

Introduction to Computational Proteomics – Open problems

Jacques Colinge

*Research Center for Molecular Medicine of the Austrian Academy of
Sciences (CeMM), Lazarettgasse 19/3, A-1090 Vienna*

&

*Upper Austria University of Applied Sciences at Hagenberg
Hauptstraße 117, A-4232 Hagenberg*

jcolinge@excite.com

Outline

Proteomics has become an important approach to analyze biological samples. This tutorial will introduce the central problem of searching mass spectrometry data against a database. Quantitative proteomics and peptide *de novo* sequencing will be covered as well. This presentation should stimulate the interest of bioinformatics researchers in other fields and provide a concise introduction to life scientists.

Part 1 (20 min): Introduction to proteomics. We start by introducing the main problems in proteomics: identify proteins in a sample, characterize modified proteins, compare samples and quantify proteins. We point out the difficulty caused by excessively complex samples with high dynamic range of protein concentrations. We then rapidly introduce the concept of mass spectrometry as an analytical method.

Part 2 (30 min): Peptide mass fingerprinting (PMF) and MALDI instruments. On the basis of the general context presented in Part 1, we introduce and detail a first proteomics method. Show a first example with a spectrum and a database search result. Explain a basic algorithm for searching PMF data against a database of protein sequences. Introduce the notion of scoring function and present classical examples, e.g. MOWSE, ProFound, MSA and OLAV-PMF.

Part 3 (20 min): Peak detection. Raw spectrum processing is rapidly covered to actually link the somewhat abstract mass lists used for searching databases with the signal generated by the MS instruments.

Part 4 (60 min): Complex samples and tandem mass spectrometry. Database sizes and sample complexity may limit the usage of PMF. Tandem mass spectrometry is a manner to obtain additional information via fragmentation. Explain the principle of fragmentation.

Present a schematic abstract mass spectrometer with ion source, fragmentation cell and mass analyzer. Present different technologies (collision induced fragmentation, post-/in-source decay). Explain on-line mass spectrometry.

Several peptide scoring functions are reviewed: MASCOT, SEQUEST, post-processing of SEQUEST, OLAV-Phenyx. The problem of scoring protein identification is then discussed.

Part5 (40 min): Other problems, other approaches. We cover several problems which are of great importance in proteomics today: eukaryote genome searches, peptide de novo sequencing, differential proteomics via quantitative and semi-quantitative methods, protein characterization by top-down techniques.

Discussion (30 min).

CONTENTS

INTRODUCTION.....	5
A typical proteomics project.....	6
Protein separation techniques.....	7
PEPTIDE MASS FINGERPRINTING.....	11
Introduction.....	11
Searching a database.....	11
Scoring functions.....	14
RAW SPECTRUM PROCESSING.....	17
Introduction.....	17
MALDI-TOF.....	17
Peak detection.....	17
TANDEM MASS SPECTROMETRY.....	19
Introduction.....	19
On-line mass spectrometry.....	19
Fragmentation cell.....	19
Mass analyzer.....	19
The fragmentation spectrum.....	20
Modified peptides.....	22
Internal fragments.....	22
MS/MS database search.....	22
MS/MS scoring functions.....	23
OTHERS.....	25
De novo peptide sequencing.....	25
Direct genome searches.....	26
Differential proteomics.....	28
REFERENCES.....	29

INTRODUCTION

Proteomics is the complete analysis of proteins. Proteomics involve numerous technologies and address numerous questions concerning the proteins:

- What are the proteins contained in a biological sample?
- What are their concentrations?
- How their expression changes in various samples?
- What are their posttranslational modifications (PTMs)?
- How do they interact with other proteins or molecules?

In this lecture we concentrate on computational aspects of protein identification. Characterization (identification of protein modifications), quantification and sample comparisons are discussed more rapidly.

The analysis of proteins is much more complicated than the analysis of DNA or RNA. The technology available is less mature and more costly, and, mainly, the proteins are much more complex and fragile molecules. Nonetheless, there are important reasons to study the proteins:

- Proteins can be modified in many ways by molecules that are bound to them (PTMs). Very often, these modifications are essential for the proteins to be active. These modifications are dynamically added or removed by the cell machinery. Protein spatial conformation may change depending on its environment (acidity, presence of water, etc.). It is not uncommon that secreted proteins reach their final conformation after having left the cell only. All these variations are not defined by the gene sequences.
- Alternative splicing may generate unexpected gene products that only a proteomics analysis may reveal. Recently, experimental evidence of protein splicing has been reported.
- The RNA concentration is not always correlated to the corresponding protein concentration. Therefore DNA-chip experiments must be completed and/or validated by protein concentration analyses.
- Important circulating bio-fluids such as plasma are not made of cells and therefore studies based on genomics or transcriptomics are not possible. Nevertheless, such fluids give a general picture of the organism state – via hormones – and may be very appropriate for diagnostic. Moreover, fluids

such as plasma (blood) and tears for instance are accessible via non invasive techniques.

- Most of the cell machinery is controlled and effectuated by proteins. Hence only the study of proteins can give the full picture.

A typical proteomics project

Most of the analyses in proteomics start from a biological sample that must be properly collected and prepared. Sample preparation is made of multiple stages, among the first ones we find the adjunction of protease inhibitors to stop protein degradation and maintain the original contents of the sample. Then one generally faces the problem that the sample is too complex: it contains proteins at very different concentrations.

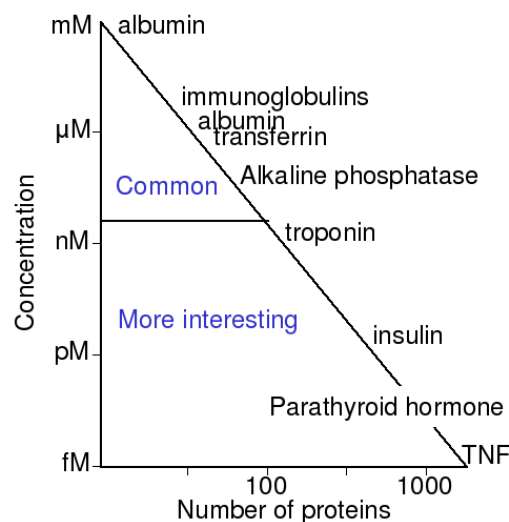


Figure 1: Protein concentrations in human plasma

The technologies available for identifying the proteins cannot deal with extremely complex samples without missing most of the low abundant proteins. There are various techniques for separating the proteins contained in the original sample and for obtaining simpler samples – of reduced complexity – that are more amenable to in-depth analysis.

The final analysis, aimed at identifying the proteins, is almost exclusively performed by mass spectrometry (MS). Former techniques such as Edman degradation are rarely used nowadays. MS produces data that are specific to the proteins analyzed and these data serve for database searching or, alternatively, to try to infer (part of)

the protein sequences directly. It is also possible to deduce information concerning protein concentrations and modifications from the MS data.

Protein separation techniques

We now review, with limited details, the more frequently used proteomics technologies.

Liquid chromatography. The samples analyzed in proteomics are mostly liquids. If the sample is not a liquid, e.g. bones, it must be solubilized by using acids for instance. Several technologies are grouped under the name “liquid chromatography” (LC). A LC station consists of a column (a tube) and a pump that pushes the sample into the column. Depending on the column interior, proteins go out of the column at different times, depending on their physico-chemical characteristics (hydrophobicity, charge, etc.), allowing us to collect simpler samples, the so-called chromatographic fractions.

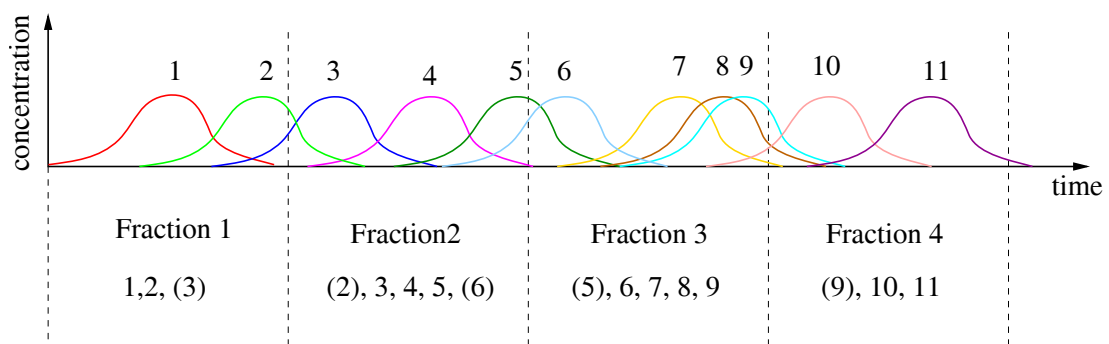


Figure 2: Fraction collection principle. Proteins elute from the LC column at different times and by collecting column output during time intervals we obtain fractions (sub-samples) of reduced complexity. In the figure we show how the concentration of 11 proteins depends on time and the effect of collecting fractions during fixed time intervals: not all 11 proteins are in each of the 4 resulting fractions.

Two main categories of LC columns exist: columns filled with a medium that influence protein elution speed and columns coated with a solid phase that interacts with the proteins. In order for the second category to impose different retention time it is generally necessary to mix the original sample with a buffer, which composition varies over time.

The word "chromatography" comes from the fact that, usually, at the end of the column there is a chromatographic measurement at a certain wavelength. This measurement is made at a wavelength that interfere with peptide bonds or certain amino acids and the intensity of the signal provides an information about the amount proteins coming out of the column. This eventually allows for estimating protein concentrations in the collected fractions.

A few abundant proteins. As already mentioned, it is possible that a few proteins represent almost all the protein mass. This extreme abundance hides less abundant and generally more interesting proteins. This is typically the situation encountered in plasma and serum, where the twelve most abundant proteins (serum albumin, immunoglobulins) comprise more than 95% of the total mass of plasma proteins. This problem is less dramatic when analyzing cell cultures or certain biopsies. There exists chromatographic columns coated with a solid phase containing antibodies that are aimed at retaining such abundant proteins by affinity with high-efficiency. The final concentration of the abundant proteins is massively reduced and the relative abundance of the minor proteins is thus augmented. Despite its obvious advantage, this technique has possible drawbacks such as variations in the amount of retained abundant proteins, which potentially introduce extra variability in the samples, and the risk to retain interesting proteins that interacts with the targeted abundant proteins.

Gel filtration chromatography. Proteins are separated according to their size. Gel filtration columns are made of a heterogeneous medium that forces small proteins to go through a longer path in order to go out of the column than larger proteins. Consequently, large proteins elute first from the column.

Ion exchange chromatography. Proteins are separated according to their charge. The column is coated with a solid phase carrying charges that interact with protein charges. If the sample is mixed with an acidic or basic buffer, there is a competition between the column coating and the buffer. As the composition of the buffer progressively changes from acidic to basic or vice versa, proteins are unbound from the column coating differentially, see Figure.

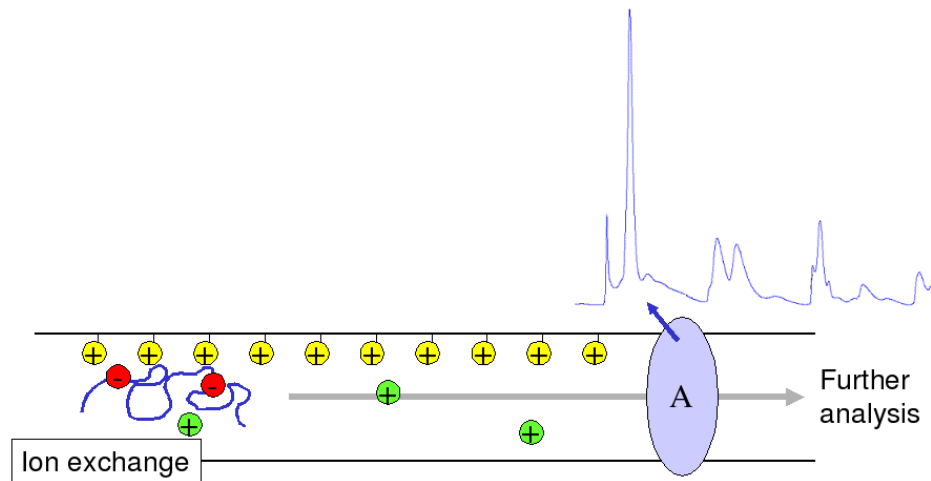


Figure 3: Ion exchange column. The different elution times are obtained by changing the buffer and hence modifying the competition between the buffer and the column coating to bind to proteins, depending on proteins charge. Hence the separation according to the proteins charges. We also represent the chromatogram as generated by the detector (A).

Reverse phase chromatography. Proteins are separated according to their hydrophobicity by a principle similar to the ion exchange chromatography, i.e. competition between column coating and the buffer.

2D Gels. Two dimensional electrophoretic gels (2D gels) is an old technique for separating proteins that is still used today as it has some unique features. It is not a liquid chromatography technology. The principle of 2D gels is to first separate proteins by their isoelectric point (pI). The isoelectric focusing (IEF) concentrates proteins at their pIs and allows proteins to be separated on the basis of very small charge differences. Under the influence of an electric field, a protein moves in a pH gradient until it reaches the position where its net charge is zero (pI). IEF is performed in strips that are then deposited at one side of a rectangular polyacrylamide gel. The second dimension of separation is obtained by separating proteins according to their length in a direction orthogonal to the IEF strip. Sodium dodecyl sulphate (SDS) is an anionic detergent which denatures proteins by "wrapping around" the polypeptide backbone. SDS confers a negative charge to the protein in proportion to its length. The application of a second electric field achieves the second separation.

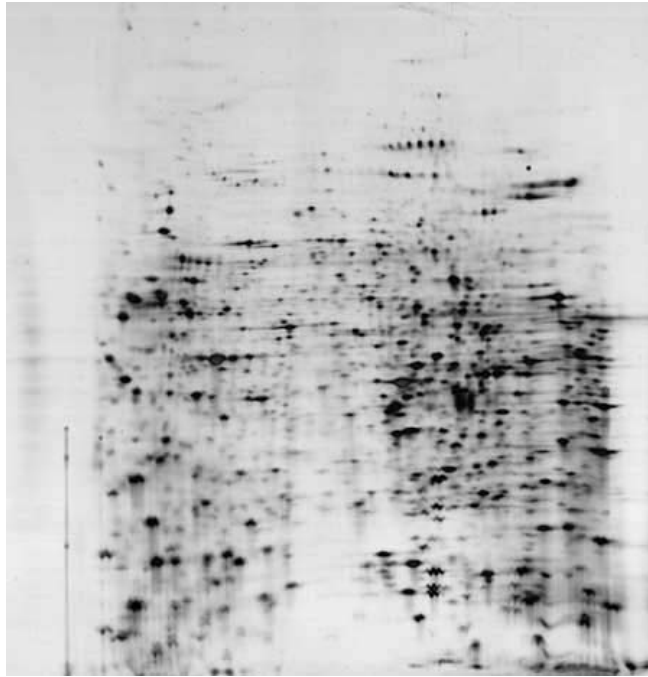


Figure 4: A typical 2D gel.

After migration, the proteins in the gel are stained to make them visible, see Figure 4. Gels are scanned and imaging techniques are used for detecting the spots. The coordinates of the spots are transmitted to a robot called a spot picker that cuts the gels to collect small pieces corresponding to the spots. These spots are simpler samples that are further analyzed by MS as chromatographic fractions would be.

Semi-quantitative information can be extracted from 2D gel images and it is possible to compare gels obtained from several tissues to do sample comparisons. Nonetheless, 2D gels are very difficult to make reproducible. Moreover, the amount of material (proteins) that can be loaded on a gel is limited and the problem of samples where a few abundant proteins constitute 95% of the protein mass is difficult to address with gels only. We mention one last limitation of 2D gels, the pI and MW range in which they work well is limited and hence part of the proteome is not amenable to 2D gels analyses.

PEPTIDE MASS FINGERPRINTING

Introduction

When one wants to search a database of protein sequences to identify proteins contained in a sample, it is necessary to have data that are specific to the proteins. Since we use MS, a first natural choice would be to use protein masses. This option does not allow identifying proteins uniquely because: (1) MS instrument precision is not absolute and several distinct proteins can have very close masses; (2) proteins are generally modified and not only one possible mass is associated to a protein but a list of possible masses, thus reducing further data specificity. We conclude that the protein mass is not specific enough, although it may be used as additional information to facilitate database searching.

There exist highly-reliable enzymes, such as trypsin or chymotrypsin, that cleave at specific locations and yield peptides of reasonable size, e.g. an average of 10-12 amino acids for trypsin. As the cleavage sites are amino acid sequence specific, the masses of the peptides are somehow correlated to the original protein sequence.

Identifying proteins on the basis of the masses of their peptides is called peptide mass fingerprinting (PMF). PMF is only possible for very simple samples because when numerous proteins are mixed together, the masses of all their peptides no longer constitute a specific set of data. There are too many masses in the spectrum and, as we do not know in advance which peptide masses correspond to distinct proteins, we have to use them all when comparing to protein sequences taken from a database. This increases the rate of possible false identifications. Moreover, ion suppression effects and instrument resolution limit the number of detectable peptides by favoring the most intense signals. Consequently, PMF is a technique that is used mainly in combination with 2D gels, whose spots contain 1-2 dominating proteins only. The classical instrument for doing PMF is a MALDI-TOF instrument. It is described in the next chapter.

Searching a database

The principle of searching a database with PMF data is as follows. Given a list of

experimental peptide masses L (obtained from an experimental spectrum), apply the enzyme cleavage rule to the database protein sequences, compute the mass of the theoretical peptides and compare with L .

The first two steps in designing a database search engine are hence *in silico* enzymatic digestion and peptide mass computation.

Theoretical digestion. The most frequently used enzyme in PMF is trypsin. The generic rule for trypsin cleavage is: cleave after lysine (Lys, K) or arginine (Arg, R), provided it is not followed by a proline (Pro, P). Although trypsin is efficient in cleaving proteins, it happens that some cleavage sites are missed. Such locations are called missed cleavages. Since multiple copies of the protein are digested simultaneously, it is possible sometimes to observe both perfect cleavages and missed cleavages. Therefore, all cases must be considered in the theoretical digestion.

Protein: MC*TM*ACTKGIPRKQWWEM*MKPCKADFCV				
Tryptic digestion (peptide, start, stop, nmc, mass):				
MCTMACTK	0	7	0	960.353715
QWWEMMKPCK	13	22	0	1381.598105
ADFCV	23	27	0	553.220625
MCTMACTKGIPR	0	11	1	1383.613105
MCTMACTKGIPRK	0	12	2	1511.708065
GIPRK	8	12	1	569.364915
GIPRKQWWEMMKPCK	8	22	2	1932.952455
KQWWEMMKPCK	12	22	1	1509.693065
KQWWEMMKPCKADFCV	12	27	2	2044.903125
QWWEMMKPCKADFCV	13	27	1	1916.808165

Figure 5: Theoretical digestion. (nmc) is the number of missed cleavages. Modified amino acids are indicated by an asterisk at their right-hand side (C is modified by iodoacetamide +57Da, M is oxidized +16Da).

Peptides with one missed cleavage are not uncommon (typically 25% of the peptides), whereas peptides with two or more missed cleavages are less frequent. they are also larger and may be difficult to ionize thus giving a weaker signal in the

spectrum. It is customary to consider peptides with one missed cleavage maximum only when searching a database.

Mass computations. Because of the peptide structure represented in Figure 6, to compute the mass of unmodified peptides is straightforward: add the individual amino acid masses and the mass of a water molecule. As a matter of fact, there is an extra hydrogen at both the C- and N-term sides, and an extra oxygen at the C-term side.

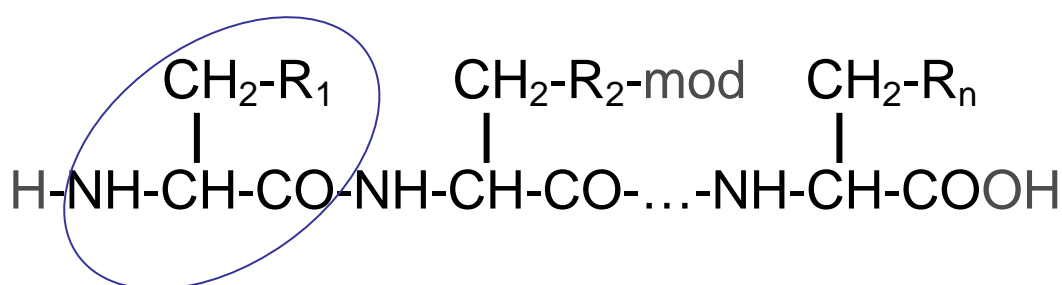


Figure 6: Peptide structure and mass computation.

As we already mentioned, proteins can be modified by PTMs or by chemical reagents such as the ones used for suppressing di-sulfur bonds between cysteines. The modifications are linked to specific amino acids in the protein sequence and the mass of these amino acids must be modified for computations.

Two types of modification must be considered. Fixed modifications are always present, i.e. the mass of the modified amino acids can be replaced by the original mass corrected by the mass delta (positive or negative) due to the modification. For instance, the reagents normally used for breaking di-sulfur bonds are very efficient and we can consider that every cysteine is modified always. Carboxyamidomethyl cysteines (Cys_CAM) have a mass augmented by 57.02146 Da, i.e. their mass is $103.00919 + 57.02146 = 160.03065$ Da.

Variable modifications are not always present and as we compute the theoretical masses of peptides we have to consider every possible combination. For instance, the oxidation of methionines is a typical variable modification that adds 15.9949 Da. Given a peptide sequence 'ARMTHLLMK' we must compute 4 theoretical masses

because there are two variable modification sites. When several variable modifications are taken into account simultaneously, the combinations to compute include all possible modification sites and their number grows very fast.

Scoring function

As we already explained the common method for searching PMF data against a database of protein sequences consists in digesting the protein sequences *in silico* and then in comparing the theoretical and the experimental masses. This comparison involves a scoring function, which role is to measure the correlation between experimental and theoretical data.

The most obvious PMF scoring function is of course the so-called shared peak count, i.e. to count the number of matching theoretical and experimental masses given a certain instrument precision. The instrument precision is specified as a mass tolerance Delta and two masses within a distance Delta are considered as identical (match). Delta can be expressed as an absolute mass error in Da or as a relative mass error in ppm. Since MALDI – and other – instruments mass precision diminishes with increasing masses, a relative error tolerance is more appropriate generally.

To associate a score to every protein in the database is one initial task of database searching. The next task is to decide which protein(s) is(are) the correct one(s) or to associate confidence levels to the protein identifications found in the database. This second task requires choosing a specific method for estimating the confidence levels and, obviously, its performance will be influenced by the performance of the scoring function. A good scoring function already discards many false positive protein identifications by giving high scores to the true positive ones.

MOWSE score. The MOWSE score is a heuristic PMF score that is used – slightly modified – by Mascot PMF search engine. It is based on a model of the typical mass of peptides. By using a database of protein sequences, we learn the typical number of enzymatic peptides in mass windows of 100 Da depending on the intact protein mass (the protein mass may eventually bias the distribution of peptide masses); protein mass window size is 10 kDa.

When a database is searched, after the list of matching experimental masses has been established for a given mass tolerance, the principle of MOWSE score is to compute the score s by combining a quantity similar to a likelihood, a scale factor and the protein mass:

$$s = \frac{5000}{M_p \prod_{m \in S} f_{\lfloor m/100 \rfloor \lfloor M_p/10000 \rfloor}}.$$

The protein mass M_p reduces the score to limit the influence of large theoretical mass lists caused by large proteins, e.g. nebuline. The sort of likelihood is used as a divisor such since more peculiar peptide masses are assumed to bring more support to protein identification.

MSA score. The heuristic MSA score is based on the observation that properly calibrated experimental masses should not deviate too much from the theoretical masses. It also includes a requirement that the protein sequence coverage should be as high as possible to validate protein identification.

When comparing the experimental masses with the theoretical masses of a database sequence, MSA applies two successive re-calibration steps and too far experimental masses are removed from the set of matched masses progressively. The standard deviation of the mass errors of the finally matched masses is used as an indicator of the match quality (the smaller the better). The number of matched masses n , as well as the sequence coverage in percent g , are the two other elements of a heuristic scoring function:

$$Z = 100 - \frac{500}{\sigma n^2 g}.$$

Reliable protein identifications should yield scores larger than or equal to 99.

ProFound score. ProFound is a popular commercial search engine for PMF spectra. The score used by this engine is derived via a Bayesian approach. For a given database sequence, we are interested in computing the probability that this sequence corresponds to the experimental spectrum at hand. The derivation of ProFound's formula relies on standard combinatorial arguments (probability to match n masses in a set of m theoretical masses, etc.) and assumed Gaussian mass errors.

OLAV-PMF score. The previous scoring functions do not include any physico-chemical properties of peptides in their underlying models. To improve over these scoring functions, it is important to introduce more complicated models aimed at capturing certain trends of peptide ionization. The price to pay is a model with more parameters, that must be trained for specific sample preparation conditions and MALDI instrument settings.

Another point which is not considered by the previous scoring functions is that the best statistics in an hypothesis test is often a likelihood ratio. Therefore, we introduce a family of scoring functions that are both designed as likelihood ratios and that model certain properties of peptides such as observed modification and amino acid composition. We also model protein sequence coverage.

RAW SPECTRUM PROCESSING

Introduction

In this chapter we explain how masses can be obtained from the experimental spectrum acquired by the MS instrument. We primarily illustrate that in the case of peptide masses measured by a MALDI-TOF instrument. The masses extracted from a spectrum compose the mass list and they are the input data of database searching.

MALDI-TOF

The classical instrument for doing PMF is a MALDI-TOF instrument. The principle of MALDI is as follows. The sample to analyze (a digested gel spot) is mixed with a reagent named a MALDI matrix. This mixture is then deposited on a metallic plate and crystallized (in vacuum). A laser is used to turn the sample into a cloud of ionized peptides that are accelerated by a constant electric field. The ions are charged positively by the gain of one proton and they fly along an empty tube. The masses of the peptides are determined by the time needed to reach the detector at the tube extremity, hence the name time-of-flight (TOF).

Peak Detection

To obtain good quality mass lists, which obviously facilitate database searching and the solution of other problems in computational proteomics, it is of prime importance that the processing of the mass spectra is properly done. This is the role of peak detection or peak picking software. Such software generally comes with the instrument and is provided by the instrument manufacturer since it is convenient that it is integrated with software controlling the instrument and data acquisition.

The signal acquired by the instrument is a sampled continuous signal that contains chemical and electronic noise. The masses to determine correspond to the top of the major peaks, whereas small peaks are the contribution of chemical (slow oscillation of roughly 1 Da frequency) and electronic (rapid oscillation) noises.

Each atom that composes a peptide may have isotopes, i.e. supplementary neutrons in its nucleus that augment its mass by 1.00728 Da. The probability to have isotopes is specific to each atom and consequently, the probability to have isotopes is specific to each peptide, depending on its atomic composition. Multiple copies of each peptide in the instrument cause a peptide signal made of several peaks. The extra peak due isotopes must be removed from the mass list before database searching.

One method of peak detection consists in (1) recognizing the individual peaks (to build the so-called peak table), and (2) de-isotope. An *ad hoc* algorithm first localizes potential peaks, which are subsequently more precisely determined by one of the following common methods:

- Finding the m/z value where the slope is equal to zero (pre-smoothing mandatory).
- Finding the m/z value where the signal is the most intense, i.e. the apex (pre-smoothing would be wise).
- Computing the centroid, i.e. the m/z value where half of the area under the peak is reached.
- Fitting the spectrum peak to a theoretical model of a peak with shape parameters such as signal intensity, width, baseline. A Gaussian is generally appropriate for that.

De-isotoping of the peak table can be achieved by looking for peaks at one Da distance with reasonable relative intensities and grouping them. Such a task may be done by applying heuristic rules or by introducing a scoring function that measures the quality of alternative peak groups in order to optimize the de-isotopization.

Instead of first processing the peaks independently and then grouping them, it is possible to define a notion of peptide signal pattern and to look for all the isotope peaks of a peptide simultaneously.

TANDEM MASS SPECTROMETRY

Introduction

When we introduced the technique of peptide mass fingerprinting we explained that, due to the enzyme specificity, the set of masses of enzymatic peptides constitute a much more specific data set for searching a database compared to the sole protein mass. Tandem mass spectrometry can be introduced in a similar way: There exist techniques to break peptides into smaller molecules, the so-called fragments, and because such fragmentation processes are governed by certain rules the set of fragment masses constitutes a more specific data set compared to the sole peptide mass.

On-line mass spectrometry

Today, the main two techniques for protein/peptide ionization are MALDI and electro-spray (ESI). ESI works in liquid phase and thus it can be combined with an LC column for peptide separation in order to analyze relatively complex samples. This is not possible with MALDI that requires more or less one protein per sample.

Fragmentation cell

A classical technique to induce peptide fragmentation is to use an inert gas such as helium to create collisions with the peptides. Such a technique is named collision induced dissociation (CID), note that the word fragmentation is sometimes replaced by dissociation.

Mass analyzer

We already described TOF detectors that generally yield high-resolution and good mass precision. Another widely used technology, though less precise usually, is quadrupole ion filters. Fourier transform ion cyclotron resonance provides high mass accuracy (1 ppm). Finally, ion trap mass analyzers deliver medium precision (500 ppm) but are very versatile and robust instruments. They dominate the market today.

The fragmentation spectrum

The fragmentation of peptides follows certain rules. Would it not be the case, we could not use the fragment masses as specific data for identifying the peptides. A situation that is similar to the rules for enzymatic digestion and peptide mass fingerprinting, although fragmentation is governed by more stochastic rules.

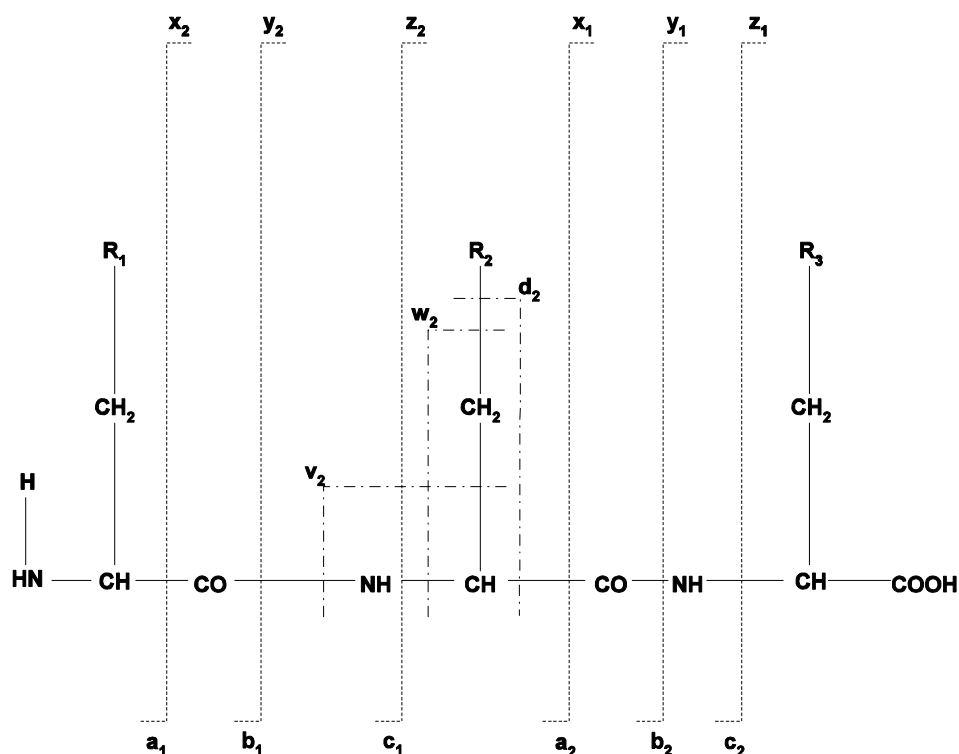


Figure 7: Schematic view of the generic fragmentation locations in a peptide. Fragments of type a, b, and c are N-terminal fragments, i.e. they include the N-terminal side of the peptide, whereas x, y, and z are C-terminal fragments. Fragments of type v, w, and d include part of the side-chain and their are only created by high-energy collisions. They may be used for distinguishing between isobaric amino acids such as leucine and isoleucine.

Since fragments of a given type always include the same atoms between residues, it is possible to compute their theoretical masses by summing the amino acid masses and by applying a correction for the atoms.

Ion type	K	V	P	Q	V	S	T	P	T	L	R
a	-	200.2	297.2	425.3	524.4	611.4	712.4	809.5	910.5	1023.6	-
b	-	228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	-
c	-	245.2	342.3	470.3	569.4	656.4	757.5	854.5	955.6	1068.6	-
x	-	1123.6	1024.5	927.5	799.4	700.4	613.3	512.3	415.2	314.2	201.1
y	-	1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
z	-	1081.6	982.5	885.5	757.4	658.4	571.3	470.3	373.2	272.2	159.1

Figure 8: Example of fragment masses.

A given type of instrument does not produce all the type of fragments usually, only a limited number of them a clearly detected in the MS/MS spectrum. Abundant fragment ions of tryptic peptides in most spectrometers are b and y. Fragments of type a yield weak signals and c, x, and z fragments are barely produced under normal conditions.

Multiply charged peptide (precursor) ions may generate multiply charged fragment ions. It is hence useful to be able to compute multiply charged fragments theoretical masses: add one proton mass for each extra charge and divide by the charge.

Ion type	K	V	P	Q	V	S	T	P	T	L	R
y	-	1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
y++	-	549.3	499.8	451.3	387.2	337.7	294.2	243.7	195.1	144.6	88.1
b	-	228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	-
b++	-	114.6	163.1	227.1	276.7	320.2	370.7	419.2	469.8	526.3	-

Figure 9: Doubly charged fragment masses.

For tryptic peptides the most common multiply charged fragments are doubly charged fragments. Triply charged fragments a normally observed for large peptides only and quadruply charged fragments are such low abundant that their signal is hidden by the noise. Recall that to observe fragments with charge z the peptide must be charged at least z times.

During the fragmentation process, certain residues (serine, threonine) may loose water or ammonia. Since such losses are not systematic we eventually may want to consider all combinations of losses. Moreover, since not all the residues may have loss, not all fragments may loose water of ammonia. For instance, to loose two water and one ammonia molecule, a fragment must include in its sequence at least two of S and T and at least one of N, Q, R.

Modified peptides

When a peptide is modified there is not only an impact on the peptide total mass, where the total mass of all the modifications must be added, but also the fragment masses are modified. Since a modification is bound to a specific amino acid of the peptide, the fragments that do not include this amino acid have their original mass unchanged. On the contrary, fragments that include the modified amino acid have their mass augmented by the modification mass (which may be negative).

Internal fragments

It may happen that fragments of the precursor re-fragment thus producing internal fragments, i.e. fragments that neither include the N- nor the C-terminal sites of the peptide. Normally such internal fragment are low abundant and they do not contribute to the observed spectrum significantly. A special type of internal fragments named immonium ions, resulting from y/a fragmentations, and which only include one residue, produce a detectable signal for certain residue. They are not detectable by every type of instrument but, when visible, they usefully give information about the peptide composition as they have fixed masses depending on the residue only.

MS/MS database search

The principle of searching a database with MS/MS data is similar to PMF database search. The main two differences are that we identify peptides (we do not identify the proteins directly) and we often do not know the charge state of the peptides (we only know the m/z of the precursors and the corresponding MS/MS spectra).

A simplified algorithm for searching a database is as follows: digest each database entry, compare peptide masses with experimental peptide masses, in case of match compute the theoretical fragmentation spectrum and determine a score. At the end of the database scan, the peptide identifications are grouped into protein identifications. A protein score is eventually computed.

MS/MS scoring functions

Shared peak count. Same as for PMF but count the fragment peaks.

Mascot score. Mascot developers (for MS/MS) never described their scoring function but they recognize that it is an adaptation of MOWSE score to fragmentation spectra. That is the parameters of the scoring function are trained to learn the probability to observe a fragment of a given mass given the mass of the precursor peptide ion. Some additional and proprietary preprocessing is applied to the experimental spectrum to normalize peak intensities and detect noise level.

SEQUEST score. SEQUEST scores are not based on a model but it rather rely on a heuristic approach. Namely, an initial and purely heuristic score is computed and the n best peptides found in the database are re-scored with another more sophisticated scoring function. The first purely heuristic scoring function takes into account the number of matched ions, their intensities, the consecutive matches in a series, and, if applicable, the presence of immonium ions. The second scoring function creates an artificial spectrum from the theoretical fragment masses and gives intensities to the peaks. This artificial spectrum is then compared via a cross-correlation function to the experimental spectrum.

OLAV score. The general approach here is, as in the PMF case, to design the scoring function as a likelihood ratio and to consider informative patterns that may be observed from the comparison of theoretical and experimental masses. We use the probabilities to observe each ion type with a given instrument, a HMM scores the consecutive matches, a model of typical intensity distributions scores the observed intensities, and there are also components of the score that depend on the amino acid founds at the ends of the fragments.

Protein identification

Besides peptide identifications we are interested in the proteins present in the biological sample. Therefore, peptide identification can be regarded as an intermediary step towards protein identification usually. To obtain reliable protein lists based on reliable peptide lists is not as straightforward as it might seem since

several complications occurs. A first problem is caused by peptides shared by several proteins, variants or paralogs:

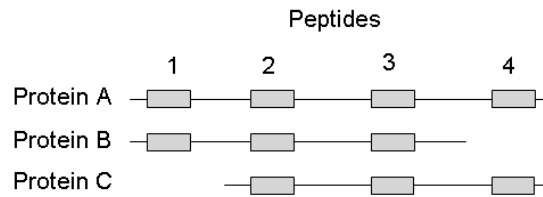


Figure 10: One example of ambiguous protein identifications. It is impossible to decide which of protein A alone or B+C or A+B+C should be considered as identified. Other such problematic patterns exist.

Errors in sequence databases can cause wrong identifications and redundancy is a source of multiple identifications or ambiguous cases such as in Figure 10. The situation is even more complicated if several databases are considered for identification. We typically have to assign them different reliability levels.

Two solutions exist for producing protein identifications. The first one is to simply rely on a set of empirical rules such as a certain number of required distinct peptides, thresholds on peptide scores, multiple occurrences of the protein in different LC fraction in case of protein separation, etc. Alternatively, we can compute a protein score and set a threshold on the latter. Protein score computation is usually performed by using a probabilistic model and by including in the model elements related to the empirical rules of the score-free approach.

Independent of the method used for obtaining protein identifications, the aim is to obtain a list of reliable identifications, where each set of proteins that are identified by the same peptides is reported as a group with one representative protein sequence. Ambiguous identifications can be reported additionally or discarded.

OTHERS

Peptide de novo sequencing

So far we identified MS data by searching a biological sequence database but there are situation where such a database is not available or not appropriate. A classical example is the analysis of a sample coming from an organism whose genome is not completely sequenced. If a significant proportion of the gene products are still unknown for this organism then to search known sequences will not explain much of the MS data. A more difficult example is the case where peptides are modified in an unexpected manner and hence are not found via the variable modifications allowed in the database search. To consider all possible modifications is not feasible and thus a method that would predict part of peptide sequences – the non modified parts – would allow to recognized candidate peptides from the database and then a dedicated processing could reveal the modifications.

To predict the peptide sequence from the MS/MS spectrum directly – de novo peptide sequencing – is a difficult problem and to predict short reliable parts of the sequence, the so-called sequence tags, is more realistic. The latter sequence tags can then be used as an incomplete but reliable sequence or they can be used for searching a database by allowing mismatches. Sequence tags of several peptides from the same protein identify the protein specifically.

To predict sequence tags can be achieved through several methods. Typically three main approaches can be identified: (1) heuristic methods; (2) graph theoretic algorithms; (3) MCMC algorithms.

Heuristic methods build solutions by enlarging previous solutions and code for many empirical knowledge of peptide fragmentation. Graph theoretic algorithms first translate the problem into a directed acyclic graph problem by representing every experimental mass as a node and by linking nodes with a mass difference close to an amino acid mass. The predicted peptide sequence is then a “longest” path in the graph. MCMC algorithms optimize a MS/MS scoring function over the space of all possible peptide sequences.

Direct genome searches

Proteomics provides experimental data that can be used for further annotating genome sequences, thereby complementing existing annotations, which are obtained in silico partially. New genes, new exons, and new splice variants can be recognized by this technique.

When dealing with a eukaryotic genome the main difficulties are the size of the search space – the translated human genome yields 7 billions amino acids – and intronic sequences that break the continuity of the coding sequence.

The extremely large size of the search space forces us to use stringent thresholds to avoid a myriad of false positive peptide identifications, thereby causing many false negatives. This limitation can be attenuated by improving MS/MS scoring functions and by introducing alternative search strategies: search gene predictions first or combine with peptide de novo sequencing.

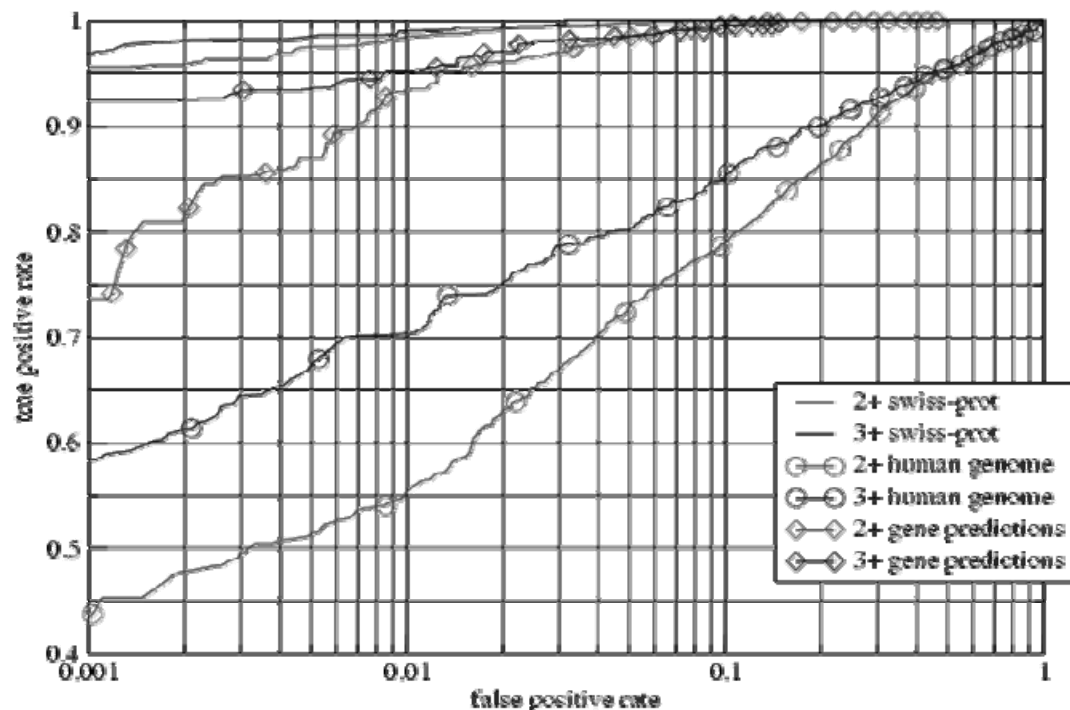


Figure 11: ROC curves for ion trap data illustrating the increase of false positives as the database size grows. Doubly and triply charged peptides searched against the human part of swiss-prot, all gene predictions obtained by genescan and hmmer, and the translated human genome. All the peptides in the dataset are contained in one exon only.

To get rid of the intronic sequences and identify peptides that are across several exons requires to couple gene structure prediction with MS data identification. Such algorithms exist and can be applied without causing much additional false positives.

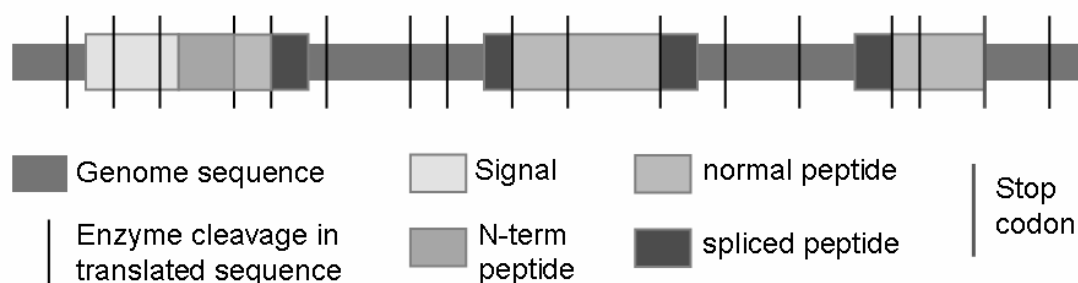


Figure 12: Tryptic peptides in the translated amino acid sequence do not coincide with splice site necessarily. Spliced peptides are coded across two exons or more.

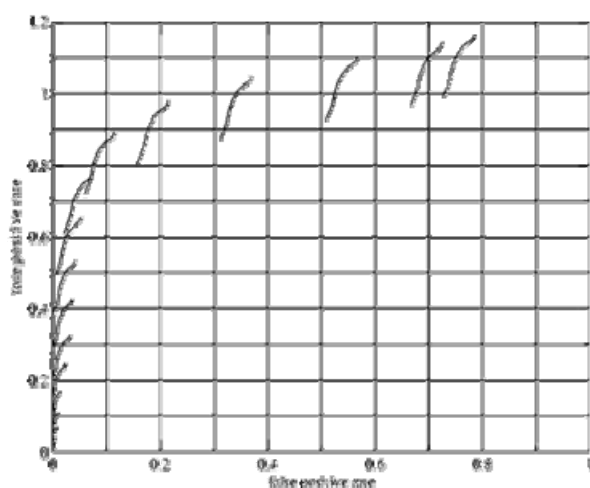


Figure 13: Additional spliced peptides can be found by locally refining a regular genome search through splice sites prediction. Results for a standard ion trap instrument, a linear ion trap coupled with a Fourier Transform instrument for exact parent mass acquisition yields spliced peptides identifications without additional false positives.

Differential proteomics

The implementation of differential proteomics analysis of biological samples can be achieved by several techniques. Classical semi-quantitative methods involve the comparison of 2DE gel images and spot volume computations. More recently, these methods have been complemented by DIGE staining, a technique that allows to stain up to three samples and then to pool them before 2DE gel production. By examining the gel at different wavelengths it is possible to compare the samples.

Nowadays a lot of comparative studies are performed without gels and they apply techniques that can be divided into two categories: label-free methods and labelling methods.

Label-free methods do not necessitate any special sample preparation and they either use areas under chromatograms or peptide counts to estimate relative/absolute peptide abundances. From the latter the protein abundances are deduced by averaging or any other method.

Labelling methods necessitate to prepare samples specifically before to pool them and to analyze them simultaneously, the relative peptide abundance being deduced afterwards as well as the protein abundance. Isotopic labelling introduces additional isotopes for certain amino acids or at specific places of the peptide. This causes the peptides to appear as pairs of peaks, labelled copies of the peptide being heavier. The relative intensities of the peaks give the relative abundances.

Another kind of labelling technique bind a cleavable label to the peptides – ICAT, iTraq – that cause shifted masses (ICAT), as for isotopic labelling, or additional reporting peaks (iTraq).

By spiking known quantities of a peptide in a sample it is possible to obtain absolute concentration estimations in a way that is very competitive compared to classical antibody-based methods.

REFERENCES

General

Aebersold, R. and Mann, M. (2003) Mass spectrometry-based proteomics, *Nature*. **422**: 198-207

Washburn, M.P., Wolters, D. and Yates, J.R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**: 242-247

Peptide mass fingerprinting

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. **20**: 3551-3567.

Zhang, W. and Chait, B.T. (2000) ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal Chem.* **72**: 2482-2489

Egelhofer, V., et al. (2000) Improvements in protein identification by MADLI-TOF-MS peptide mapping. *Anal Chem.* **72**: 2741-2750

Magnin, J., Masselot, A., Menzel, C. and Colinge, J. (2004) OLAV-PMF: a novel scoring scheme for high-throughput peptide mass fingerprinting. *J Proteome Res.* **3**: 55-60

Tandem mass spectrometry

Perkins, D.N., Pappin, D.J., Creasy, D.M. and Cottrell, J.S. (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*. **20**: 3551-3567.

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

Keller, A., Nesvizhskii, A.I., Kolker, E. and Aebersold, R. (2002) Empirical statistical model to estimate the accuracy of peptide identification made by MS/MS and database search. *Anal Chem.* **74**: 5383-5392.

Colinge, J., Masselot, A., Giron, M., Dessingy, T. and Magnin, J. (2003) OLAV: Towards high-throughput MS/MS data identification. *Proteomics.* **3**: 1454-1463

Colinge, J., et al. (2004) High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics.* **4**: 1977-1984

Craig, R., Cortens, J.P. and Beavis, R.C. (2004) Open source system for analyzing, validating, and storing protein identification data. *J Proteome res.* **3**: 1234-1242

Geer, L.Y., et al. (2004) mass spectrometry search algorithm. *J Proteome Res.* **3**: 958-964

Nesvizhskii, A.I. and Aebersold, R. (2005) Interpretation of shotgun proteomics data. *MCP.* **4**: 1419-1439

Cargile, B.J., Bundy, J.L. and Stephenson, J.L. (2004) Potential for false positive identifications from large databases through tandem mass spectrometry. *J Proteome Res.* **3**: 1082-1085

De novo peptide sequencing

Taylor, J.A. and Johnson, R.S. (2001) Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Anal Chem.* **73**: 2594-2604

Ma, B., et al. (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.* **17**: 2337-2342

Frank, A. and Pevzner, P. (2005) PepNovo: de novo peptide sequencing via probabilistic network modelling. *Anal Chem.* **77**: 964-973

Tanner, S., *et al.* (2005) *Anal Chem.* **77**: 4626-4639

Skilling, J. (1999) Improved methods of identifying peptides and protein by mass spectrometry. European Patent Application EP 1,047,107,A2

Shevchenko, A., *et al.* (2001) Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal Chem.* **73**: 1917-1926

Heredia-Langner, A., *et al.* (2004) Sequence optimization as an alternative to de novo analysis of tandem mass spectra. *Bioinformatics.* **20**: 2296-2304

Genome searches

Kuster, B., Mortensen, P, Andersen, J.S. and Mann, M. (2000) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics.* **1**: 641-650

Chen, T (2001) Gene-finding via tandem mass spectrometry. Proceedings of RECOMB 2001, Montreal, Canada

Colinge, J., *et al.* (2005) Experiments in searching small proteins in unannotated large eukaryotic genomes. *J Proteome Res.* **4**: 167-174

Differential proteomics

Liu, H., Sadygov, R.G. and Yates, J.R. (2004) A model for random sampling and estimation of relative protein abundance. *Anal Chem.* **76**: 4193-4201

Colinge, J., *et al.* (2005) Differential Proteomics via probabilistic peptide identification scores. *Anal Chem.* **77**: 596-606

Julka, S. and Regnier, F. (2004) Quantification in proteomics through stable isotope coding: a review. *J Proteome Res.* **3**: 350-363

Web resources

<http://www.matrixscience.com>

<http://phenyx.vital-it.ch>

<http://insilicospectro.vital-it.ch>

<http://prospector.ucsf.edu>

<http://www.proteomecommons.org>

<http://www.isb.org>

<http://www.ebi.ac.uk/pride/>

<http://www.systemsbiology.org>


14th Annual International Conference On Intelligent Systems For Molecular Biology

ISMB 2006 Fortaleza, Brazil
August 6-10, 2006
and 2nd Annual AB3C Conference: X-Meeting

Introduction to Computational Proteomics – Open problems

Jacques Colinge


CeMM, Vienna
University of Applied Sciences at Hagenberg,
jcolinge@excite.com

Ce-M-M- 

ISMB 2006 Fortaleza, Brazil August 6-10, 2006


Outline

- Introduction to proteomics
- Peptide mass fingerprinting
- Raw spectra processing
- Tandem mass spectrometry
- Other problems

 Introduction to computational proteomics Ce-M-M- 130

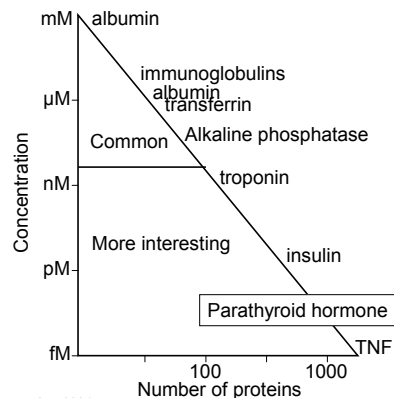
Intro :: Main problems

- To identify proteins in a biological sample
- To compare samples
- To quantify proteins
- To characterize proteins (modifications)
- Interactions, localization, protein structure


 Introduction to computational proteomics Ce-M-M- 131

Intro :: Sample complexity

- Many samples contain proteins at *very different concentrations*
- A few proteins represent most of the protein total mass
- Extreme case: plasma
- *No PCR !*
- Need for *protein separation techniques*

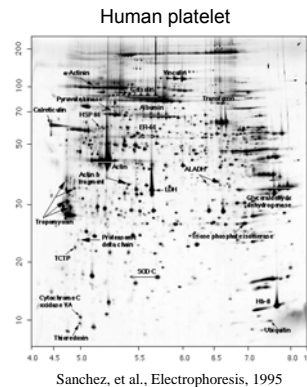


Rose *et al.*, Proteomics, 2004

 Introduction to computational proteomics Ce-M-M- 132

Intro :: 2DE-Gels

- Proteins are separated according to their *pI* and their *size*:
 - Isoelectric focusing → *pI*
 - SDS-PAGE → *size*
- Staining
- Spots are detected, picked, and *further analyzed*

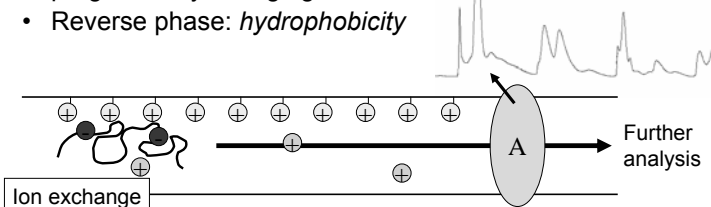


Introduction to computational proteomics

Ce-M-M- 133

Intro :: Liquid chromatography

- Proteins *elute* in a medium *with different speeds* depending on their chemico-physical properties
- Gel filtration: *size*, small proteins follow longer paths
- Ion exchange: *charge*, interaction with the wall of a column, competition with column coating by progressively changing the buffer
- Reverse phase: *hydrophobicity*

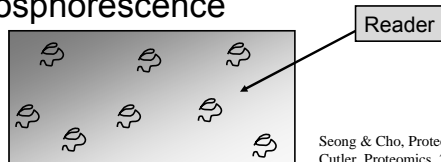


Introduction to computational proteomics

Ce-M-M- 134

Intro :: Chips

- Antibodies chips: *localized* and *specific* interaction – similar to gene chips
- Affinity chips: surface with *varying affinity properties* – example Ciphergen Chip™
- Different methods of reading the chip, e.g. SELDI, phosphorescence



Seong & Cho, Proteomics, 2003
Cutler, Proteomics, 2003



Introduction to computational proteomics

Ce-M-M- 135

Intro :: Mass spectrometry

- After sample complexity reduction one usually wants to *identify the proteins*
- Mass spectrometry *measures molecules masses*
- Masses *may constitute specific data sets*



Introduction to computational proteomics

Ce-M-M- 136

Intro :: To identify

- Identify proteins *from MS data*
- Database searching:
 - Protein databases
 - mRNA or DNA databases after translation
- Prediction *de novo*
- Databases are not complete for every organism → additional homology searches are performed sometimes



Introduction to computational proteomics

Ce-M-M- 137

Intro :: To compare

- Sample comparisons are *essential* to understand biological processes
- Need for *differential proteomics*
- *Many* relative (semi-) quantitative *methods*:
 - 2D gels through image comparisons
 - Protein chips
 - Chromatography through area comparisons
 - Labels (introduced later on)



Introduction to computational proteomics

Ce-M-M- 138

Intro :: To quantify

- Initial discoveries require *validation*
- Absolute quantitation by *MS is an alternative to ELISA*
- No specific antibodies necessary
- May be *very sensitive and precise*



Introduction to computational proteomics

Ce-M-M- 139

Intro :: To characterize

- Find *posttranslational modifications* such as phosphorylations and glycosylations
- Elucidate *glycans structures*
- Discover new (active) fragments
- Help in 3D structure determination



Introduction to computational proteomics

Ce-M-M- 140

PMF :: Introduction

- Protein masses are generally *not specific enough*
- Mass spectrometry is simpler with *small molecules* (500-4500 Da)
- Peptide mass fingerprinting (PMF):
 - Digest proteins by an enzyme (trypsin)
 - Measure resulting *peptide masses*

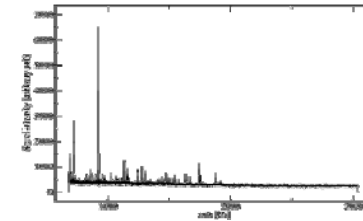


Introduction to computational proteomics

Ce-M-M- 141

PMF :: Introduction

- Sample assumed to contain *one protein only*
- Digest and measure *peptide masses*



- Extract masses from raw spectrum



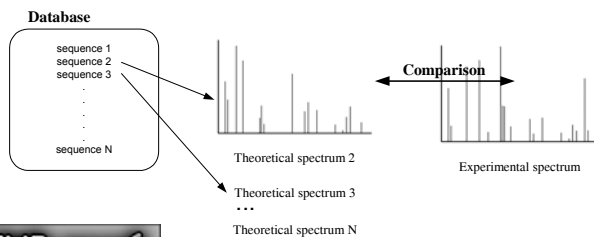
Introduction to computational proteomics

Ce-M-M- 142

PMF :: Introduction

Search a database by

- (1) Digesting database sequences theoretically
- (2) Compute theoretical peptides masses
- (3) Compare with experimental data
- (4) Output the best match(es)



Introduction to computational proteomics

Ce-M-M- 143

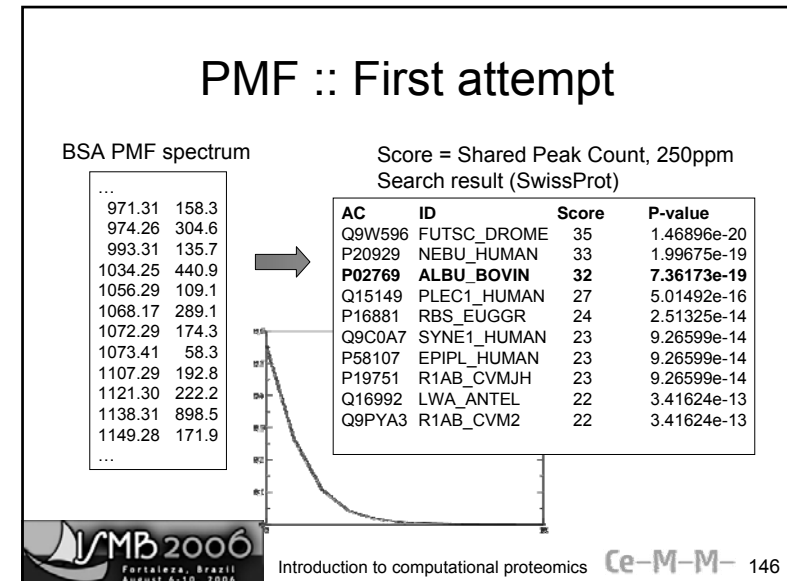
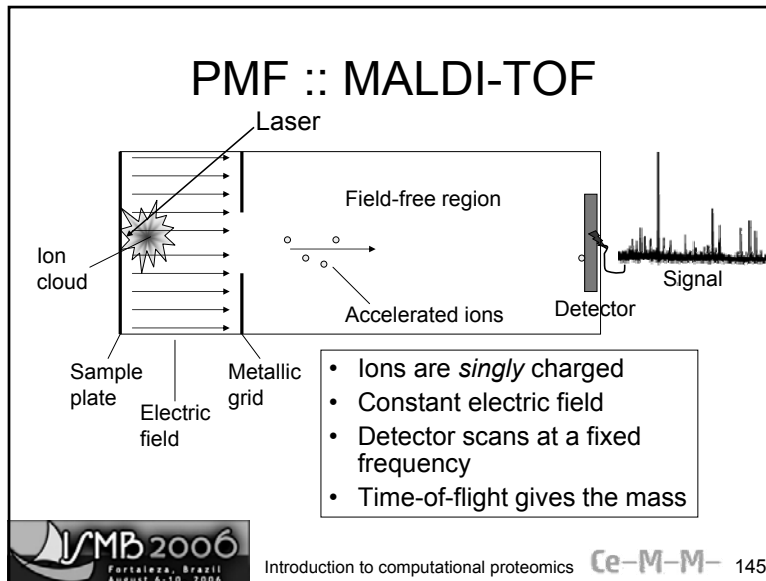
PMF :: MALDI-TOF

- How to obtain peptide masses?
- Digested samples are *mixed with a matrix* (reagent)
- Then *deposited on a metallic plate*
- Then *ionized*: Matrix Assisted Laser Desorption Ionization
- The *masses of the peptides are measured*



Introduction to computational proteomics

Ce-M-M- 144



PMF :: Theoretical digestion

- *Trypsin rule*: cleave after K and R, except when followed by P
- There are *missed cleavages*: include them in the theoretical spectrum
- Example: ATESKILTRPQSURHIS
 - No missed cleavage: ATESK, ILTRPQSUR, HIS
 - 1 missed cleavages: ATESKILTRPQSUR, ILTRPQSURHIS
 - 2 missed cleavages: ATESKILTRPQSURHIS

IMB 2006 Fortaleza, Brazil August 6-10, 2006 Introduction to computational proteomics Ce-M-M- 147

PMF :: Peptide masses

One amino acid

$$\text{H}-\text{NH}-\text{CH}(\text{CH}_2-\text{R}_1)-\text{CO}-\text{NH}-\text{CH}(\text{CH}_2-\text{R}_2-\text{mod})-\text{CO}-\dots-\text{NH}-\text{CH}(\text{CH}_2-\text{R}_n)-\text{COOH}$$

- Unmodified peptides ($p=a_1 \dots a_n$):

$$m(p) = \sum_{i=1}^{\text{len}(p)} m(a_i) + m(\text{H}_2\text{O})$$
- Modified peptides ($p=a_1 \dots a_{\{\text{mod}_j\}} \dots a_n$):

$$m(p) = \sum_{i=1}^{\text{len}(p)} m(a_i) + m(\text{H}_2\text{O}) + \sum_{j=1}^{\#\text{mod}} \Delta(\text{mod}_j)$$

IMB 2006 Fortaleza, Brazil August 6-10, 2006 Introduction to computational proteomics Ce-M-M- 148

PMF :: Variable modifications

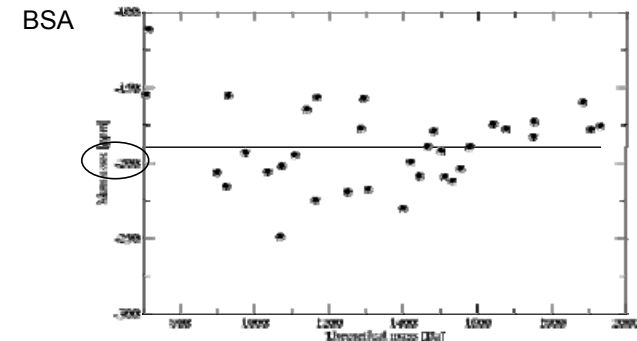
- Variable Ξ not always present
- All combinations must be computed and added to the theoretical spectrum because they have different masses
- HISTM{Oxi}C{CAM}UM{Oxi}LIK{BIOT}:
 - $(2+1)(1+1)=6$ combinations:
 - 1xCAM, 1xCAM+Oxi, 1xCAM+2xOxi
 - 1xCAM+BIOT, 1xCAM+BIOT+Oxi
 - 1xCAM+BIOT+2xOxi



Introduction to computational proteomics

Ce-M-M- 149

PMF :: Calibration



250 or 50 ppm ??



Introduction to computational proteomics

Ce-M-M- 150

PMF :: Search parameters

- Mass precision: instrument dependent, impacts specificity
- Noise level: impacts specificity
- Missed cleavages: reduce specificity
- Modifications: variable modifications increase search space
- Database size: reduce by taxonomy or estimation of pI/MW (gel)



Introduction to computational proteomics

Ce-M-M- 151

PMF :: Scoring function

- Measures the *correlation between experimental and theoretical spectra*
- The example with BSA shows that shared peak count (SPC) is not an option!!
- For small proteins, need to identify a protein with *5-6 peptide masses* in human SwissProt

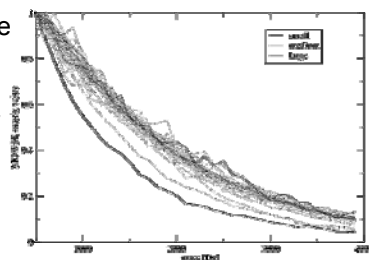


Introduction to computational proteomics

Ce-M-M- 152

PMF :: MOWSE

- Digest a protein database
- Learn frequencies of peptide masses per protein mass windows of 10 kDa
- More peculiar masses convey more information



$$\text{Score} = \frac{5000}{M_p \prod_{m \in S} f_{\lfloor m/100 \rfloor, \lfloor M_p/10000 \rfloor}}$$

Pappin, et al, Curr.Biol., 1993

Mascot implements a “probabilistic” MOWSE score, Free web server at <http://www.matrixscience.com>

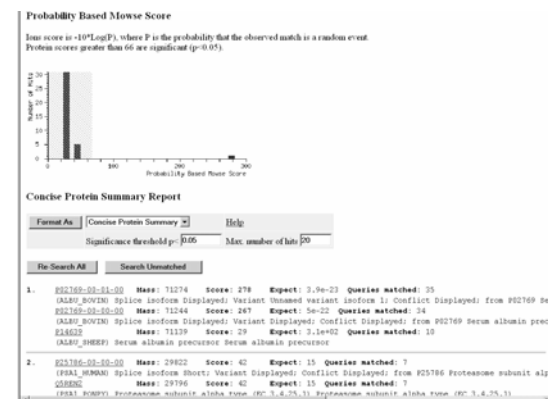


Introduction to computational proteomics

Ce-M-M- 153

PMF :: Mascot (~MOWSE)

BSA



Perkins, et al., Electrophoresis, 1999



Introduction to computational proteomics

Ce-M-M- 154

PMF :: ProFound

- Score = *Probability that the match between experimental and theoretical spectra is correct given the data at hand*
- Bayesian approach

$$\mathbb{P}(H_1|D, I) = \frac{\mathbb{P}(H_1|I)\mathbb{P}(D|H_1, I)}{\mathbb{P}(D|I)}$$

- Purely “combinatorial” model

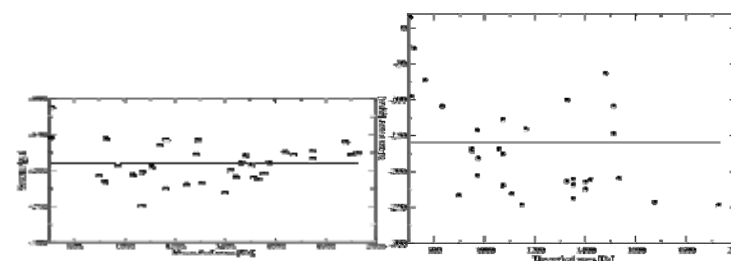
Zhang & Chait, Anal.Chem., 2000



Introduction to computational proteomics

Ce-M-M- 155

PMF :: MSA



BSA spectrum against BSA
sequence: *mean ± 50 ppm*

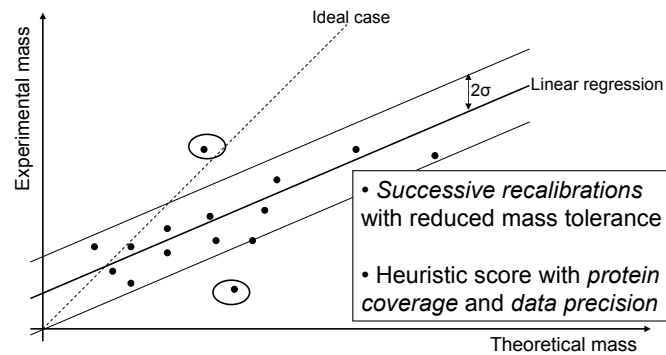
BSA spectrum against NEBULIN
sequence: *mean ± 100 ppm*



Introduction to computational proteomics

Ce-M-M- 156

MSA



Egelhofer, et al., Anal.Chem., 2000

Introduction to computational proteomics

Ce-M-M- 157

PMF :: OLAV-PMF

- Signal detection theory: best score is a *likelihood ratio*

$$L = \log \left[\frac{\mathbb{P}(M|H_1)}{\mathbb{P}(M|H_0)} \right]$$

- Collect *informative observations*
- Assume their *independence*

$$L(s) = \log \left[\mathcal{L}_{\text{cov}}(s, P) \prod_{p \in P} \mathcal{L}_{\text{comp}}(p) \frac{1}{\#M(p)} \sum_{m \in M(p)} \mathcal{L}_{\text{mod}}(m) \right]$$

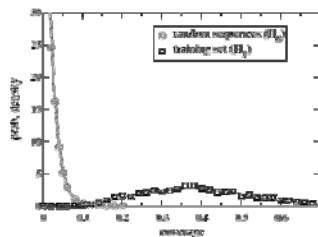


Magnin, et al., J.Prot.Res., 2004

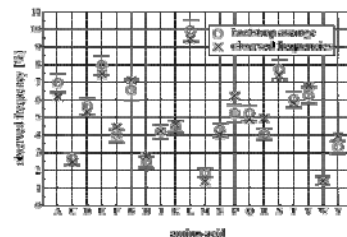
Introduction to computational proteomics

Ce-M-M- 158

PMF :: OLAV-PMF



$\mathcal{L}_{\text{cov}}(s, P) = \text{Gaussian/exponential}$



$\mathcal{L}_{\text{comp}}(p) = \text{independent Bernoulli's}$

$\mathcal{L}_{\text{mod}}(m) = \text{Binomials}$

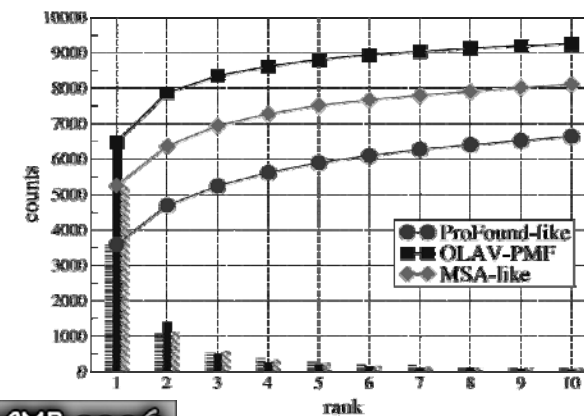
M_i	$R(M_i)$	P_i, H_1	P_i, H_0
Cys, CAM	C	0.89	0.5
Oxidation	M	0.93	0.5
Oxidation	H	0.02	0.5
Oxidation	W	0.21	0.5



Introduction to computational proteomics

Ce-M-M- 159

PMF :: "Limited" comparison



Introduction to computational proteomics

Ce-M-M- 160

PMF :: Observations

- *Re-calibration helps* all the scoring functions
- Re-calibration makes p-value estimations more complicated: only a few random scores available
- Statistics-based methods have *more potential*
- BUT *their parameters must be tuned*
- Otherwise use MSA
- Robustness not studied so far
- Other similar scorings exist, meta-scorings



Introduction to computational proteomics

Ce-M-M- 161

PMF :: Open problems

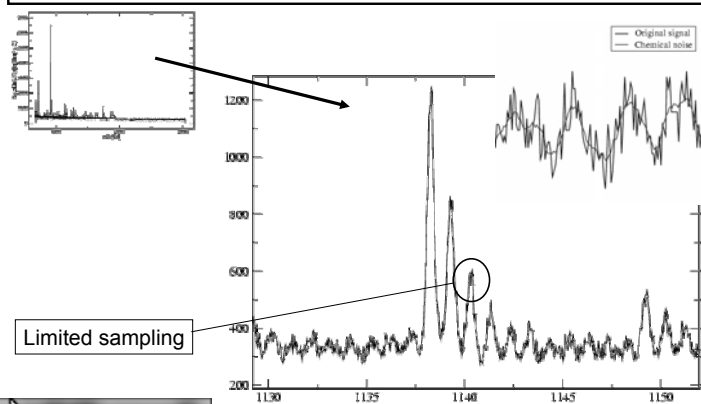
- Scoring, always scoring
- *P-value*, *E-value* estimations
- Several proteins in a spectrum



Introduction to computational proteomics

Ce-M-M- 162

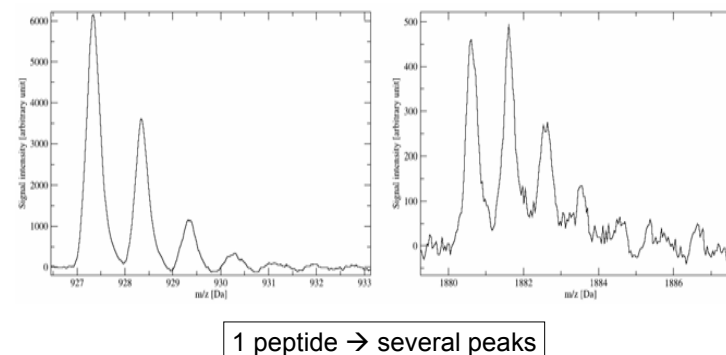
Raw :: Noisy continuous signal



Introduction to computational proteomics

Ce-M-M- 163

Raw :: Isotopes



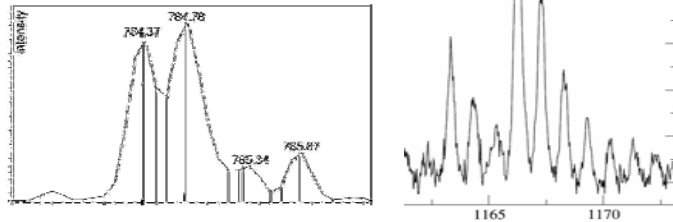
Introduction to computational proteomics

Ce-M-M- 164

Raw :: More difficult

Limited resolution
Multiple charges

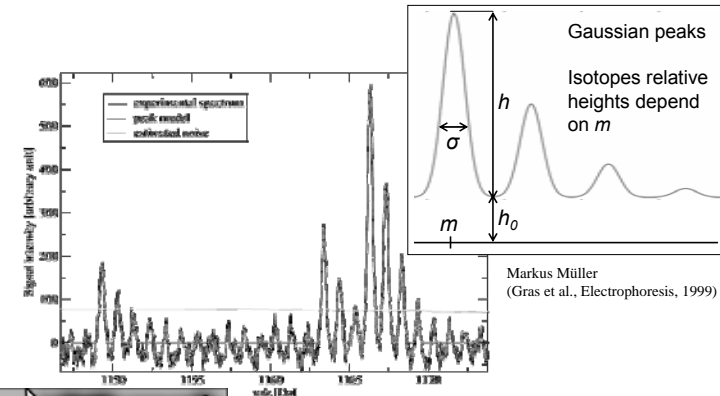
Overlapping peptides



Introduction to computational proteomics

Ce-M-M- 165

Raw :: Isotopic distribution model



Markus Müller
(Gras et al., Electrophoresis, 1999)



Introduction to computational proteomics

Ce-M-M- 166

Raw :: Open problems

- Need for sound theoretical approaches
- *Parameter-free* algorithms
- *Fast* algorithms
- Works fine with *limited resolution* and *multiple charges*
- Reliable *charge state* determination
- Eliminate *noisy peaks*
- Eliminate *low quality spectra*
- Add spectra from the same peptide



Introduction to computational proteomics

Ce-M-M- 167

MS² :: Limitations of PMF

- *Lack of specificity*: requires many peptides, *problem with small proteins*
- Needs highly separated proteins: *LC technologies are usually not applicable*
- The above limitations are *not due to MALDI* !



Introduction to computational proteomics

Ce-M-M- 168

MS² :: Specificity

- Peptide fragmentation (MS/MS or MS²) provides more information on peptides:
 - Peptides are *broken into smaller molecules*, the so-called *fragments*
 - *Fragment masses* are measured
- *One spectrum per peptide*
- Peptide can be (ideally) identified *individually*

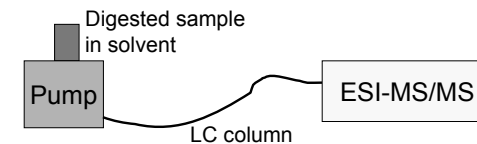


Introduction to computational proteomics

Ce-M-M- 169

MS² :: Sample complexity

- *Peptides* in digested samples are *separated* by LC in liquid phase (peptides too small for gels)
- MS² specificity → we *do not need* all the peptides of a protein *in the same spectrum*
- *Electrospray ionization* (ESI) can be performed *on-line* after peptide LC separation

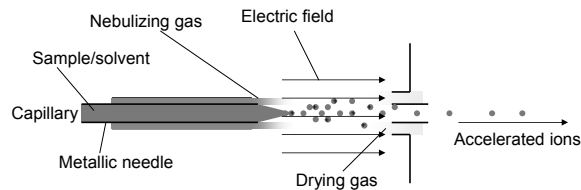


Introduction to computational proteomics

Ce-M-M- 170

MS² :: ESI

Electrospray ionization works in *liquid phase*



Peptide ions are often *multiply charged*

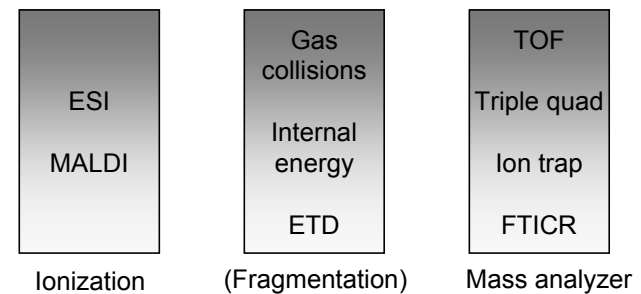


Introduction to computational proteomics

Ce-M-M- 171

MS² :: Generic mass instrument

Three main components

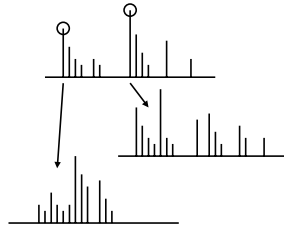


Introduction to computational proteomics

Ce-M-M- 172

MS² :: LC-ESI-MS-MS/MS

- Peptides enter the instrument *continuously*
- The instrument acquires an *MS spectrum* (peptide masses)
- It *selects* the most intense peaks
- The instrument successively acquires *MS/MS spectra* for each selected peak
- Repeat



The complete analysis of a sample (all the spectra) constitutes an *MS/MS run* or a *LC run*



Introduction to computational proteomics

Ce-M-M- 173

MS² :: Database search

- For each protein sequence
 - *Digest* the protein sequence
 - Compute *peptide masses*
 - *Finds matching* experimental masses
 - Compute *theoretical fragmentation* spectra
 - *Compare* with experimental spectra
 - Store high-scoring peptide matches
- *Group peptide matches* into protein identifications



Introduction to computational proteomics

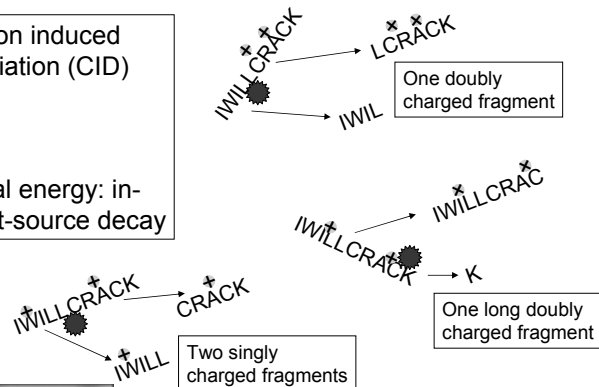
Ce-M-M- 174

MS² :: Peptide fragmentation

Collision induced dissociation (CID)

ETD

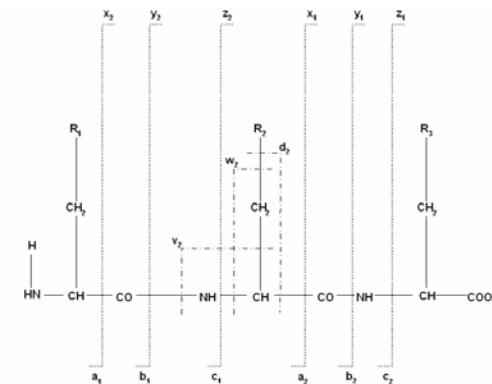
Internal energy: in- or post-source decay



Introduction to computational proteomics

Ce-M-M- 175

MS² :: Peptide fragmentation



Introduction to computational proteomics

Ce-M-M- 176

MS² :: Fragment masses

- For a peptide $p=s_1s_2\dots s_n$ we have

$$\text{mass}(b_k) = \sum_{i=1}^k m(s_i) + m(H)$$

$$\text{mass}(y_k) = \sum_{i=n-k+1}^n m(s_i) + m(H_3O)$$

- Example:

Ion type	K	V	P	Q	V	S	T	P	T	L	R
a	-	200.2	297.2	425.3	524.4	611.4	712.4	809.5	910.5	1023.6	-
b	-	228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	-
c	-	245.2	342.3	470.3	569.4	656.4	757.5	854.5	955.6	1068.6	-
x	-	1123.6	1024.5	927.5	799.4	700.4	613.3	512.3	415.2	314.2	201.1
y	-	1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
z	-	1081.6	982.5	885.5	757.4	658.4	571.3	470.3	373.2	272.2	159.1



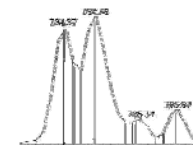
Introduction to computational proteomics

Ce-M-M- 177

MS² :: Multiply charged fragments

- Peptide ions carrying several charges can yield *multiply charged fragment ions* ($z > 1$)
- Mass analyzers normally *only "see" m/z values*
- Theoretical spectrum in m/z scale

$$m(f, z) = \frac{m(f) + (z - 1)m(\text{proton})}{z}$$



Ion type	K	V	P	Q	V	S	T	P	T	L	R
y	-	1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
y++	-	549.3	499.8	451.3	387.2	337.7	294.2	243.7	195.1	144.6	88.1
b	-	228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	-
b++	-	114.6	163.1	227.1	276.7	320.2	370.7	419.2	469.8	526.3	-



Introduction to computational proteomics

Ce-M-M- 178

MS² :: Modified peptides

- As for PMF, all the combinations of variable modifications must be considered (with position)
- The fragment masses must be adjusted
- KVPQVSTphosPTphosLR (phos=79.9663):

Ion type	K	V	P	Q	V	S	T	P	T	L	R
y	-	1257.6	1158.5	1061.4	933.4	834.3	747.3	566.3	469.2	288.2	175.1
b	-	228.2	325.2	453.3	552.4	639.4	820.4	917.4	1098.5	1211.5	-

- KVPQVSTPTLR:

Ion type	K	V	P	Q	V	S	T	P	T	L	R
y	-	1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
y++	-	549.3	499.8	451.3	387.2	337.7	294.2	243.7	195.1	144.6	88.1
b	-	228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	-
b++	-	114.6	163.1	227.1	276.7	320.2	370.7	419.2	469.8	526.3	-



Introduction to computational proteomics

Ce-M-M- 179

MS² :: Neutral losses

Ion type	K	V	P	Q	V	S	T	P	T	L	R
b	-	228.2	325.2	453.3	552.4	639.4	740.4	837.5	938.5	1051.6	-
b-H ₂ O	-	-	-	-	-	621.4	722.4	819.5	920.5	1033.6	-
b-2(H ₂ O)	-	-	-	-	-	-	704.4	801.5	902.5	1015.6	-
b-3(H ₂ O)	-	-	-	-	-	-	-	-	884.5	997.6	-
b-NH ₃	-	-	-	436.3	535.3	622.4	723.4	820.5	921.5	1034.6	-
b-H ₂ O-NH ₃	-	-	-	-	-	604.3	705.4	802.4	903.5	1016.6	-
b-2(H ₂ O)-NH ₃	-	-	-	-	-	-	687.4	784.4	885.5	998.6	-
b-3(H ₂ O)-NH ₃	-	-	-	-	-	-	-	-	867.5	980.6	-
y	-	1097.6	998.6	901.5	773.5	674.4	587.4	486.3	389.3	288.2	175.1
y-H ₂ O	-	1079.6	980.6	883.5	755.4	656.4	569.3	468.3	371.2	-	-
y-2(H ₂ O)	-	1061.6	962.5	865.5	737.4	638.4	551.3	-	-	-	-
y-3(H ₂ O)	-	1043.6	944.5	847.5	719.4	620.4	-	-	-	-	-
y-NH ₃	-	1080.6	981.5	884.5	-	-	-	-	-	-	-
y-H ₂ O-NH ₃	-	1062.6	963.5	866.5	-	-	-	-	-	-	-
y-2(H ₂ O)-NH ₃	-	1044.6	945.5	848.5	-	-	-	-	-	-	-
y-3(H ₂ O)-NH ₃	-	1026.6	927.5	830.5	-	-	-	-	-	-	-

Water: S, T; ammonia: Q, R, N

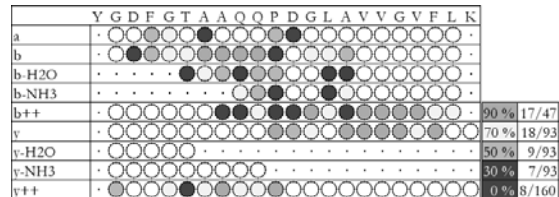


Introduction to computational proteomics

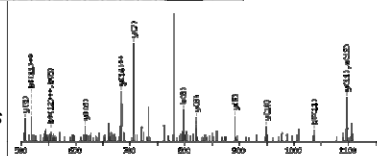
Ce-M-M- 180

MS² :: A match

- Mass tolerance → matched peaks



- Standard quality criteria:
 - intense peaks should match
 - as many as possible peaks should match
 - series of contiguous matches



Introduction to computational proteomics

Ce-M-M- 181

MS² :: Mascot

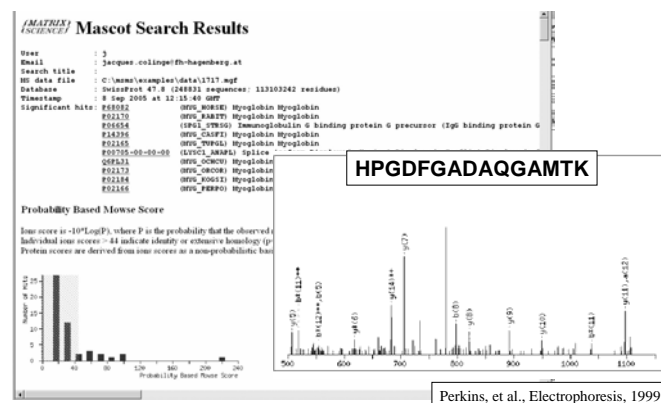
- Mascot is a commercial program that implements a *MOWSE-like score for MS/MS*: distribution of fragment masses depending on the peptide mass
- Mascot *estimates p-values*
- Mascot does some *spectrum pre-processing* to adapt to diverse data types automatically
- Latest versions include a *proprietary peak detection algorithm*



Introduction to computational proteomics

Ce-M-M- 182

MS² :: Mascot search result



Introduction to computational proteomics

Ce-M-M- 183

MS² :: Sequest

- Sequest is a commercial program that uses a *heuristic approach*
- A *preliminary scoring function* is used for rapidly scanning the database:

$$S_p = (\sum i_m) n_m (1 + \beta) (1 - \rho) / n_r$$

$\sum i_m$ is the sum of matched ion intensities
 n_m is the number of matches ions
 n_r is the total number of ions
 β is for the continuity of the match
 ρ is the presence of immonium ions

- Stores the 200 best peptides for each experimental spectrum

Eng, et al., J.Am.Soc.Mass Spectrom., 1994



Introduction to computational proteomics

Ce-M-M- 184

MS² :: Sequest

- Assign empirical intensities to the theoretical masses to create an *artificial spectrum* a
- Rescore the best matches by a second function X_{corr} (e is the experimental spectrum):

$$(a * e)(t) = \frac{\sum_{i=0}^N (a_i - \mathbb{E}(a))(e_{i+t} - \mathbb{E}(e))}{\sqrt{\text{Var}(a)\text{Var}(e)}}$$

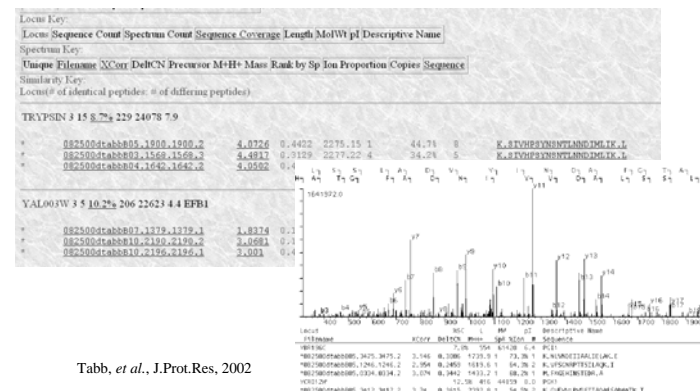
$$X_{\text{corr}} = (a * e)(0) - \mathbb{E}((a * e)(t)) - 75 \leq t \leq 75$$



Introduction to computational proteomics

Ce-M-M- 185

MS² :: Sequest search result (DTAselect)



Tabb, et al., J.Proc.Res, 2002



Introduction to computational proteomics

Ce-M-M- 186

MS² :: Post-processed Sequest (Mascot)

- Use exported data (score, p-value, initial score for Sequest, etc.) to build a statistical model of correct/wrong matches
- Tons of papers (SVM, Bayesian, NN, ...)
- Significant improvement*, especially for Sequest
- Some people *use both Sequest and Mascot* and only keep common identifications
- Decoy database to estimate FP rate

Moore, et al., J.Am.Soc.Mass Spectrom, 2002
MacCoss, et al., Anal.Chem., 2002
Nesvizhskii, et al., Anal.Chem., 2003

Sadygov & Yates, Anal.Chem., 2003
Keller, et al., Anal.Chem., 2002



Introduction to computational proteomics

Ce-M-M- 187

MS² :: OLAV

- As for PMF:
 - Collect *useful observations*
 - Build "sub-scores" as *likelihood ratios* for each observation
 - Assume *independence* and multiply

$$\text{Score } L = L_1 L_{\text{int}} L_{\text{succ}} L_{\text{pair}}$$

- OLAV is Phenyx, free web server at <http://phenyx.vital-it.ch/>

Colinge, et al., Proteomics, 2003 & 2004



Introduction to computational proteomics

Ce-M-M- 188

MS² :: Phenyx search result

Database/AC/Peptide viewJob (Job 205)

Summary Compound view Result pages: 1 2 3

AC	ID	Score	#Peptides	% Cov	Description
R0561	MYSA_DROME	125.8	14	9.3	Myosin heavy chain...
Q0505	ATPB_DROME	39.8	5	14.6	ATP synthase beta ch...
P05381	ATPA_DROME	39.1	4	11.9	ATP synthase alpha c...
P15999	ATPA_RAT	28.3	3	8.4	ATP synthase alpha c...
Q06764	TPH_CHK1	18.6	2	11.2	Tropomyosin (Alderge...
P03957	ACTS_DROME	16.2	2	10.2	Actin, indirect fil...
Q04911	THSD_DROME	12.6	1	6.3	Tropomyosin 2 (Chro...
P03958	CYC_RAT	12.5	1	13.5	Cytochrome c, somat...

Subtotal

R0561-VS2 in unprot_sprot (Protein Details)

R0561-VS8 in unprot_sprot (Protein Details)

R0561-VS14 in unprot_sprot (Protein Details)

R0561-VS7 in unprot_sprot (Protein Details)

R0561-VS10 in unprot_sprot (Protein Details)

R0561-VS11 in unprot_sprot (Protein Details)

R0561-VS5 in unprot_sprot (Protein Details)

R0561-VS13 in unprot_sprot (Protein Details)

R0561-VS15 in unprot_sprot (Protein Details)

R0561-VS6 in unprot_sprot (Protein Details)

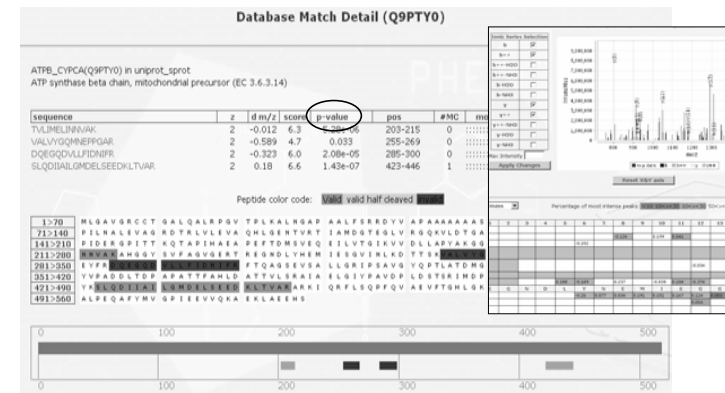
Seq. Sequence	z	m/z	d m/z	Score	P-Value	Pos.	#MC	Modif.	SN	Compound
+ F/QVEAEIVAAINLAK/R	2	821.356	0.083	13.6	1.06e-37	1880-1894	0		0	proteome7_184.txt.198
+ F/AHLEQTLDELEDEP/R	2	945.453	-0.474	11.6	8.06e-27	1024-1039	1		0	proteome7_184.txt.136
+ F/LRAGAGATAGAGLAK/R	2	815.338	0.08	10.9	1.56e-23	1149-1164	0		0	proteome7_184.txt.251
+ F/GAVVEEQQLAAWR/R	2	790.264	-0.391	10.4	4.96e-21	1485-1495	0		0	proteome7_184.txt.181
+ F/LARAEITISLNQK/R	2	788.152	-0.253	10.2	3.27e-20	1396-1409	0		0	proteome7_184.txt.71
+ F/LSTIRDELQLEDFR/A	2	823.399	-0.181	9.5	2.6e-17	1420-1433	0		0	proteome7_184.txt.104
+ F/LNADGVADQLQDQDSN	2	972.011	-0.513	9.5	2.6e-17	1503-1520	1		0	proteome7_184.txt.249
+ F/LANALQNELESE/T	2	688.132	-0.795	8.8	2.45e-14	1677-1688	0		0	proteome7_184.txt.154
+ F/LAIGAGNADLR/R	2	622.999	-0.171	8.4	1.06e-12	1060-1070	0		0	proteome7_184.txt.200
+ F/RELEEEVABQ/Q	2	798.208	-0.358	7.6	6.04e-10	1114-1125	0		0	proteome7_184.txt.73
+ F/AQLELSQVR/Q	2	522.73	-0.434	7.6	9.01e-10	1560-1568	0		0	proteome7_184.txt.89
+ F/LESELTLVADQLLELWLR/N	2	1149.382	-0.303	7	3.45e-8	1729-1748	0		0	proteome7_184.txt.227
+ F/LNADGVADQLQDQDSN	3	648.408	-0.407	6.6	8.46e-7	1503-1520	1		0	proteome7_184.txt.46
+ F/TALLDLSGR/LG	2	567.533	-0.227	6.5	0.00000332	884-894	0		0	proteome7_184.txt.59



Introduction to computational proteomics

Ce-M-M- 189

MS² :: Phenyx search result



Introduction to computational proteomics

Ce-M-M- 190

MS² :: OLAV (L_1)

- Each type of fragment θ has a certain *probability to be detected* in a correct match p_θ and in a random match r_θ
- Depends on the peptide charge z

Dancik, et al., JCB, 1999

	z=2			z=3		
	b	y	y ⁺⁺	b	y	y ⁺⁺
correct	0.57	0.61	0.17	0.35	0.40	0.39
random	0.13	0.11	0.09	0.10	0.10	0.16

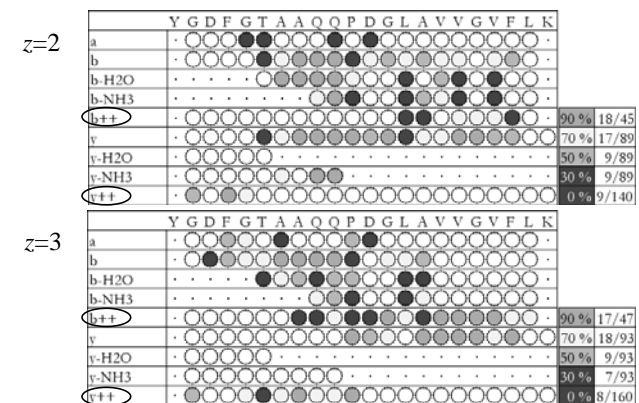
Ion trap instrument



Introduction to computational proteomics

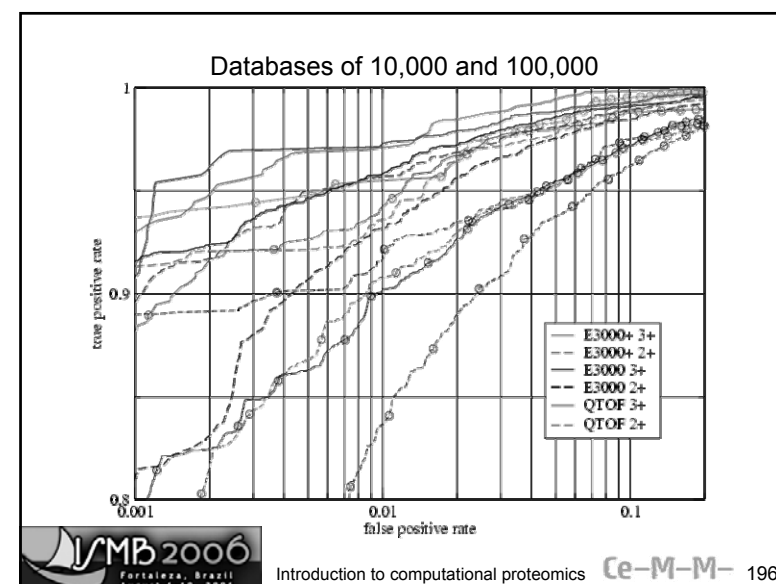
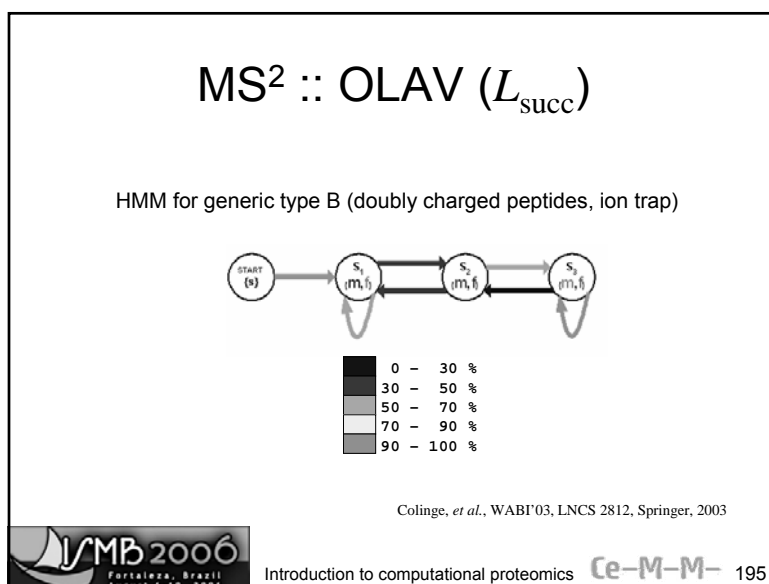
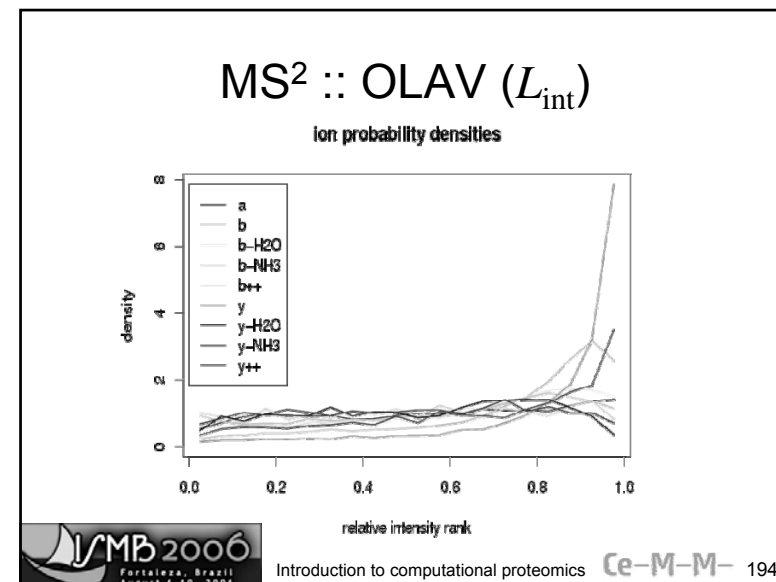
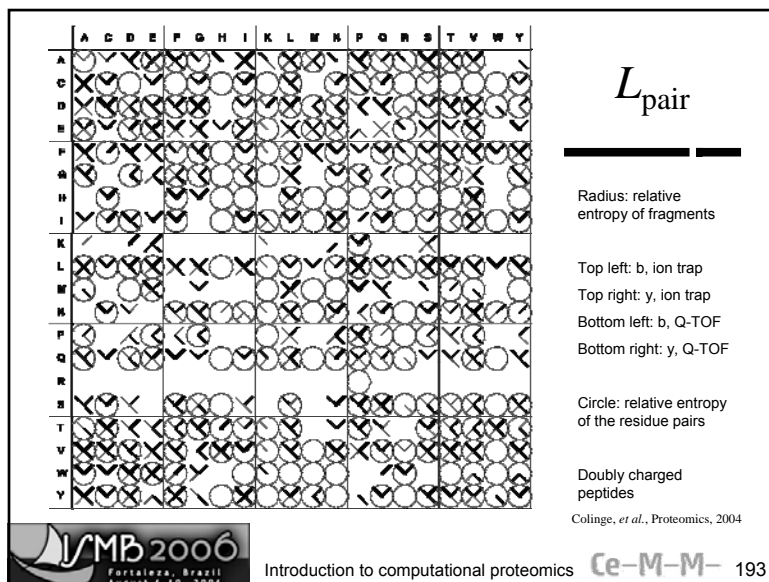
Ce-M-M- 191

MS² :: OLAV (L_1)



Introduction to computational proteomics

Ce-M-M- 192



MS² :: “Limited” comparison

- At high *sensitivity* levels: 90% to 95%
- ESI-IT, ESI-Q-TOF, MALDI-TOF/TOF
- OLAV yields *less than 1% false positive* peptide identifications (database size 10,000-100,000)
- *Improves* over Mascot and post-processed Sequest *by a factor 3-5* at this sensitivity level

Open-source: Craig & Beavis, Bioinformatics, 2004
Geer, *et al.*, J.Prote.Res., 2004

Colinge, *et al.*, Proteomics, 2004



Introduction to computational proteomics

Ce-M-M- 197

MS² :: Back to proteins

- To identify peptides reliably does not yield unambiguous protein identifications automatically.
- Peptides shared by several proteins:



Nesvizhskii & Aebersold, MCP, 2005

- Redundancy and errors in databases



Introduction to computational proteomics

Ce-M-M- 198

MS² :: To score or not to score

- Lists of proteins only by using rules. For instance:
 - Two distinct peptides identified
 - Deals with ambiguities in shared peptides
 - Multiple protein occurrences in distinct LC fractions (protein separation)
 - Peptide tryptic termini
- Computation of a protein score: same “ingredients”.
- Complications: protein length not always known, combine databases.

Cargile, *et al.*, J Proteome Res, 2004
Nesvizhskii, *et al.*, Anal Chem, 2003
Allet, *et al.*, Proteomics, 2004



Introduction to computational proteomics

Ce-M-M- 199

MS² :: Open problems

- Peptide scoring function
- Protein scoring function
- Variants: splice, polymorphism
- Variable modifications
- Database representation (suffix tree)
- Search results representation and integration
- Visualization



Introduction to computational proteomics

Ce-M-M- 200

Others :: *de novo* sequencing

- Goal: to *infer the peptide sequence* (or part of it) from the MS/MS spectrum *directly*
- No database!
- Motivations:
 - *Incomplete databases* for certain organisms
 - *Unexpected modifications*
 - Save search time with large databases (?)
- Types of algorithms: *empirical, optimization, evolutionary computations*

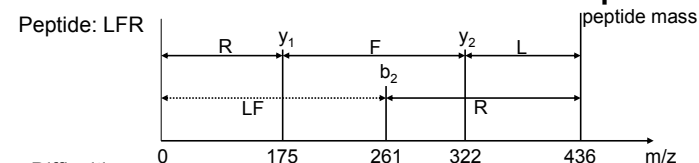
Shevchenko, et al., Anal Chem, 2001
Tanner, et al., Anal Chem, 2005



Introduction to computational proteomics

Ce-M-M- 201

Others :: *de novo* :: Principle



Difficulties:

- We do not know whether a mass is a *N- or C-term fragment*
- Some masses *may be missing*
- b_1 is not detected
- $I=L$, $K=Q$, $F \approx \text{oxi-M}$
- Many pairs of amino acids share the same mass

Solutions:

- Use *N- and C-terminal fragments in a combined manner*
- Focus on *partial safe predictions*: the so-called *sequence tags*
- Sufficient *mass accuracy*



Introduction to computational proteomics

Ce-M-M- 202

Others :: *de novo* :: Heuristic

- Build the peptide sequence *by extending* it one amino acid at a time
- Generally maintain a *population of many candidates*
- Elimination of candidate sequences by a *set of rules* or a direct global comparison with the spectrum

Taylor & Johnson, Anal.Chem., 2000
Tabb, et al., Anal.Chem., 2003



Introduction to computational proteomics

Ce-M-M- 203

Others :: *de novo* :: Optimization

- Several possible *optimization* formulations
- Example: find the optimal path in a graph
 - Nodes are peaks
 - Creates additional nodes by assuming the masses are from C-terminal fragments
 - Vertices when the mass differences are close to amino acid masses
 - Find the best path by using a scoring function
 - Eventually consider sub-optimal solutions as well as partial solutions (sequence tags)

Chen, et al., JCB, 2001; Frank & Pevzner, Anal Chem, 2005; Ma, et al., Rapid Comm.Mass Spectrom., 2003



Introduction to computational proteomics

Ce-M-M- 204

Others :: *de novo* :: Evolutionary computation

- Take a scoring function
- *Optimize over the space of possible peptide sequences*
- Sequence tags: locate reliable regions

Skilling, European patent, 1999
Heredia-Langner, et al., Bioinformatics, 2004



Introduction to computational proteomics

Ce-M-M- 205

Others :: Genome searches

- Direct genome searches
- Motivations: *Incomplete or inaccurate annotations*
- Difficulties: *size and/or spliced peptides (eukaryotes)*
- Approaches:
 - Search gene predictions
 - Peptide *de novo* sequencing + homology search
 - *Adapted database search strategy*
 - Combination

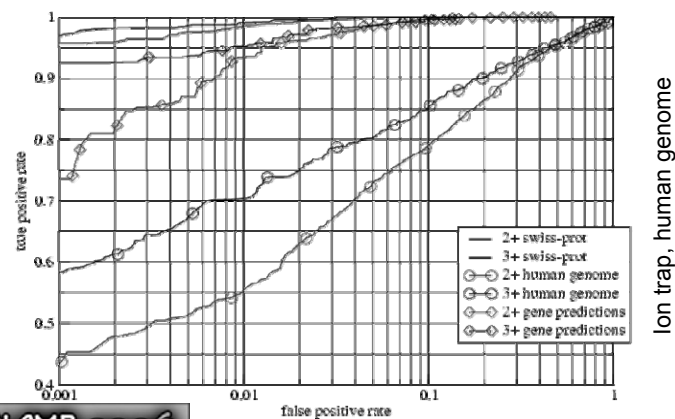
Kuster, et al., Proteomics, 2001
Jaffe, et al., Proteomics, 2004



Introduction to computational proteomics

Ce-M-M- 206

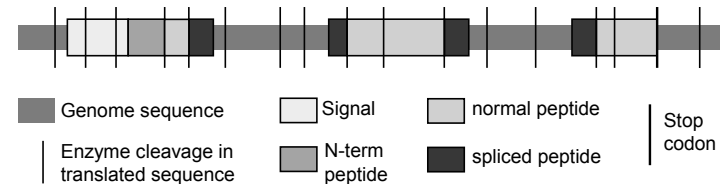
Others :: Genome :: Size



Introduction to computational proteomics

Ce-M-M- 207

Others :: Genome :: Spliced peptides



Introduction to computational proteomics

Ce-M-M- 208

Others :: Genome :: Spliced peptides

- *Predict donor sites and store up-stream sequences of length $< L_{max}$*
- *Predict acceptor sites and store down-stream sequences of length $< L_{max}$*
- *Combine up- and down-stream sequences of donor/acceptor sites at distance $< D_{max}$*
- *Search this virtual database with MS data*



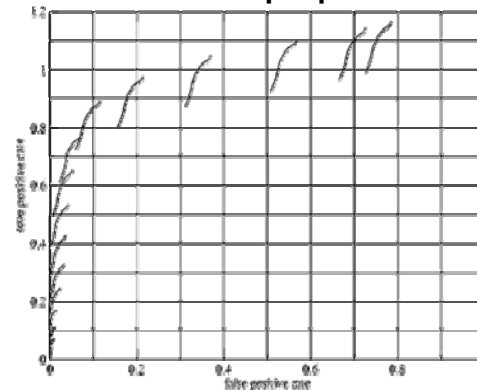
Chen, RECOMB 2001



Introduction to computational proteomics

Ce-M-M- 209

Others :: Genome :: Spliced peptides



Standard genome search followed by Chen's algorithm.

1FP/5TP at
medium sensitivity.

1FP/3TP at high sensitivity.

Several examples
of corrected or
completed genescan
predictions.

Colinge, *et al.* J.Prot.Res., 2005

Introduction to computational proteomics

Ce-M-M- 210

Others :: Genome :: Spliced peptides

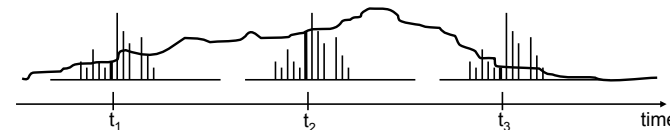
[illegible]

Introduction to computational proteomics

Ce-M-M- 211

Others :: Differential proteomics

- 2DE-Gel images comparisons: spot volumes may provide semi-quantitative information
 - Samples direct mass spectrometry profiles
- Proteomics 2003 3(9); Appel, *et al.*, Electrophoresis
- Ion chromatograms: area is semi-quantitative

Proteomics 2003 3(9): Appel, *et al.*, Electrophoresis, 1997

Wang, et.al, Anal.Chem., 2006



Introduction to computational proteomics

Ce-M-M- 212

Others :: Differential :: Counting peptides

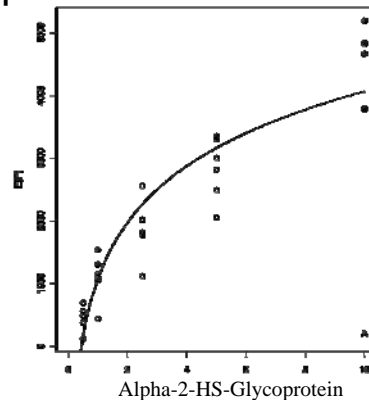
Spiking of 0.5 ml human plasma with purified proteins. Low-nanomolar concentrations.

Add a statistical test.

2.5- to 5-fold changes are detected with 90-95% confidence in human plasma, with 2-3 repetitions.

7.5-10% false positives.

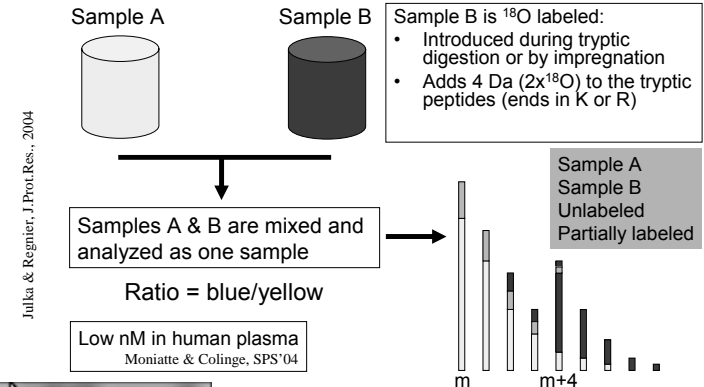
Liu, *et al.*, Anal. Chem., 2004
Colinge, *et al.*, Anal. Chem., 2005



Introduction to computational proteomics

Ce-M-M- 213

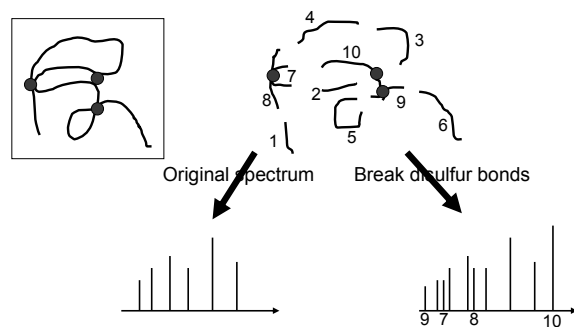
Others :: Differential :: Labels



Introduction to computational proteomics

Ce-M-M- 214

Others :: Structure or modifs



Also applicable to certain PTMs (phosphorylation, glycans, etc.)



Introduction to computational proteomics

Ce-M-M- 215

Others :: Open problems

- Peptide *de novo* sequencing
- Automatic detection of modifications
- Structure elucidation
- Analysis of protein complexes
- Differential expression analysis
- Genome annotation



Introduction to computational proteomics

Ce-M-M- 216

Open source

- Generic Perl library at <http://insilicospectro.vital-it.ch>
- Digestion and mass computations
- Peptide LC elution time predictions
- Graphical display (also LaTeX)
- XML description of atoms, amino acids, modifications, fragment types.



Introduction to computational proteomics

Ce-M-M- 217

End_of_totutorial Acknowledgements

- Alexandre Masselot (GeneBio SA)
- Erik Pitzer (UAS Hagenberg)
- Lydie Bougueleret, Marc Moniatte and former GeneProt colleagues
- Ron Appel & Denis Hochstrasser (Uni Geneva, SIB)

- OÖFH Basisfinanzierung



- GeneBio SA



Introduction to computational proteomics

Ce-M-M- 218