Proteomics and Mass Spectrometry

The exposition is based on the following sources, which are all recommended reading:

1. Aebersold,R. and Mann, M. Mass spectrometry-based proteomics. Nature 422, 198-207 (2003)
2. MacCoss,M.J., Matthews,D.E. Quantitative mass spectrometry for proteomics: Teaching a new dog old tricks. Analytical Chemistry. 77, 294A-302A (2005)
3. de Hoffmann and Stroobant: Mass Spectrometry, Principles and Applications, Wiley

Some history

Mass spectrometry is not a new technology, the foundations were developed more than 90 years ago.

- 1897: English scientist J.J. Thompson discovers the electron (Nobel prize 1906).
- The modern version of the mass spectrometer was devised by Arthur J. Dempster and F.W. Aston in 1918 and 1919, respectively.
- In 2002, the Nobel prize for chemistry was shared by Jonn B. Fenn and Koichi Tanaka (and Kurt Wüthrich) for ”their development of ionisation methods for mass spectrometric analyses of biological macromolecules”.

What is it good for?

The applications of mass spectrometry are manifold. They comprise so diverse fields such as drug testing, pharmocokinetics, space exploration and basic biological research.

This lecture will focus on mass spectrometry-based *proteomics*, i.e. on applications of mass spectrometry to quantify and identify proteins. But mass spectrometry is a general technique and can be used in many applications.

Sample preparation

Mass spectrometry can only measure over a limited mass range. Therefore, proteins are usually *digested* to peptides using a protease (usually Trypsin).

Other common preparation steps are addition of Urea (to break up non-covalent protein bounds) or the removal of highly abundant proteins (that might ”suppress” proteins of lower abundance).

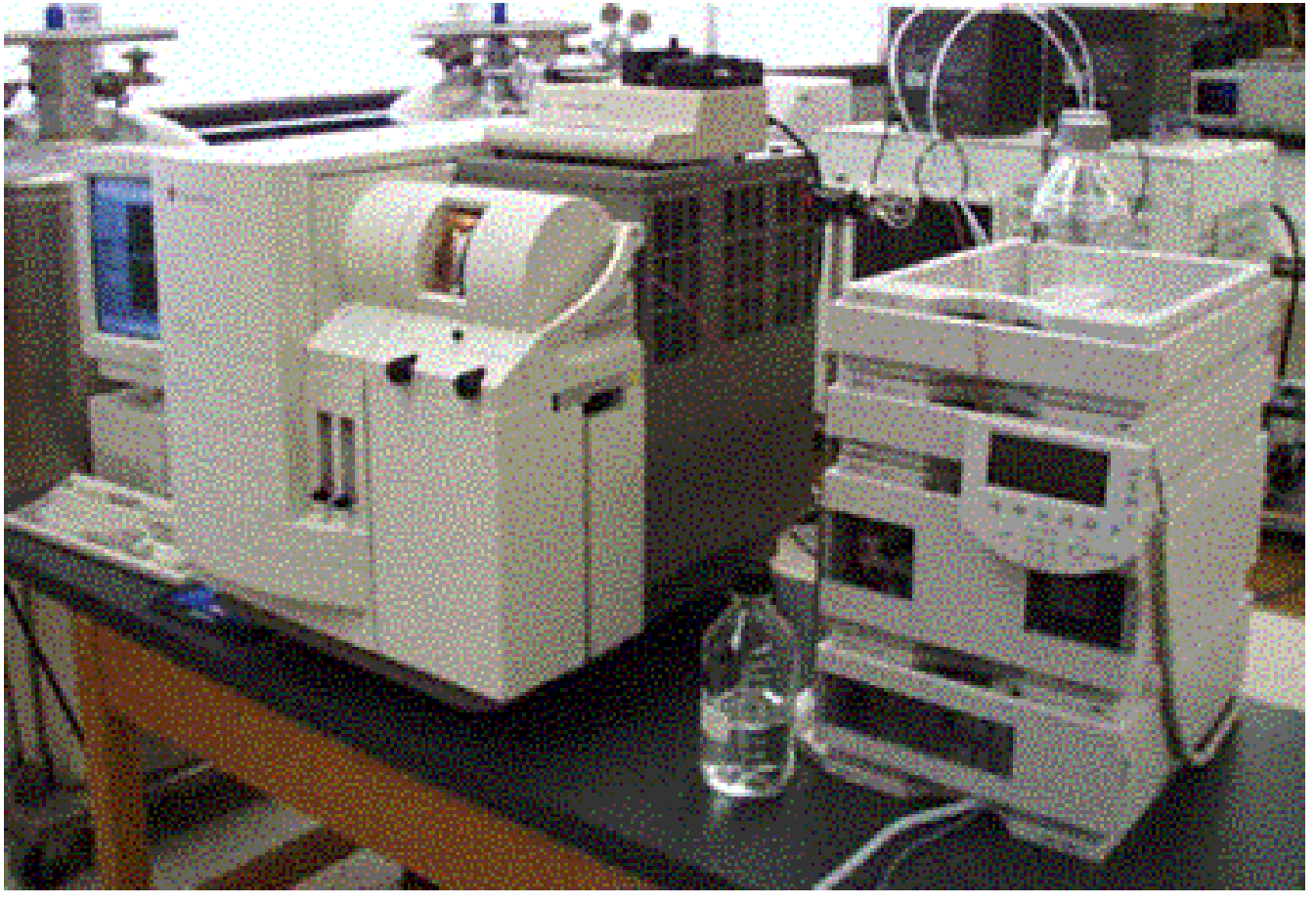
Technology of mass spectrometers

Mass spectrometers are devices that measure the mass of ions relative to their units of charge that is the *mass over charge* m/z.

They usually contain the following elements:

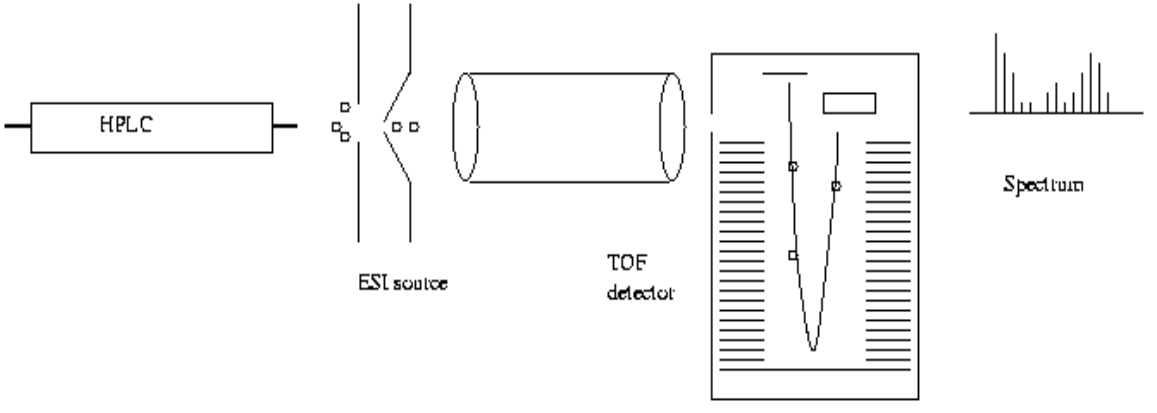
1. a device to introduce the compound (e.g. a chromatograph) into the analyzer
2. a source to produce ions from the compound
3. one or more analyzers to separate the various ions according to their m/z
4. a detector to count the ions emerging from the last analyzer
5. a computer to process the detector data.

Technology of mass spectrometers (2)



Technology of mass spectrometers

A schematic overview:



Using this setup, we can perform a (relative) quantification of the peptides in a sample.

It is possible to combine several steps of mass spectrometry (LC-MS/MS, LC-MS/MS/MS etc.). This allows us to measure the masses of peptide fragments and thus, their amino acid sequences (more details in the next lecture).

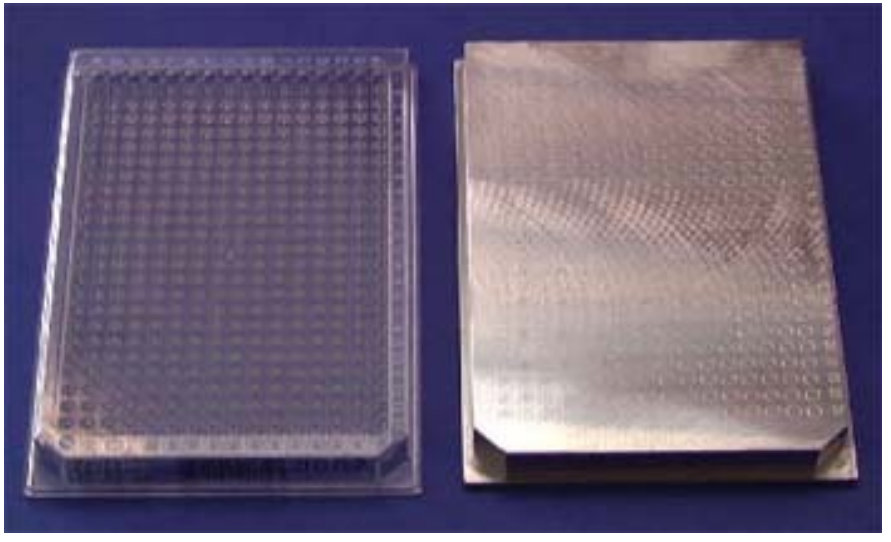
Introducing the analyte

The most common device (for protein MS) to introduce the compound into the analyzer is *HPLC* (high performance liquid chromatography). The peptides are separated while traveling through the chromatographic *column* depending e.g. on their hydrophobicity.

The analytes can then directly be subjected to the ionization phase or they can be *spotted* on a *MALDI target* (Matrix Assisted Laser Desorption Ionization).

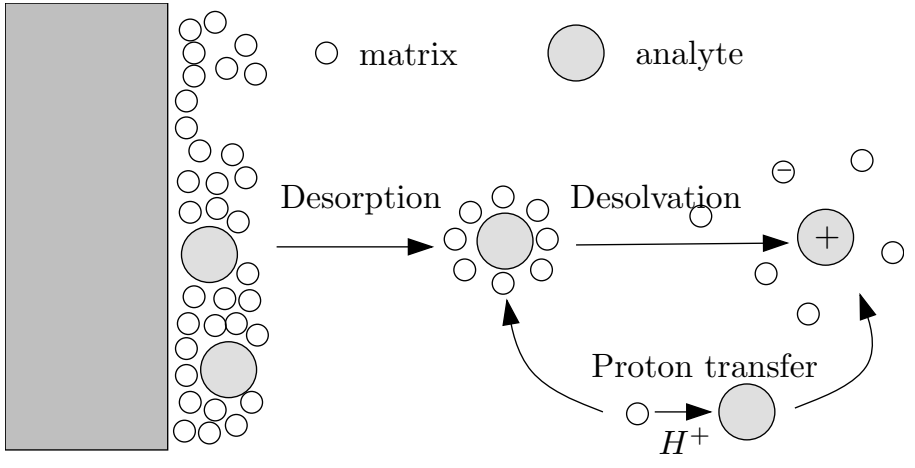
Then the compound is ionized before entering the analyzer. The two most common devices (for protein MS) to ionize the compounds are *MALDI* and *ESI* (Electrospray Ionization).

Introducing the analyte (2)



Ionization using MALDI

In MALDI the compound is mixed in a solvent containing small organic molecules in solution called *matrix*, which has a strong absorption at the laser wavelength. The mixture is then dried and results in the compound molecules being completely isolated in the matrix. Then laser pulses are shot at the mixture which results in rapid heating of the matrix and its expansion into the gas phase. The analyte molecules are set free (and stay intact) and are ionized. The process is not completely understood, but involves probably gas phase proton transfer. Usually the charge is 1 in MALDI.

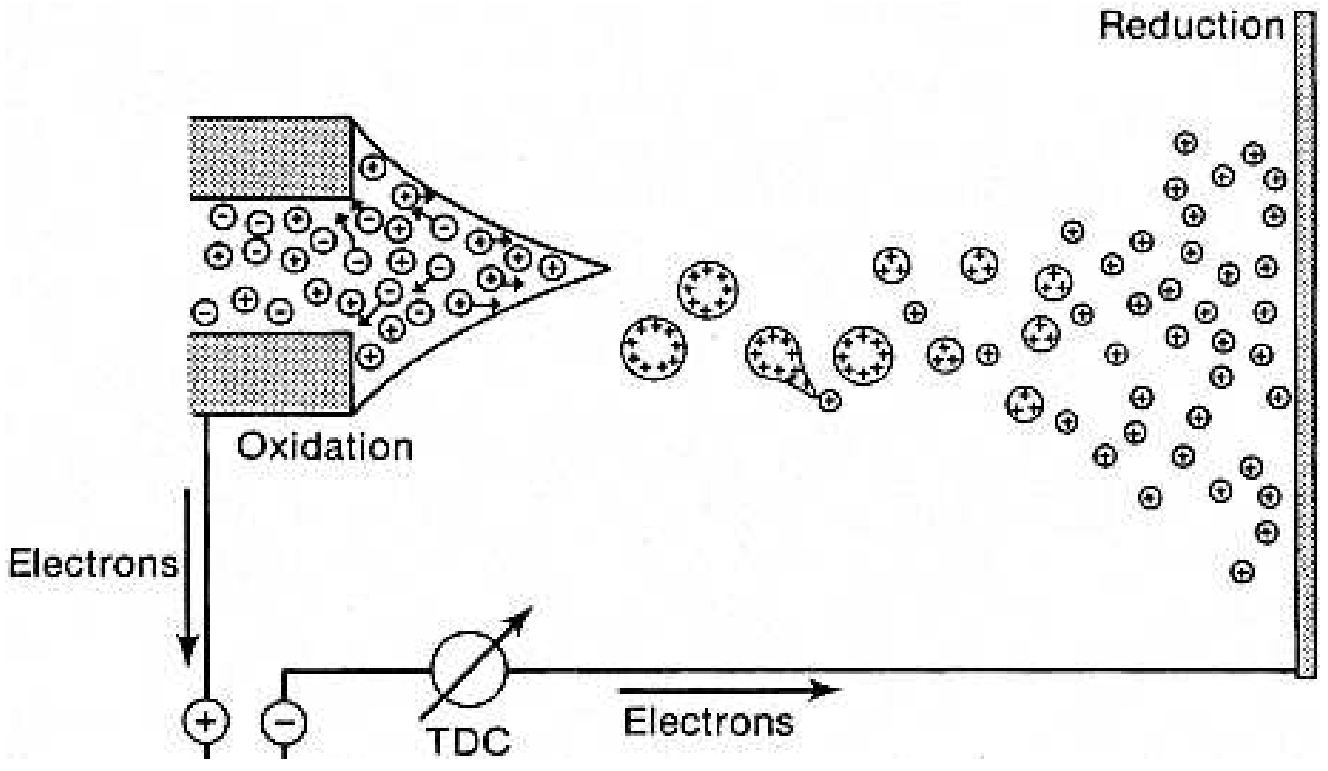


Electrospray Ionization (ESI)

In ESI, the compound is ionized at atmospheric pressure which increases the ionization efficiency greatly but poses problems at coupling the ion source to the low pressure analyzer compartment.

An electrospray is produced by applying a strong potential difference ($3 - 6kV$) to a capillary through which the liquid containing the analyte flows. This results in an electric field of the order of $10^6 V/m$. The field causes charge accumulation at the surface of the liquid which will break to form highly charged droplets.

The droplets are then dried leaving one or more (usually less then six) charge units on the analytes.



If MS is used to identify peptides, both MALDI and ESI are used. If a quantitative measurement is required, ESI has some advantages.

Mass analyzers

There are various types of mass analyzers like *quadrupole* analyzers, *quadrupole ion trap* analyzers, *time-of-flight* analyzers, *ion cyclotron resonance and fourier transform* analyzers etc.

We shortly describe the time-of-flight (TOF) analyzer. Here the ionized analyte with charge ze ($e = 1.6 \cdot 10^{-19}$ coulomb) is accelerated by a potential V_s and flies a distance d through the chamber (field-free) until it hits the detector. The analyzer measures the time until the ion hits. Then it holds that

$$t^2 = \frac{m}{z} \cdot \left(\frac{d^2}{2eV_s}\right)$$

The $\frac{m}{z}$ of an ion stands in a quadratic relationship to the flight time through the chamber. The lower the m/z of an ion, the faster it reaches the detector.

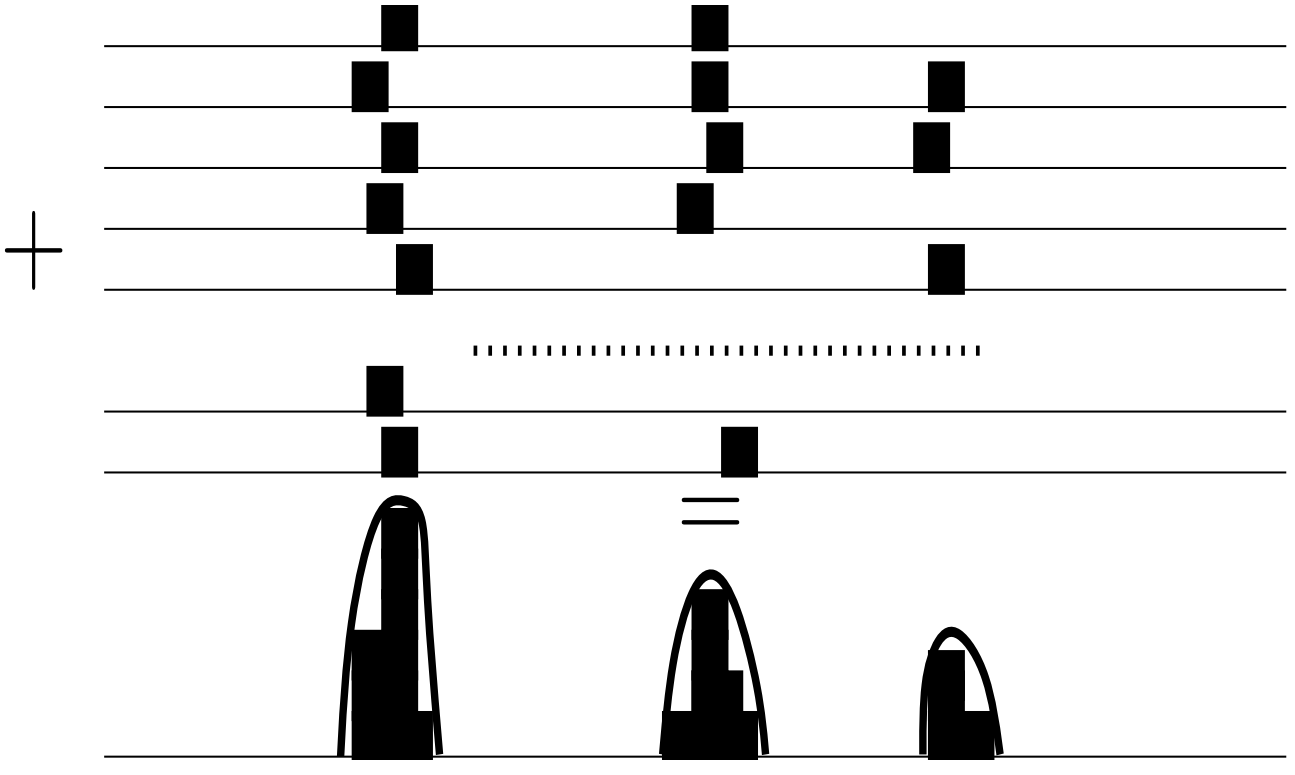
Detectors

As detectors usually *photographic plates*, *faraday cylinders*, or *array detectors* are used in conjunction with electron or photon multipliers to increase the intensity of the signal.

What is interesting for us is that certain detectors, need some time to recover after an ion hit (called *dead time*) until it can measure the next ion. Hence there is always a *suppression effect* that needs to be considered.

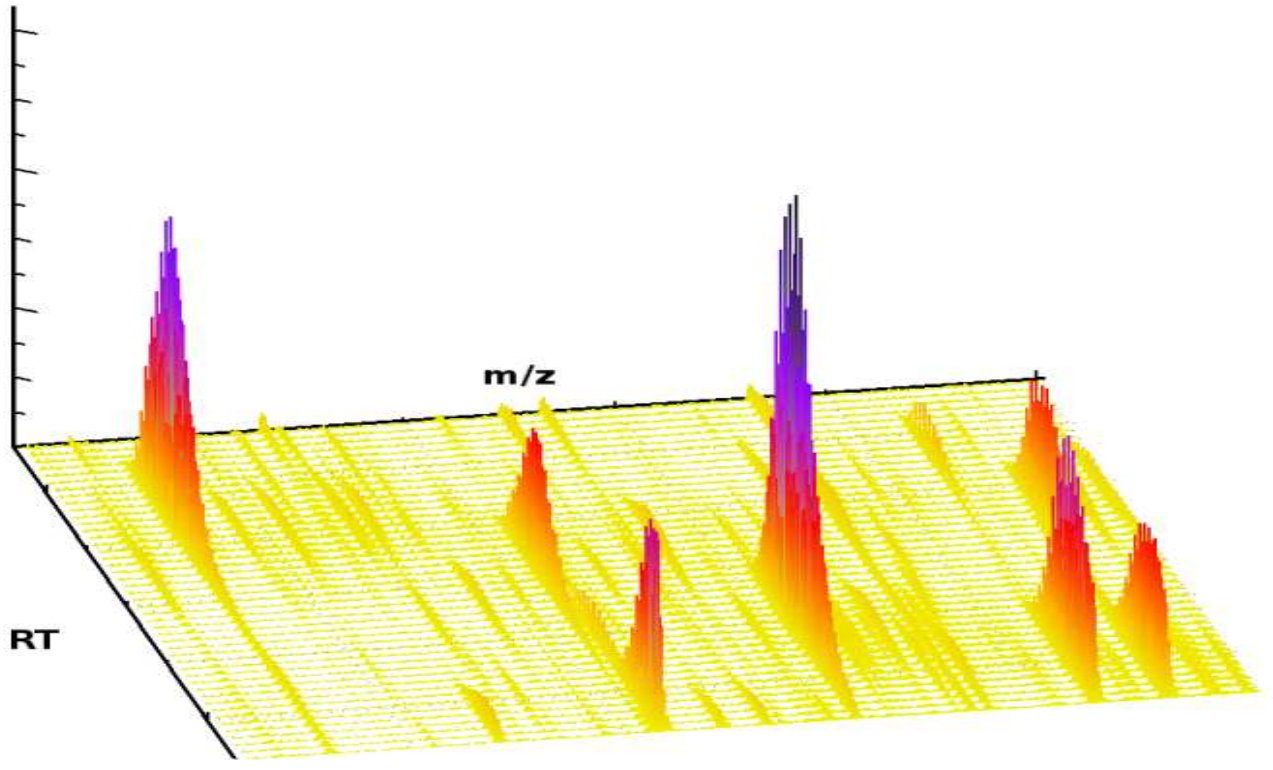
The machine software takes with a certain frequency measurements memorizing the time an ion hit (i.e. its m/z measured in Thomson or Dalton, $1Da = 1.665402 \cdot 10^{-27}kg$) and combines these measurements into a so called *scan* or *spectrum*. These raw spectra are for all practical purposes the ”rawest” data we see.

Detectors (2)



Due to the abovementioned inaccuracies in the mass measurement, we do not see exactly discrete signals in a mass spectrum, but rather Gaussian-shaped peaks as pictured above.

LC-MS raw data



Remember that we are talking about *LC-MS* (Liquid Chromatography coupled to Mass Spectrometry). That is, we do not record only a single scan for our sample but many (several thousands), one for each subfraction of the sample as it elutes from the column.

This two-dimensional data is what we call a *LC-MS map*.

What comes next ?

Data analysis ! For each peptide in our sample, we want to know its mass, charge, elution time and abundance (i.e. signal strength).

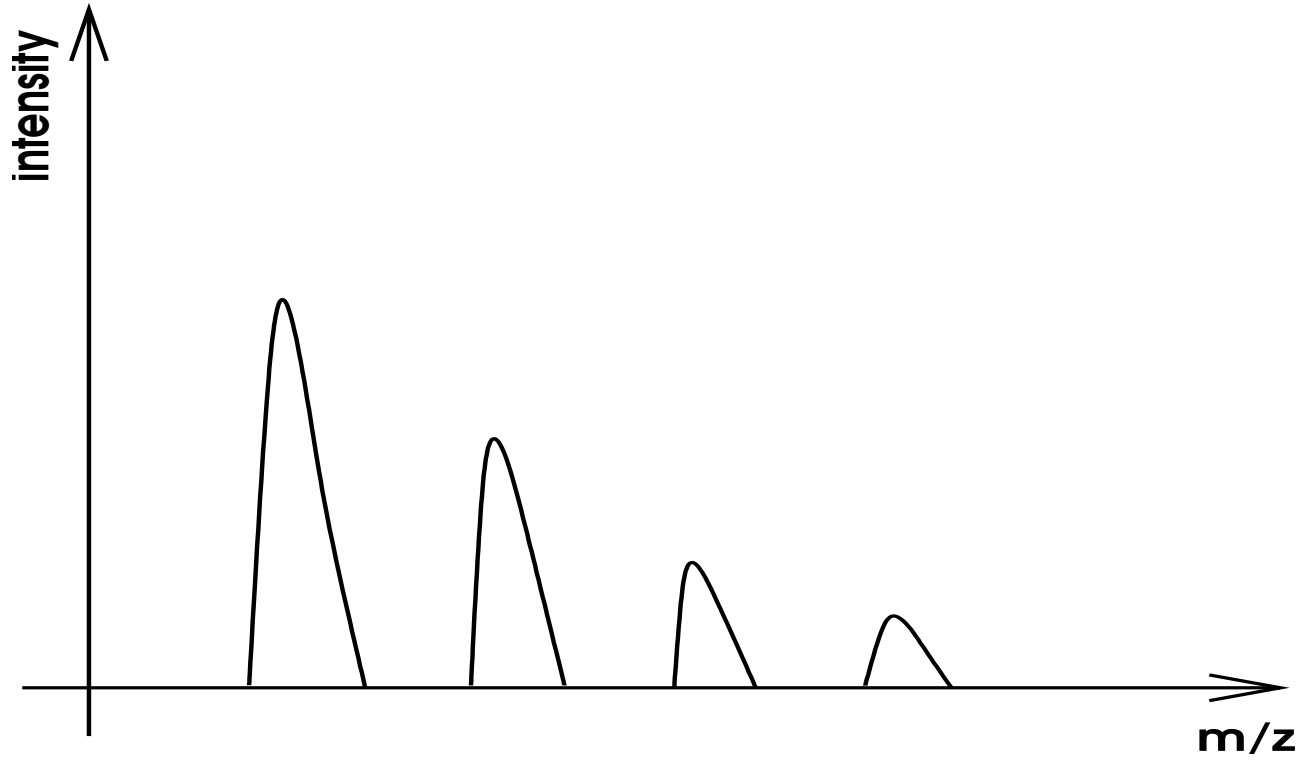
Important steps of LC-MS data analysis are:

- Low-level preprocessing: baseline reduction, mass calibration
- Peptide feature detection: detect and extract peptide signals in LC-MS data
- Comparison of samples: alignment, statistics.

We will only sketch each of these steps, more details follow in advanced courses....

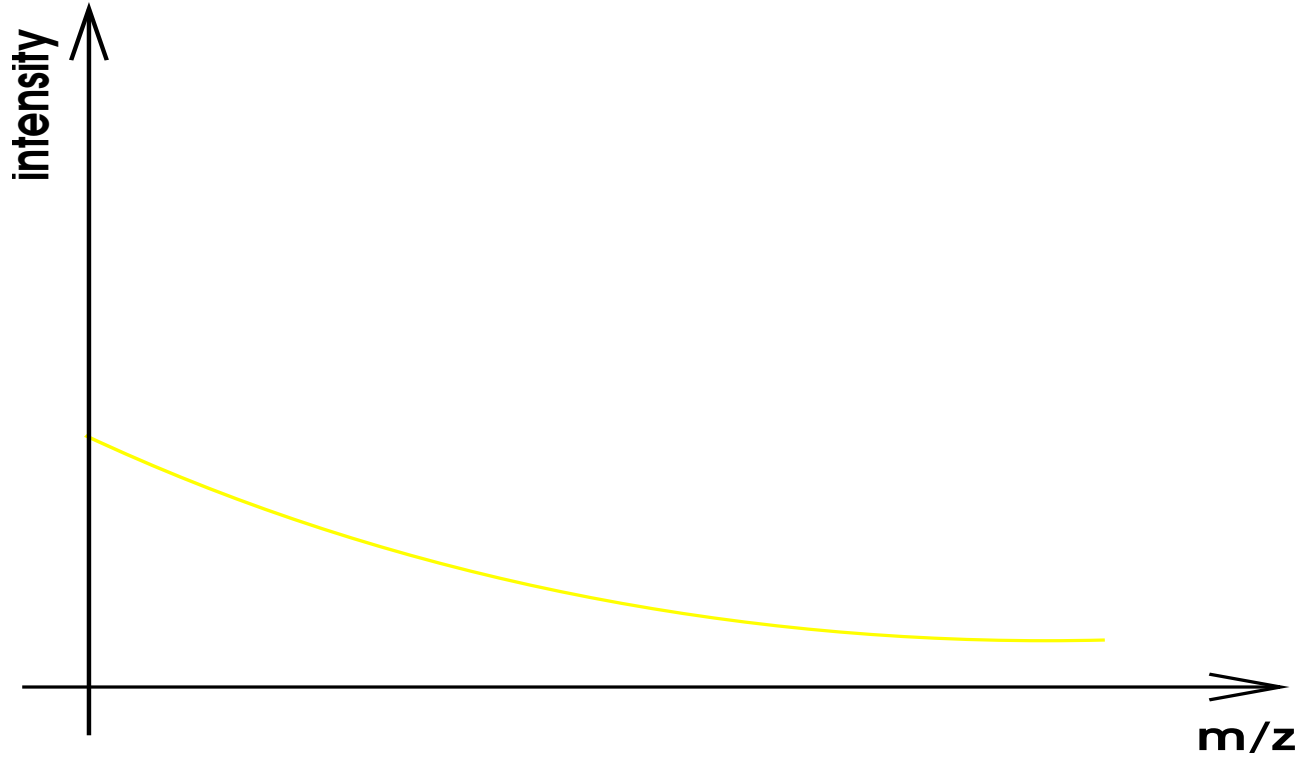
Preprocessing data

But what does a raw spectrum contain? As we argued it contains the *signal*



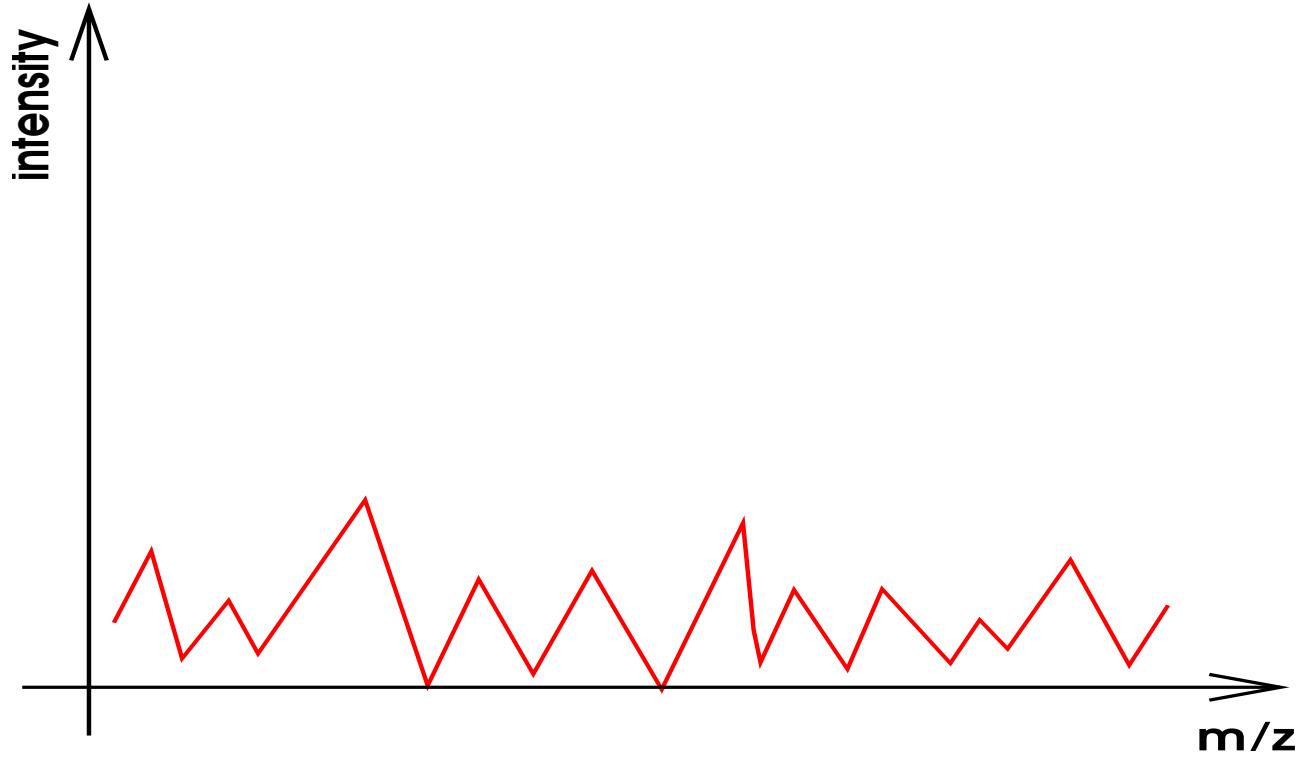
Preprocessing data (2)

It contains a *baseline* of the spectrum which represents the level of 0 intensity measured by the instrument. The baseline is caused by chemical noise in the MALDI matrix or by ion overloading.



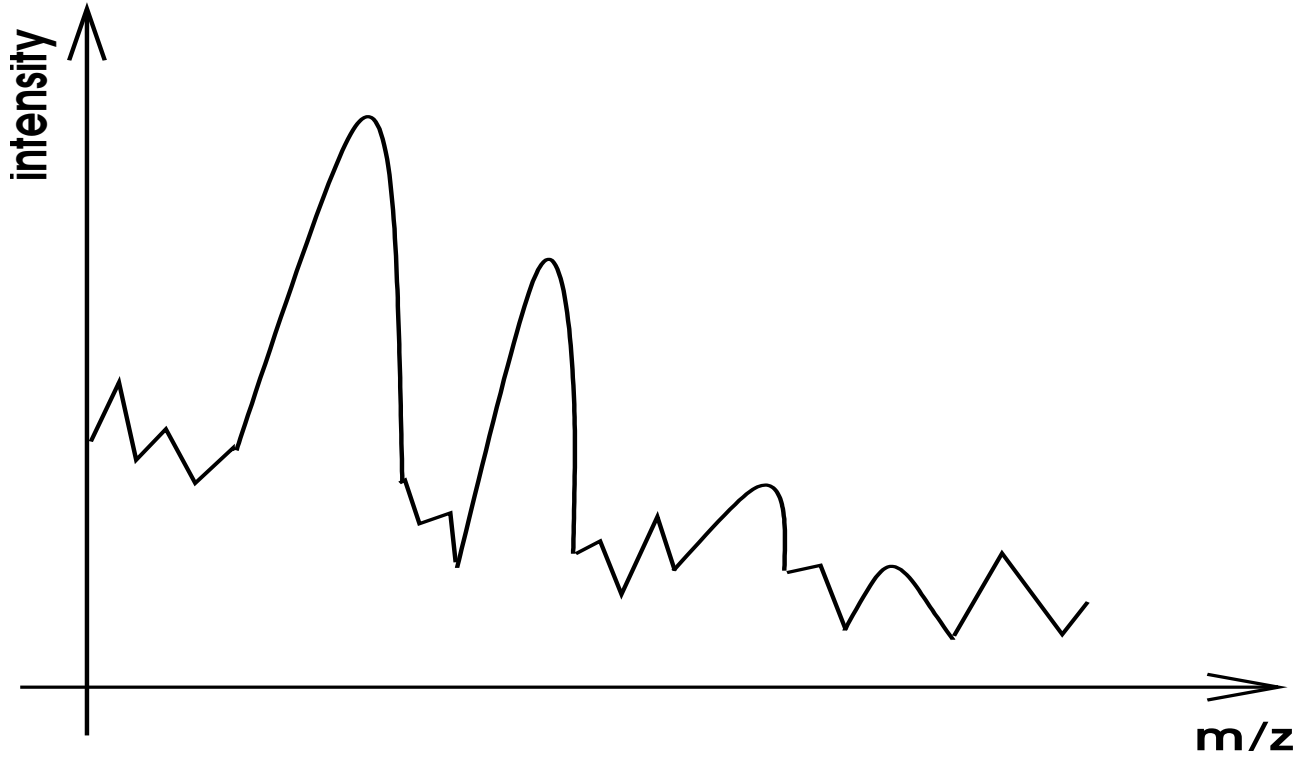
Preprocessing data (3)

In addition we have some *random* and *chemical* noise.



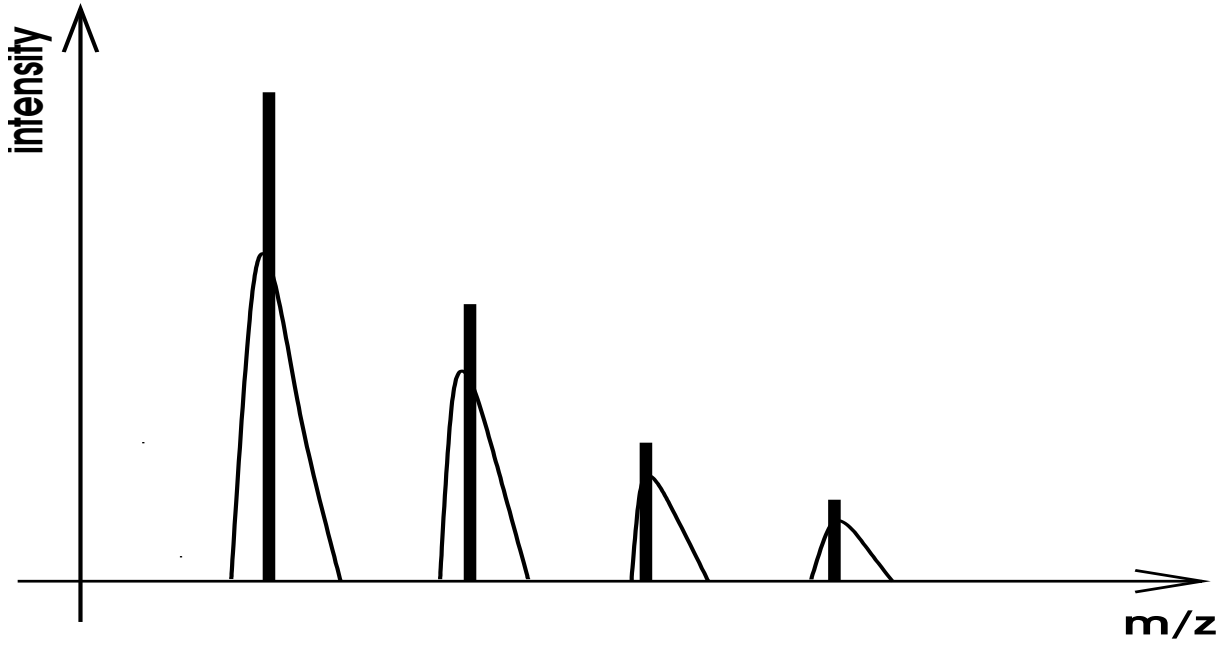
Preprocessing data (4)

Adding all these signals together we have the raw spectrum.



Preprocessing data (5)

The goal is to convert this spectrum into a *stick spectrum* where each stick corresponds to a signal peak and baseline and noise are removed.



Preprocessing data (6)

The conversion of the raw spectrum into a stick spectrum can be done by the software of the mass spectrometer. But it makes sense to do it by yourself if you really want to know what your data contains.

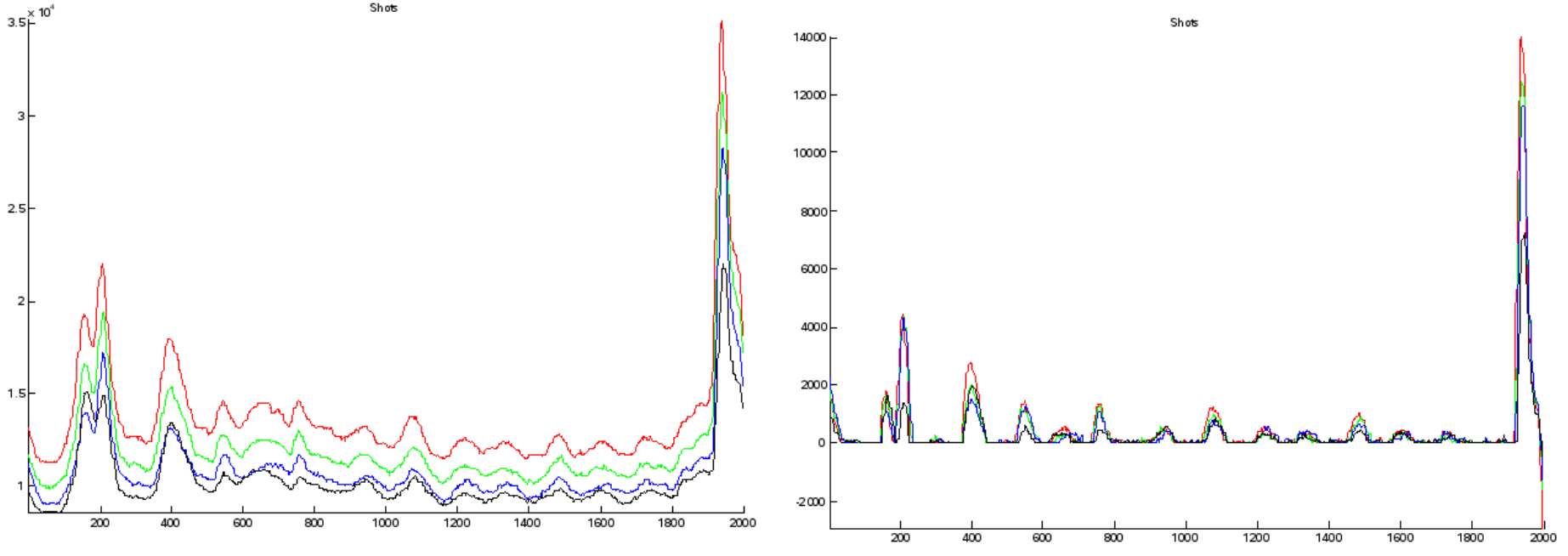
The conversion raw to stick spectrum includes the following steps:

1. Baseline reduction
2. Calibration
3. Peak Detection

We give now a few examples of these steps.

Baseline reduction

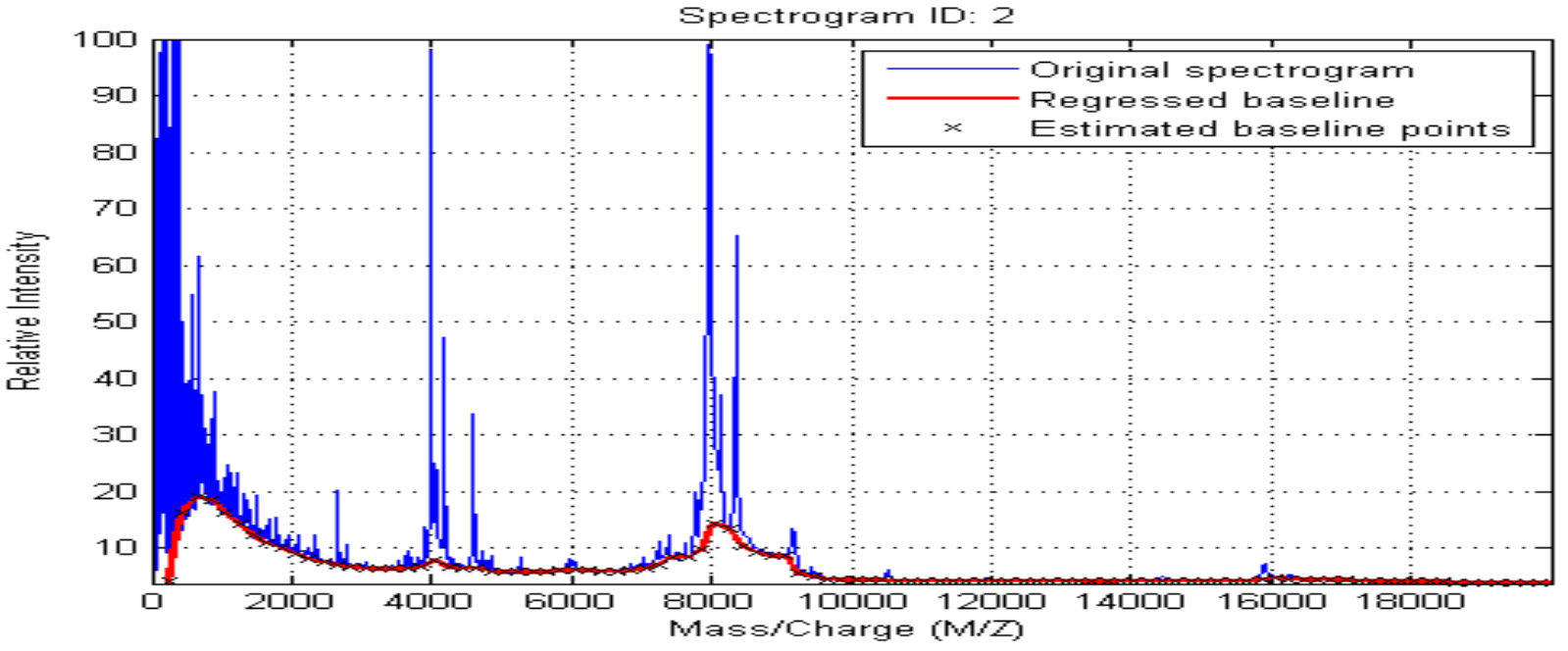
On possible way for eliminating unwanted structures in a signal is to use mathematical morphology, which is the analysis of spatial structures. The basic idea of a morphological filter is to inhibit selected signal structures. Such structures could be noise or some irrelevant signal structures like the baseline.



The so-called *top-hat filter* estimates a baseline by successively applying two morphological operations (dilation and erosion).

Baseline reduction (2)

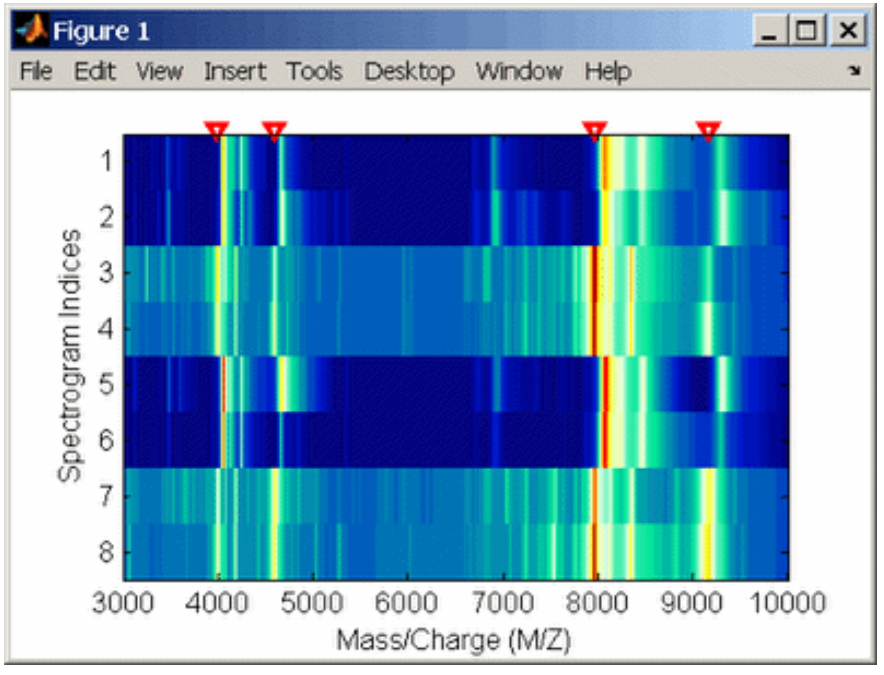
Another possibility is to estimate baseline using (robust) regression.

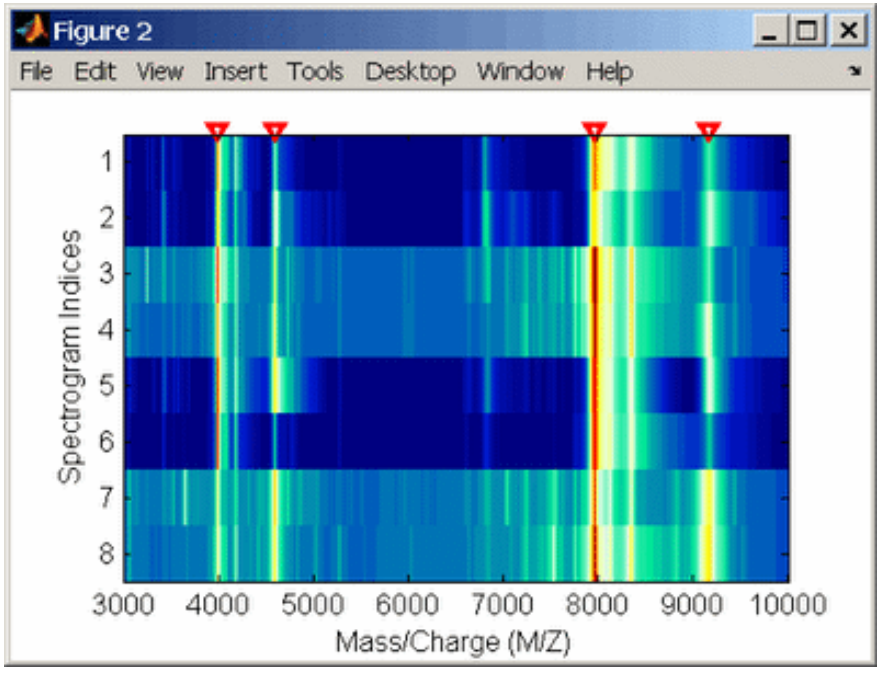


Plot taken from `http://www.mathworks.com/`. To remove the baseline, the algorithm simply subtracts the estimated baseline from the signal.

Calibration

Another problem that occurs in the measurement is that the data might be *uncalibrated*, i.e. the peaks have a systematic shift that can often be described as an affine function.





(Figures from: http://www.mathworks.com/)

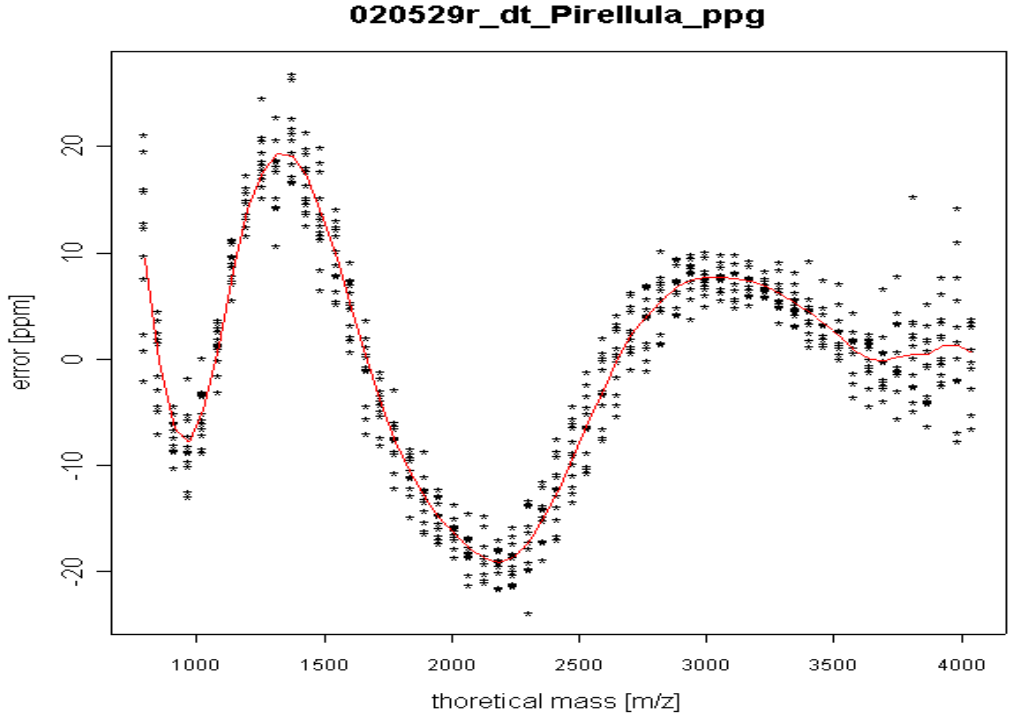
These figures show an alignment of 8 measured spectra against 4 reference masses.

Calibration (2)

The algorithm illustrated by these figures is as follows: First, a synthetic spectrum is built with Gaussian pulses centered at the masses specified by the reference mass vector. Then for each measurement individually, an affine map is determined so as to maximize the *cross-correlation* between the measured spectrum and the synthetic spectrum. The rescaled spectrum is calculated by piecewise cubic interpolation of the data points and shifting. This method preserves the shape of the peaks.

Calibration (3)

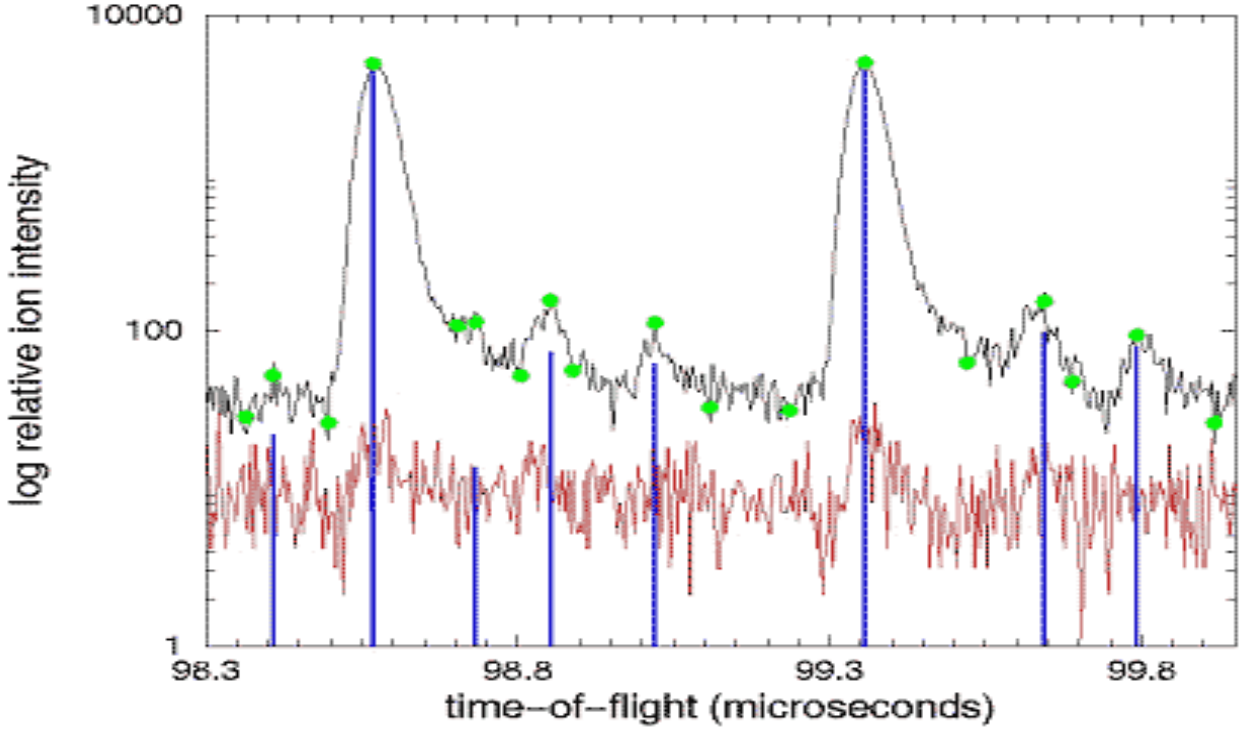
But also nonlinear calibration errors are possible. The following figure shows the dependency of the remaining error *after* affine calibration for some peptide mass fingerprint data from *Pirellula*.



(Figure taken from Eryk Wolskis website.)

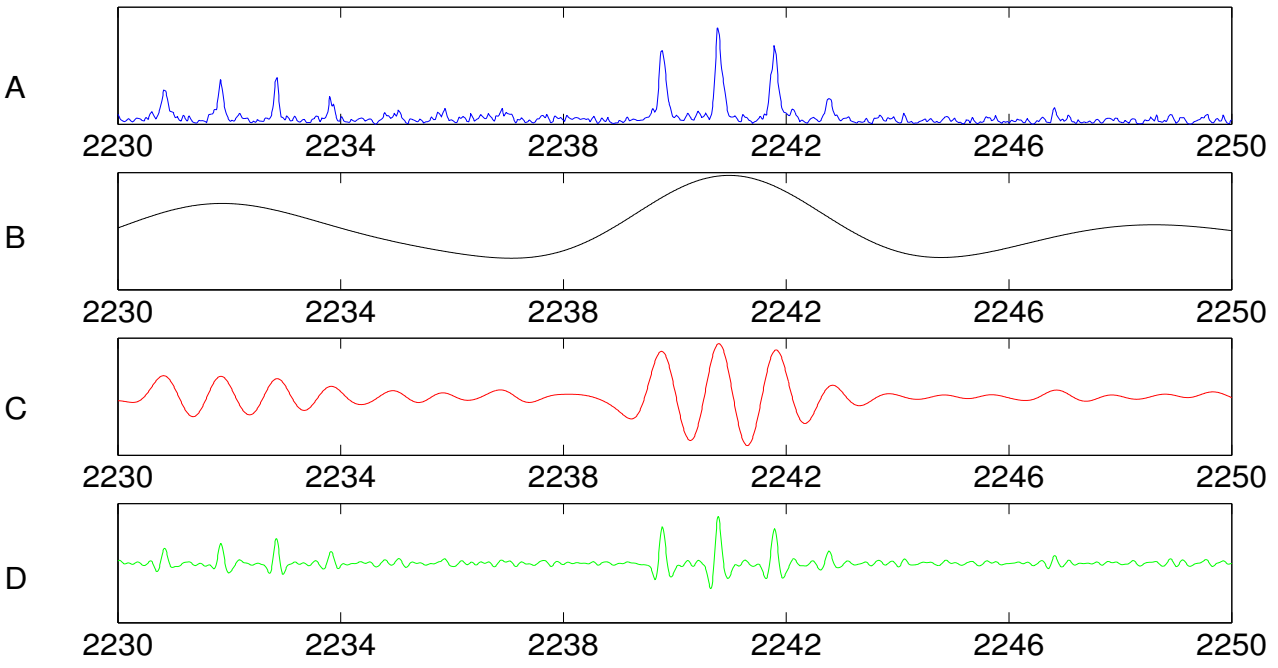
Peak detection

The peak detection routine has now to determine the centroid of the signal peak and subsequently integrate the area beneath it, since this an approximation of the total ion count.



Again, there are numerous algorithms to perform this step. Most of them are based on some sort of pattern-matching or filtering (e.g. using wavelets) since we know more or less the ideal shape of a peak.

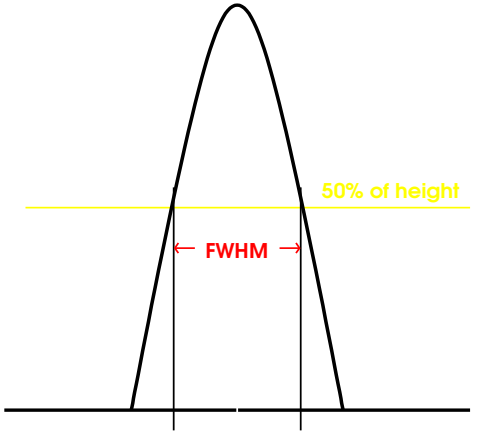
Peak detection (2)



Wavelet decomposition of a mass spectrum. From top to bottom: mass spectrum (A) and wavelet-transformed signal for different scales of the wavelet (B,C,D). As you can see, different scales pronounced different frequency ranges of the spectrum. By computing a wavelet-transform with a scale matching a typical peak shape, we can pronounce peaks in the spectrum and suppress noise structures.

Peak detection (3)

Most algorithms store some meta information with each peak such as the centroid m/z and the intensity (=area under the peak). Usually the software also computes some more characteristic values of each peak and stores it, like the height of the peak, its full width at half max (FWHM), etc.



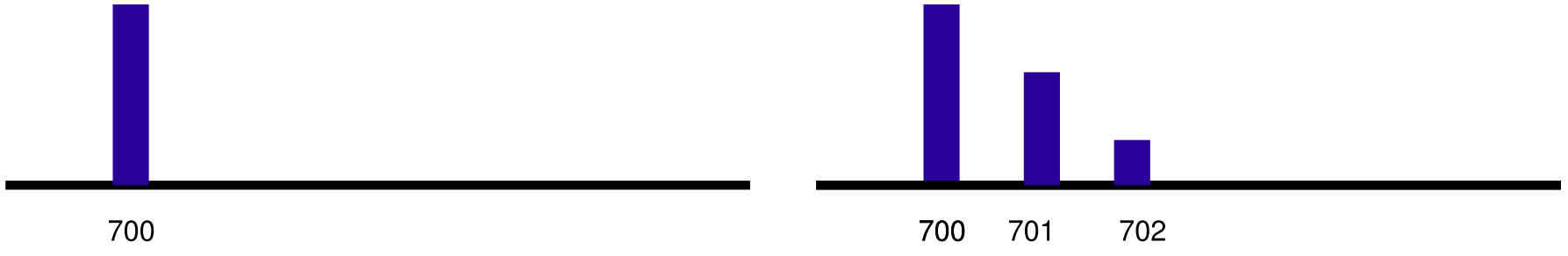
Mass Spectrometry of Proteins

Remember that we are actually interested in the peptides in our sample. It would help us if we could model the typical pattern that a peptide causes in a mass spectrum. Using such a model, we could separate peaks caused by peptides from noise peaks.

But first, we need to repeat some basic chemistry...

Mass Spectrometry of Proteins

Consider a hypothetical peptide of mass 700.



You would expect to see one peak at 700 m/z (left).But instead you will several peaks, all caused by the same peptide (right).

Quiz: How can we determine the charge state of a peptide feature?

Mass Spectrometry of Proteins (2)

Peptides are almost exclusively made out of the elements nitrogen (N), oxygen (O), carbon (C), and hydrogen (H).

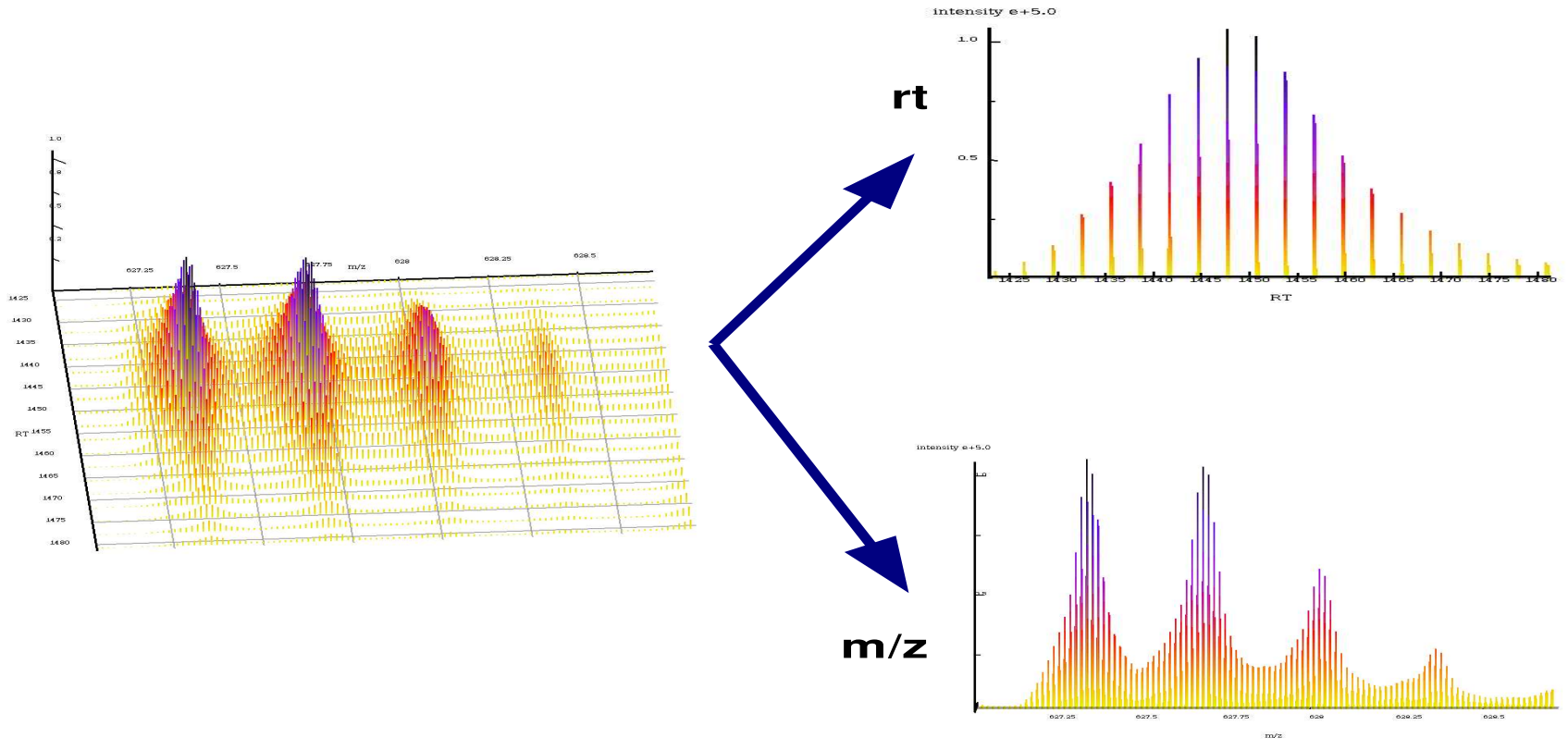
All these elements occur in nature in *different isotopes* that differ in their mass. The following table gives all masses of atoms in amino acids and their isotope proportions:

Mass Spectrometry of Proteins (3)

sym	isotopic mass	Rel. %	avg. mass
C	12.000000 13.003355	98.9 1.112	12.011
H	1.007825 2.014	99.985 0.015	1.00794
N	14.003074 15.000108	99.63 0.37	14.00674
O	15.994915 16.999133 17.999169	99.76 0.04 0.20	15.9994
S	31.972970 32.971456 33.967866 35.967080	95.03 0.75 4.22 0.02	32.066

Given the *atomic composition* of a molecule we know its numbers n_c, n_h, n_n n_o and n_s of C, H, N, O, and S atoms.

Isotope Patterns



A peptide feature and its projections on m/z and rt. Note that a peptide usually elutes over several scans hence its isotopic pattern will appear in several scans.

Mass Spectrometry of Proteins (2)

As you can see, C-12 is the most common isotope of Carbon. The second isotope, C-13, occurs roughly with probability 1%.

Let us assume that our hypothetical peptide with mass 700 has 12 C Atoms. What is the probability of seeing a single C-13?

$$b(12, 1, 0.01) = \binom{11}{1} \times 0.01 \times (0.99)^{10} \quad (1)$$

One copy of C-13 will shift the peak by 1. But each atom in the peptide might exist in a heavier state and multiple copies of a peptide will result in an *isotopic distribution*.

Isotope Patterns

The full isotope pattern of a peptide can be computed using the following formula:

The rank i peak of an isotope-cluster (peak r_i) is the peak with i Da of additional mass compared to the monoisotopic peak. Given the number of C, N, O and S atoms in a peptide, the relative height of r_i can be computed using a *binomial convolution*.

Isotope Patterns (2)

Consider a list L of isotope contributors (i. e. ^{13}C, ^{15}N, ^{17}O, ^{18}O, the influence of S and H isotopes is negligible). For each element $l \in L$, let p_l be the frequency of occurrence of an isotope, w_l be the integer offset weight of the isotope (1 for ^{13}C, 2 for ^{18}O, etc.), and n_l be the number of atoms.

Let P denote the joint isotope distribution, where $P(L, k)$ is the probability of seeing peak r_k, given a list L of isotope contributors. Then it holds

$$P(L,k) = \sum_{l \in L} \sum_{\substack{i=0, \\ i \% w_l = 0}}^{k} b\left(n_l, \frac{i}{w_l}, p_l\right) P(L - \{l\}, k - i). \tag{2}$$

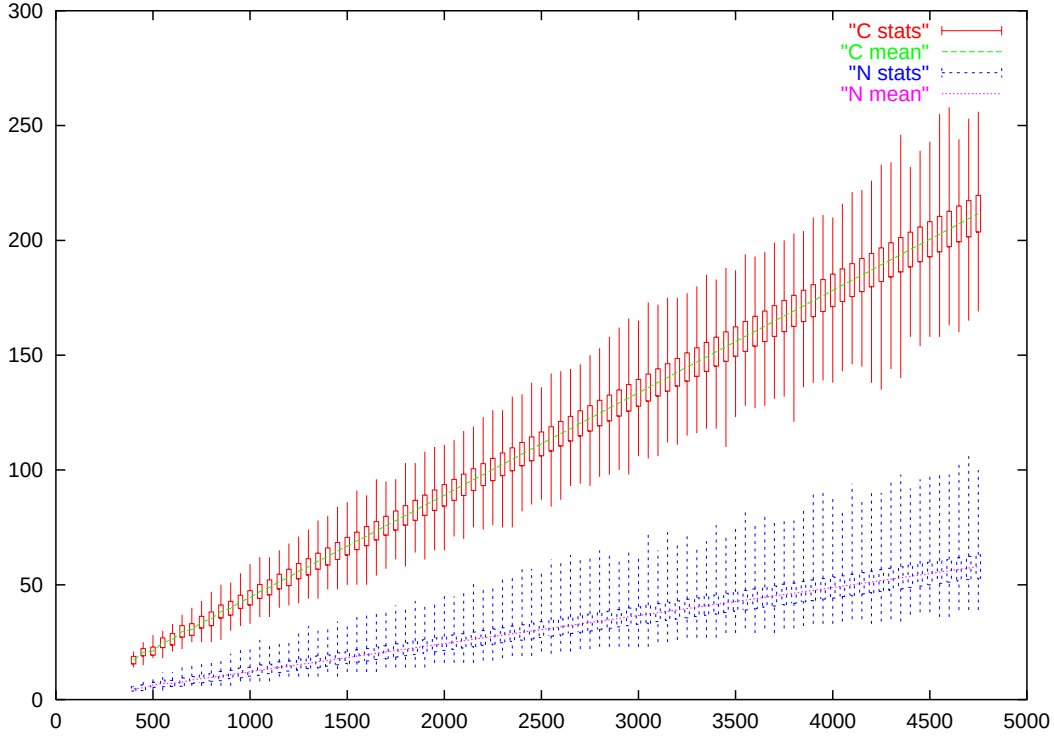
Isotope Patterns (3)

Note that we are making the approximation that the differences in masses between different isotopes are integers. This is not quite true, but in practice we compensate by looking for the peak at difference k over a range $(k - \delta, k + \delta)$.

So we can compute the isotope pattern for a peptide if we know its atomic composition. But for feature detection is MS spectra, this does not help much since for a given signal in a spectrum, we don't know its atomic composition.

Averagines (4)

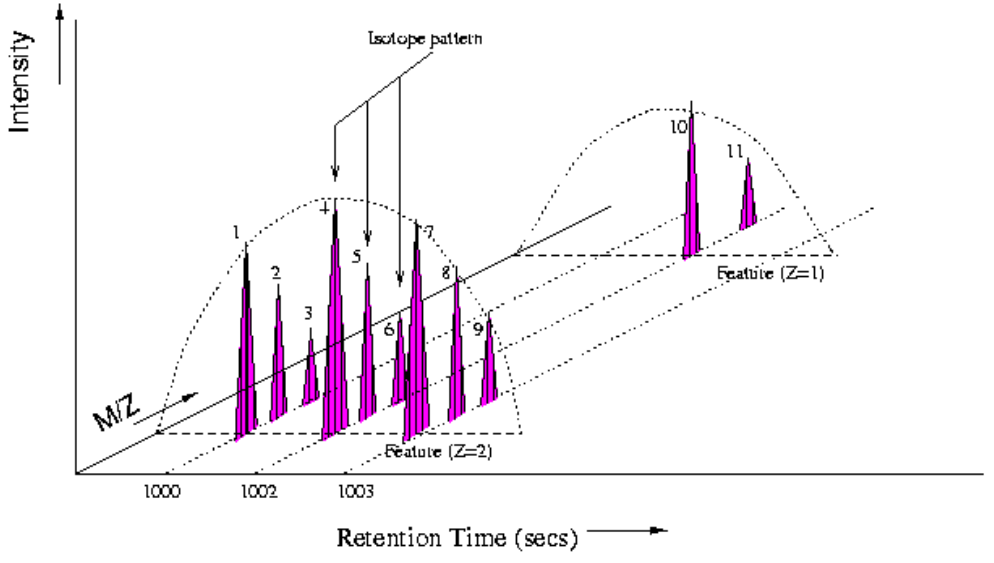
If one plots the atomic content of proteins in some protein database (e.g. SwissProt) it becomes evident, that the number of atoms for each type grows roughly linearly. The picture shows on the x-axis the molecular weight and on the y-axis the number of atoms of a type.



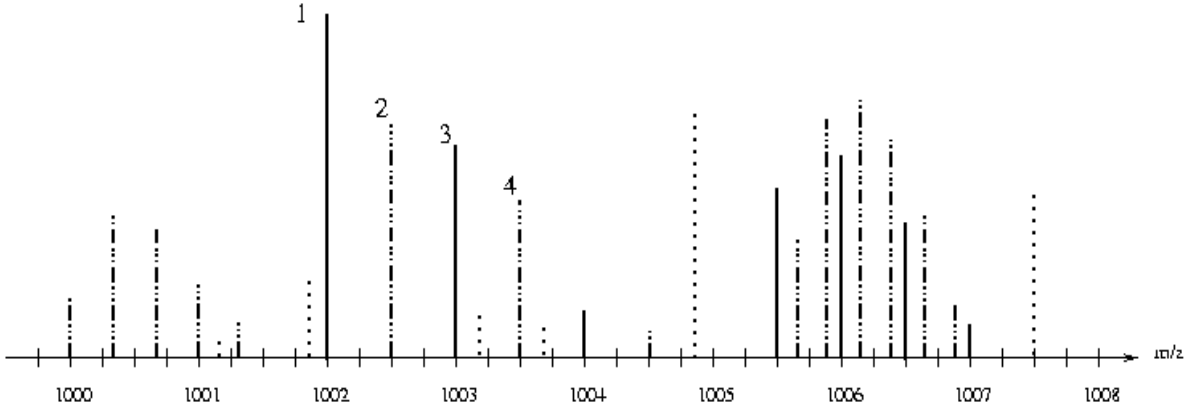
Averagines (5)

Since the number of C,N, and O atoms grows about linearly with the mass of the molecule it is clear that the isotope pattern changes with mass.

mass	P(k=0)	P(k=1)	p(k=2)	p(k=3)	p(k=4)
1000	0.55	0.30	0.10	0.02	0.00
2000	0.30	0.33	0.21	0.09	0.03
3000	0.17	0.28	0.25	0.15	0.08
4000	0.09	0.20	0.24	0.19	0.12



Note that the isotope pattern is dependent on the mass *not* on the m/z. It can be used to distinguish peptide content from non-peptide content and to deconvolve the signals.



This is a spectrum with interleaved peptide peaks and noise. In this example, real peptide peaks are solid or dashed. Noise is dotted. Peaks 1 to 4 belong to two interleaved features of charge 1 with molecular weights of 1001 and 1001.5 Da, respectively.

Peptide feature detection

Where do we stand ?

- Given the atomic composition of a peptide, we can compute the relative intensities of its peaks in a mass spectrum (the isotopic pattern).
- Since there is a very nice *linear relationship* between peptide mass and its atomic composition, we can estimate the average composition for peptide of a given mass.
- We can use this knowledge for *feature detection* i.e. to summarize isotopic pattern into peptide features and to separate them from noise peaks.

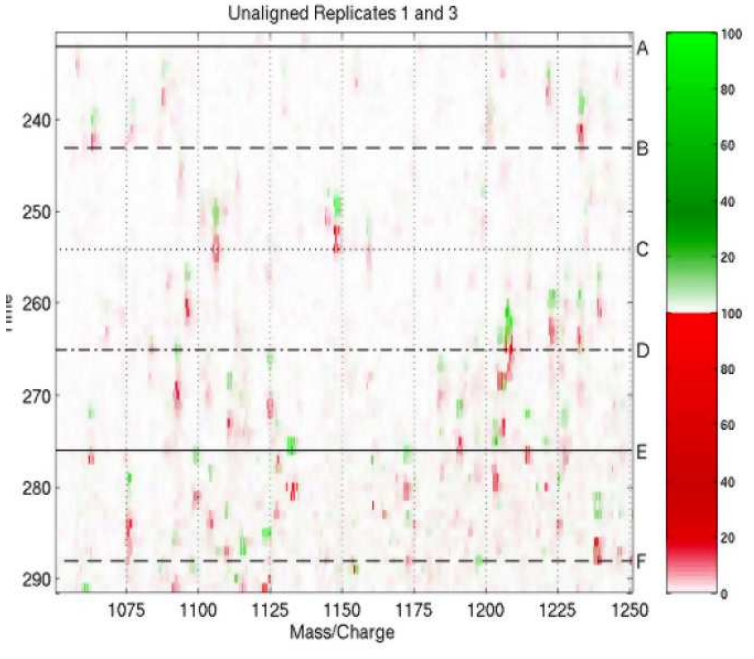
Comparative Mass Spectrometry

We have denoised and calibrated our LC-MS map. We have detected single peaks and summarized these peaks into peptide features. We estimate the abundance the abundance of each peptide by the sums of its isotopic peaks.

But we still cannot compare different LC-MS samples since systematic shifts in retention times are very common. That is, a peptide will not always elute at the same time from the column.

Alignment of LC-MS data

Two LC-MS raw maps overlayed (green and red). We can see that there is a significant difference in retention time (y-axis) between the maps.



After preprocessing and feature detection, we need to correct for this shift before doing any comparisons.

Alignment of LC-MS data

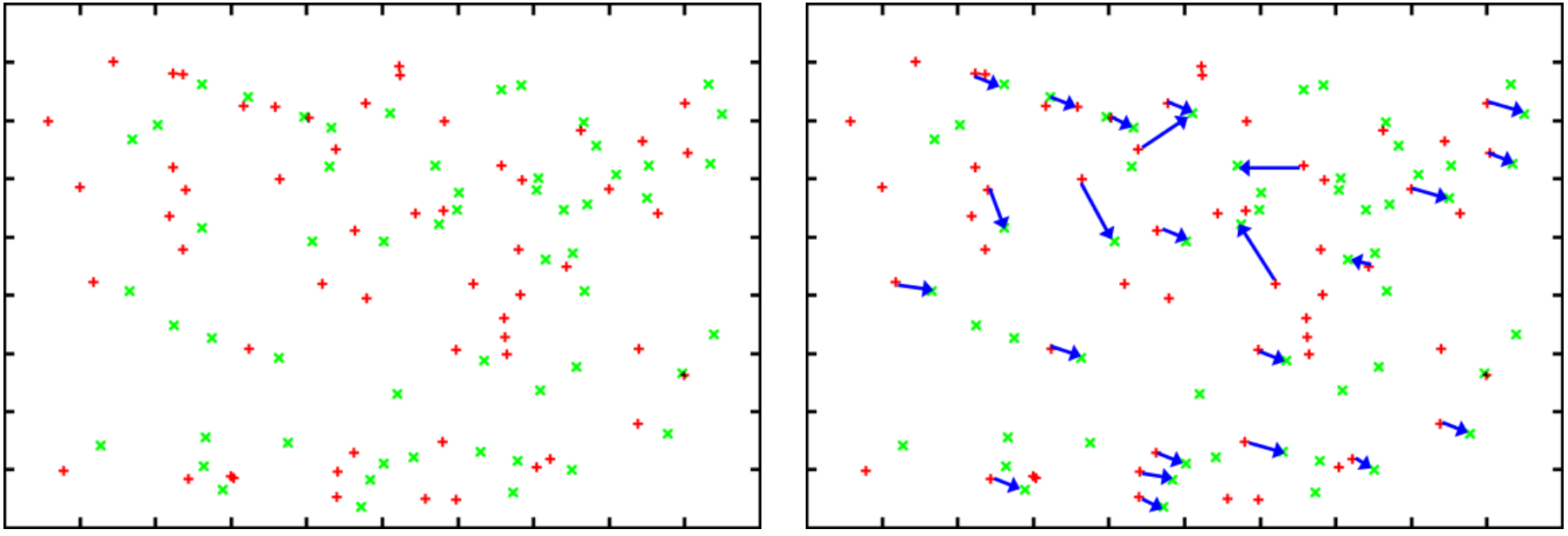
Again, there is a plethora of algorithms available. Examples are

- Dynamic Programming (think of sequence alignment)
- Iterative non-linear regression
- Methods that borrow from Computer Vision (Geometric Hashing)

We will give a brief illustration of the last approach, based on Geometric Hashing (Lange et al. *Bioinformatics* 2007).

LC-MS Alignment using Geometric Hashing

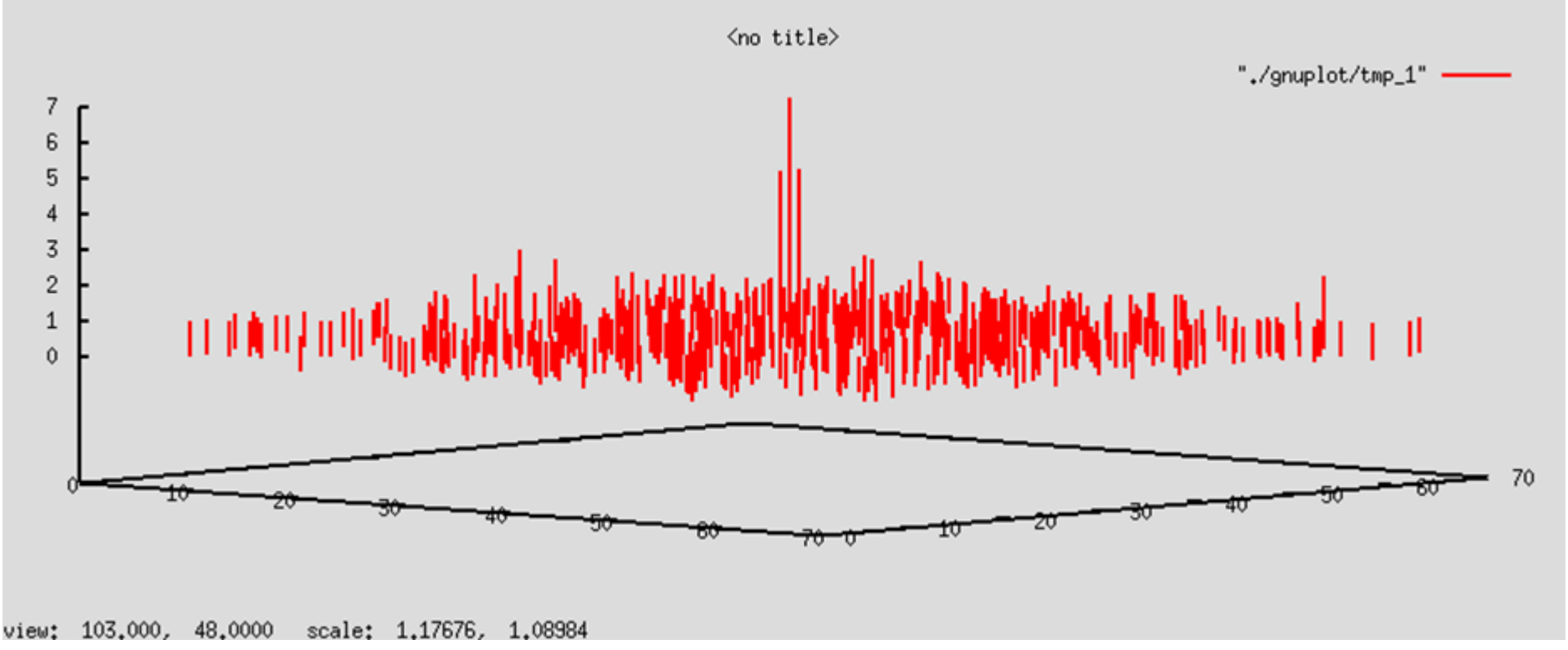
Our aim is to *find corresponding features* in two LC-MS maps. Consider all difference vectors "$\longrightarrow$" from a "+" feature to a "×" feature:



Observation: the differences between non-corresponding features are distributed fairly randomly over a wide range, whereas the differences between corresponding features are all more or less the same.

LC-MS Alignment using Geometric Hashing (2)

Key idea: use a *hashing function* that discretizes (bins) the difference vectors.



In this example, most bins have received only 0 or 1 vectors, whereas the maximum is 7. There is a group of four bins in the middle that sticks out clearly. We estimate a mapping function using a weighted average of these bins and use it to align the LC-MS maps. Details will be covered in an advanced lecture.

Interested in this problem? Sign up for our ”Softwarepraktikum” next semester.

Difference detection

After alignment, we end up with a list of peptide ions (features) found in each LC-MS experiment, maybe with missing values.

Subsequent analysis steps are *decharging* (summarization of ions of the same peptide) and *identification*.

But finally, we want to find proteins that are differentially expressed. Here, standard statistical methods come into play (t-tests, linear models, clustering).... similar to problems as in Microarray data analysis and not covered here.

Summary

- Mass spectrometry is not a new technology but has recently become popular to analyze proteins in complex samples.
- You should remember some key facts about the technology: ESI, MALDI, etc.
- The resulting data is huge (easily 1-2 GB per sample) and noisy.
- Currently very popular playground for Computer Scientists, Mathematicians and Bioinformaticians.
- Typical steps during the data analysis are baseline removal, peak detection, isotopic pattern detection and alignment. You should remember some key methods.
- Next lecture will cover LC-MS/MS data and protein identification.