of shifts τ.

A general MS/MS scoring schema (2)

The XCorr score is dependent on peptide length and spectral quality. Newer versions try to correct for these dependencies by dividing by the auto-correlation of the experimental spectrum or similar measures.

Another important quantity reported by Sequest is the normalized difference of XCorr values, δC_n between the best scoring sequence and each of the other sequences.

SCOPE: Scoring of Tandem Mass Spectra

Sequest was developed at the University of Washington in Seattle, US. The algorithm was bought by Thermo Finnigan and Sequest is now widely used in the mass spectrometry community.

It works well, but is known to produce *many false-positive hits*. For that reason, many scientists try to improve parts of this algorithm and to develop more sophisticated methods to compare MS/MS spectra.

We will introduce *SCOPE*, a scoring function for MS/MS spectra which was developed by Vineet Bafna and Nathan Edwards at Celera Genomics (see reference 3).

SCOPE: Scoring of Tandem Mass Spectra

SCOPE is an scoring function (not a full algorithm) with several nice features. We will present its key concepts.

1. It models explicitly the fragmentation depending on the peptide and experimental setting,

2. models explicitly the measurement error

3. and models noise peaks.

SCOPE (2)

The SCOPE algorithm models the process of MS/MS spectrum generation by a two-step stochastic process.

1. The first step involves generation of fragments from a peptide, according to a probability distribution estimated from many training samples.

2. The second step involves the generation of a spectrum from the fragments according to the distribution of the instrument measurement error.

Definitions

We need to introduce some terminology.

MS/MS Spectrum: A MS/MS spectrum $S \in \mathbf{R}_+^k$ is a vector of positive real numbers specifying the k observed mass-charge ratios of the spectral peaks.

Peptide: A peptide $p \in \mathcal{A}^n$ is a sequence of n amino-acid residues over the alphabet of amino-acid symbols, $\mathcal{A} = \{A, C, ..., Y\}$.

Fragment Space: An enumeration $\mathcal{F}(p)$ of all fragment mass-charge ratios that a peptide p might produce. Each element of $\mathcal{F}(p)$, then, is a fragment-charge state pair. Thus,

$$\begin{aligned}\mathcal{F}(p) \quad = \quad \{&(a_1, i), (b_1, i), (y_1, i), \ldots, \\ &(a_n, i), (b_n, i), (y_n, i), i = 1, 2, 3, \ldots\}\end{aligned}$$

Denote the mass-charge ratio of a fragment $f \in \mathcal{F}(p)$ by $(m/z)(f)$.

Definitions (2)

Fragmentation Space: The fragmentation space $\phi(p)$ of a peptide p is the set of all fragmentation patterns of p. That is,

$$\phi(p) = \{F : F \subseteq \mathcal{F}(p)\}$$

Noise: We consider any peak of S for which $F(p)$ provides no explanation to be a noise peak.

Two step process

Measurement: Each fragment with a particular mass-charge ratio generates a mass-charge ratio observation close to, but not precisely at its true mass-charge ratio. The observation of many fragments with the same mass-charge ratio leads to the formation of a distinctive peak close to the true mass-charge ratio of these fragments.

The observed peak can then by represented by a single real number, an estimate of the true mass-charge ratio of the fragments that generated it. The deviation of this mass-charge ratio of a peak from its true value is modeled according to a probability distribution, typically the normal distribution.

Scoring spectra

Let $\psi(S \mid p)$ denote the probability density function for the random vector S representing the MS/MS spectrum, given peptide p. Typically, we are searching a database for the peptide p^* that satisfies

$$p^* = \arg\max_p \psi(S \mid p)$$

Scoring spectra

A formal description of the two-step model of fragmentation followed by measurement is given by:

$$\psi(S \mid p) = \sum_{F \subseteq \mathcal{F}(p)} \psi(S \mid F, p) \Pr(F \mid p)$$

The quantity $\Pr(F \mid p)$ represents the probability of a particular fragmentation pattern of a peptide. It is in the computation of $\Pr(F \mid p)$ that the complex process of fragmentation can be modeled.

Fragmentation probability estimation

The SCOPE algorithm does not explicitly implement an automatic algorithm to estimate the probabilities $Pr(F \mid p)$ but relies on the judgment of the user.

For example, experienced operators know that the presence of acidic amino-acids in a peptide makes the neutral water loss ion type cleavages much more likely.

In general those probabilities could also be learned from sample spectra of known peptides given a specific experimental setup.

"... we have chosen probabilities in consultation with experienced mass spectrometer operators." (V. Bafna and N. Edwards, inventors of SCOPE)

Computing $\psi(S \mid F, p)$

The probability density $\psi(S \mid F, p)$ describes the probability of observing a collection of spectral peaks, given a particular fragmentation pattern of a peptide p.

Unfortunately, it is not at all obvious which fragment(s) are responsible for which peak(s), and which peaks should be considered noise. In order to compute $\psi(S \mid F, p)$, we need to either sum over all the possible explanations of each peak (which is not feasible) or use our understanding of the mass spectrometer to limit the number of terms.

Computing $\psi(S \mid F, p)$ (2)

We assume the following:

1. Each unique mass-charge ratio in the fragment space generates at most one spectral peak.

2. Each spectral peak is the observed mass-charge ratio of at most one of the (unique) mass-charge ratios in the fragment space.

3. The assignment of spectral peaks to fragments must be *non-crossing*. For all fragments f_1, f_2 and spectral peaks S_1, S_2, if $(m/z)(f_1) < (m/z)(f_2)$ and $S_1 < S_2$, then peak S_1 must have been generated by fragment f_1 and peak S_2 must have been generated by fragment f_2.

Computing $\psi(S \mid F, p)$ (3)

In addition, we augment the fragment space $\mathcal{F}(p)$ with *noise fragments*, one for each spectral peak. Each noise fragment has the same mass-charge ratio as its spectral peak. We denote this augmented fragment space $\mathcal{F}'(p)$ and the corresponding fragmentation space $\phi'(p)$.

Due to the addition of noise fragments all spectral peaks must either be assigned to a unique fragment from our original fragment space $\mathcal{F}(p)$ or to a noise fragment. Therefore we can make the following observation:

Only fragmentation patterns $F \subseteq \mathcal{F}'(p)$ with $|F| = k$ have non-zero probability.

Computing $\psi(S \mid F, p)$ (4)

However, we can say something even stronger. Let $S_i \overset{M}{=} f$ denote the event that peak S_i is generated by fragment f, and $S = (S_1, S_2, \ldots, S_k)$ be a tandem MS spectrum ordered by mass-charge ratio. Further, let $F \subseteq \mathcal{F}'(p), \mid F \mid = k$ be an arbitrary fragmentation pattern, whose observed fragments $f_1, f_2, \ldots, f_k \in F$ are ordered by mass-charge ratio.

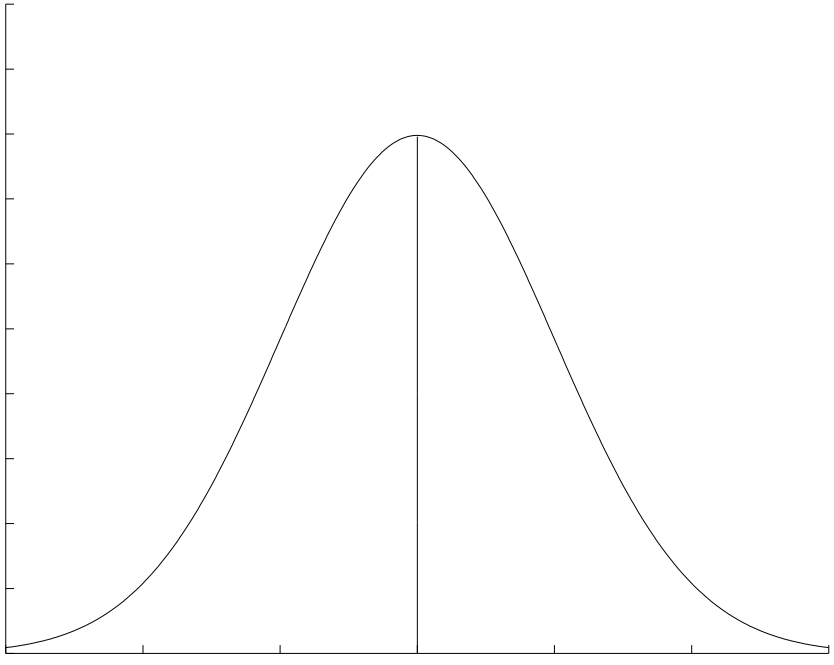
By the non-crossing and uniqueness assumptions, only one assignment of spectral peaks to fragments has non-zero probability mass. All of the probability mass for $\psi(S \mid F, p)$ is captured by this unique non-crossing assignment. We write:

$$\psi(S \mid F, p) = \psi(S \mid \cap_{i=1}^{k} [S_i \overset{M}{=} f_i], F, p)$$

In other terms: we aim for a *peaked* probability distribution, that assigns a high probability to well-matching peak assignments, but decreases rapidly if we move away from this assignment.

Computing $\psi(S \mid F, p)$ (5)

In isolation, the distribution of one measured mass-charge ratio about its true value is independent of any other measured mass-charge ratio about its true value. We model the distribution of the measured mass-charge ratios as normal distributions centered at the fragment mass-charge ratio and the distribution of the measured mass-charge ratio of noise fragments by an impulse function at the mass-charge ratio of its spectral peak.



Computing $\psi(S \mid F, p)$ (6)

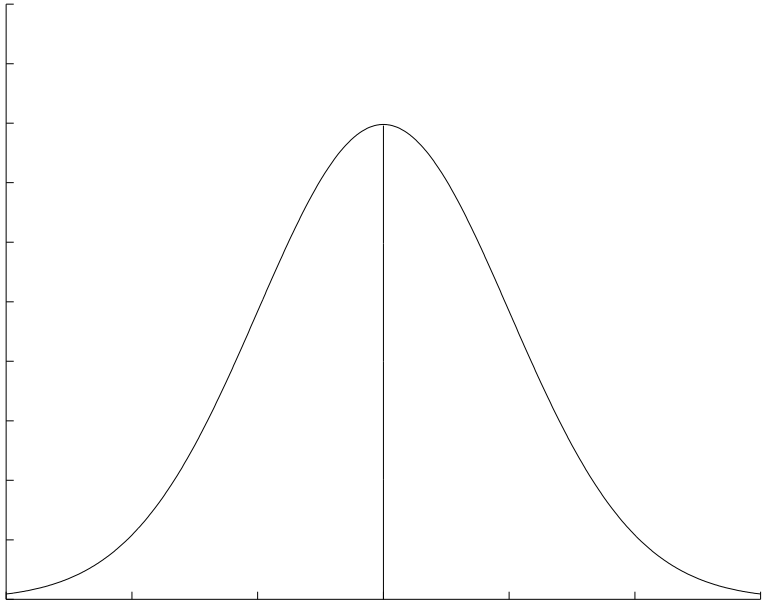
We expand $\psi(S \mid \cap_{i=1}^{k}[S_i \stackrel{M}{=} f_i], F, p)$ into its components in order to compute it.

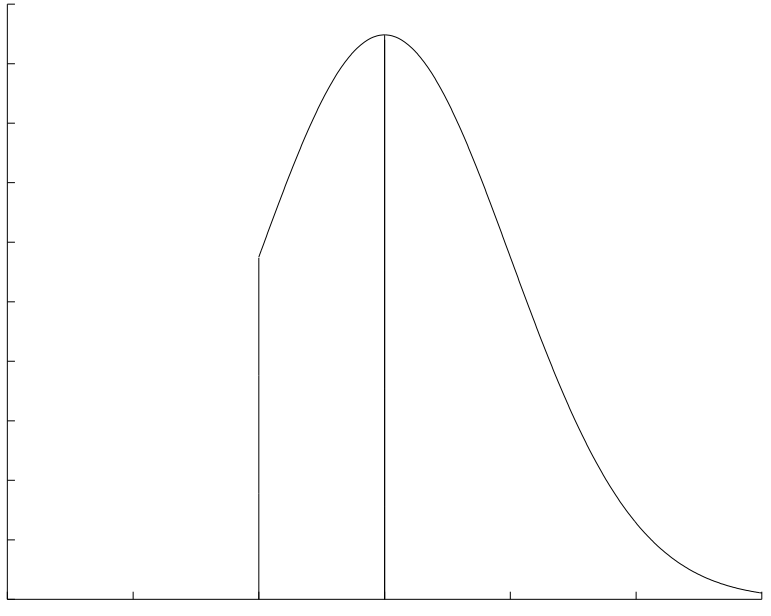
$$
\begin{aligned}
&\psi(S \mid \cap_{i=1}^{k}[S_i \stackrel{M}{=} f_i], F, p) \\
&\quad = \psi(S_1 \mid \cap_{i=1}^{k}[S_i \stackrel{M}{=} f_i], F, p) \times \\
&\qquad \Pi_{j=2}^{k}\psi(S_j \mid S_1, \ldots, S_{j-1}, \cap_{i=1}^{k}[S_i \stackrel{M}{=} f_i], F, p) \\
&\quad = \psi(S_1 \mid [S_1 \stackrel{M}{=} f_1], F, p) \times \\
&\qquad \Pi_{j=2}^{k}\psi(S_j \mid S_{j-1}, [S_j \stackrel{M}{=} f_j], F, p).
\end{aligned}
$$

In the last term S_j is only dependent on the previous S_i for $i < j$.

Computing $\psi(S \mid F, p)$ (7)

In the last term S_j depends only on the previous S_i for $i < j$. To simplify this we truncate the left-hand tail of the measurement distribution and rescale its total probability density to one.





Computing $\psi(S \mid F, p)$ (8)

Let ρ_f be the distribution of the observed peak about the true mass-charge ratio of fragment f. Then

$$\psi(S_1 \mid [S_1 \stackrel{M}{=} f_1], F, p) = \rho_{f_1}(S_1)$$

$$\psi(S_j \mid S_{j-1}, [S_j \stackrel{M}{=} f_j], F, p)$$

$$= \begin{cases} \dfrac{\rho_{f_j}(S_j)}{\int_{S>S_{j-1}} \rho_{f_j}(S)}, & S_j > S_{j-1}; \\ 0, & \text{otherwise.} \end{cases}$$

Computing $\psi(S \mid p)$

We now show how this choice of ψ allows an efficient algorithm for computing $\psi(S \mid p)$.

We want to avoid the computation of an exponential number of terms in the expression $\psi(S \mid p) = \sum_{F \subseteq \mathcal{F}(p)} \psi(S \mid F, p) \Pr(F \mid p)$. To do this we need another assumption, namely that the probability of observing f must be independent of the observation of other fragments. This is not always true but allows us to compute $\psi(S \mid p)$ efficiently by dynamic programming.

Given the spectrum $S = (S_1, \ldots, S_k)$ and the fragments $\mathcal{F}'(p) = \{f_1, \ldots, f_m\}$ ordered by mass-charge ratio, we define $\mathcal{F}'_j(p) = \{f_1, \ldots, f_j\}$ to be the first j fragments of $\mathcal{F}'(p)$.

Computing $\psi(S \mid p)$ (2)

The dynamic programming recurrence function $\Phi(i, j)$ represents the probability mass associated with the event that the first i peaks were generated by i fragments from the first j fragments of $\mathcal{F}'(p)$. Clearly, $\Phi(k, m) = \psi(S \mid p)$ is the value we are interested in. The following recurrence holds:

$$\Phi(i,j) = \begin{cases} 1, & \text{if } i = 0, \\ 0, & \text{if } i > j, \\ \Phi(i-1, j-1) & \\ \quad \times \psi(S_i \mid S_{i-1}, S_i \overset{M}{=} f_j) & \\ \quad \times \Pr(f_j \mid p) & \\ + \ \Phi(i, j-1) \ \Pr(\overline{f_j} \mid p), & \text{otherwise.} \end{cases}$$

Computing $\psi(S \mid p)$ (3)

The above recursion corresponds to a special sequence alignment problem. We align the spectrum with all possible fragments in the augmented fragment space.

The first term in the sum is for the case that the j-th fragment is assigned to a spectral peak. The probability of that is the probability of assigning the first $i-1$ spectral peaks to $i-1$ fragments among the first $j-1$ fragments [$\Phi(i-1, j-1)$] times the probability that S_i is assigned f_j [$\psi(S_i \mid S_{i-1}, S_i \overset{M}{=} f_j)$] times the probability of f_j given p [$\Pr(f_j \mid p)$].

The second term in the sum describes the probability that f_j is not assigned to any peak in S. This might be large if we do not expect a fragment to occur!

Computing $\psi(S \mid p)$ (4)

The most likely assignment F^* is given by

$$F^* = \arg \max_{F \subseteq \mathcal{F}'(p)} \psi(S \mid F, p) \Pr(F \mid p)$$

For aesthetic reasons, the score is reported in "$-\log(p)$" form.

Quiz: What is a potential problem of identification algorithms based on sequence database searching?

SCOPE: conclusions (5)

SCOPE has some interesting features (e.g. model of fragmentation process and measurement error), but it is mainly a theoretical concept and has not directly been used.

However, there are some algorithms such as InSpect or OLAV, which implement modified versions.

Even with sophisticated scoring functions such as SCOPE, most algorithms are known to produce large numbers of false positives and are unable to identify more difficult spectra (with e.g. low mass resolution, mutated peptides, post-translational modifications).

Summary

We covered:

- Peptide identification using Peptide Mass Fingerprinting (PMF) and Tandem Mass Spectrometry (MS/MS)
- A popular algorithm for PMF (MOWSE) and a scoring function for MS/MS (SCOPE).
- We did not cover: other MS/MS algorithms (X!Tandem, OMSSA), hybrid approaches (InSpect) and de-novo sequencing using MS/MS (PepNovo, Lutefisk).
- Many more questions: How to get from peptide sequence to protein sequence? How to estimate error rates for identifications? etc.