Algorithms for Peptide Mass Spectrometry

Ole Schulz-Trieglaff

Max Planck Research School for Computational Biology
and
Free University Berlin, Germany

joint work with Rene Hussong, Clemens Gröpl,
Andreas Hildebrandt and Knut Reinert

International
Max Planck Research School
for Computational Biology
and Scientific Computing

Freie Universität Berlin

Outline

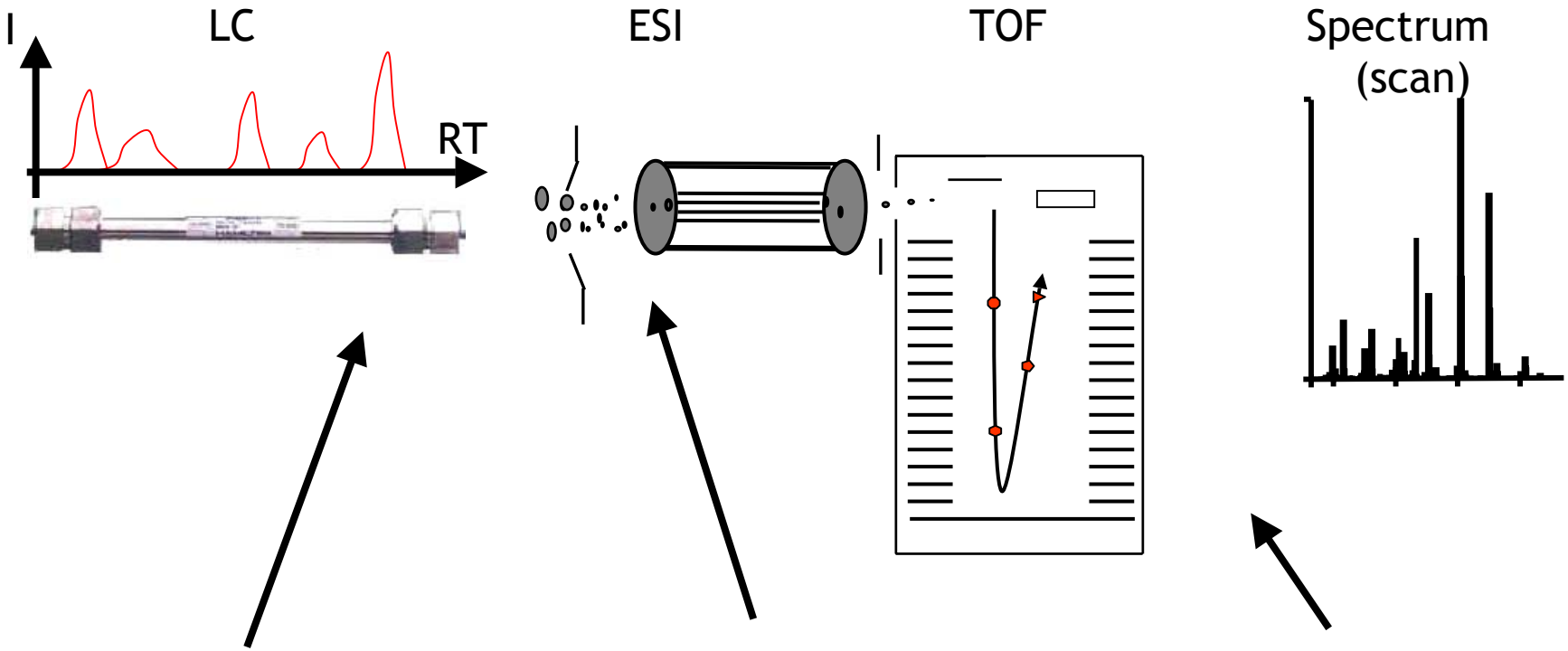
- Computational quantification of peptides

  How to obtain quantitative information about peptides in a biological sample?

- Quality control

  How good is my result ?

Liquid Chromatography-Mass Spectrometry (LC-MS)

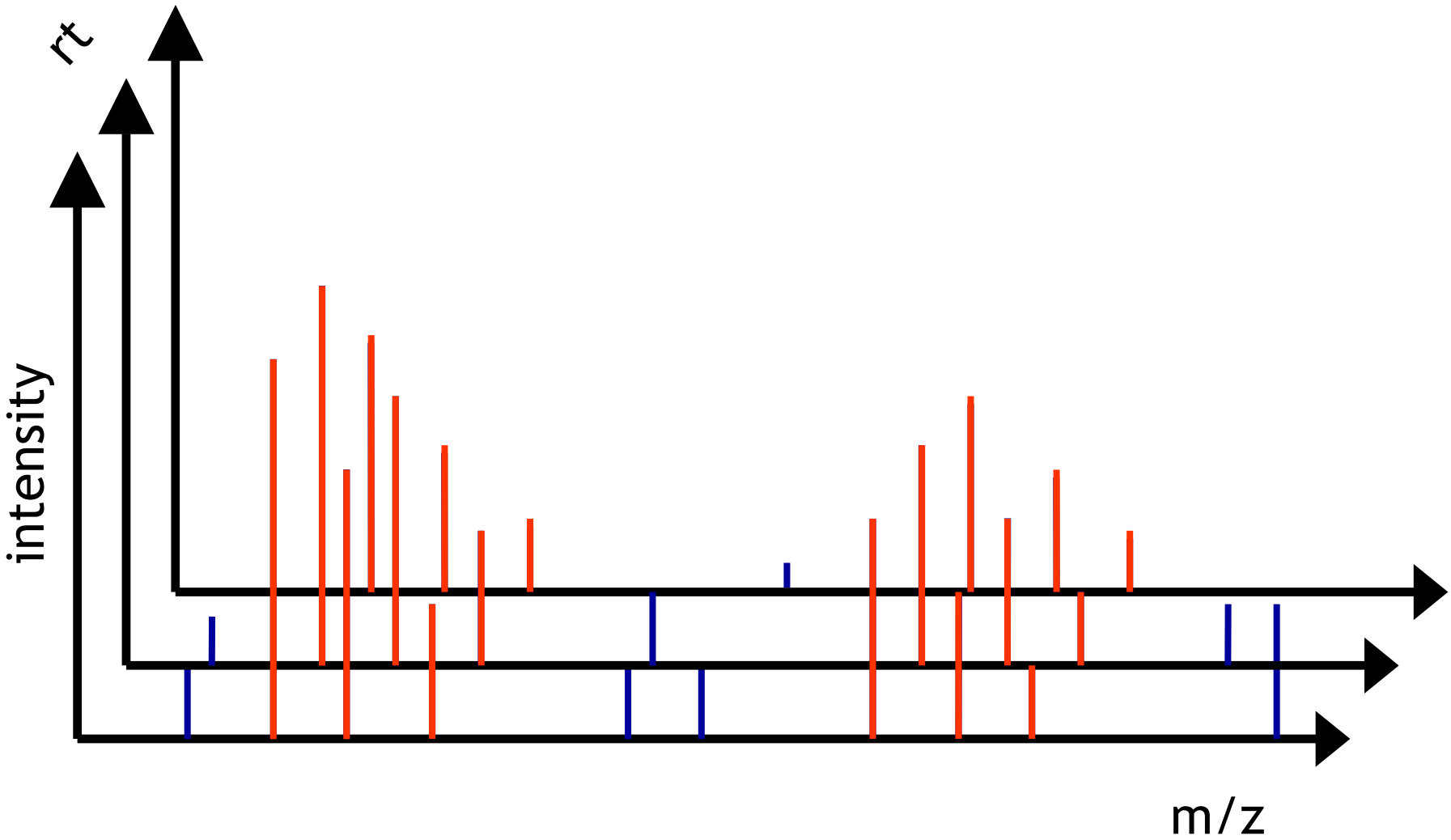


Separation 1
Different peptides have different retention time (rt).

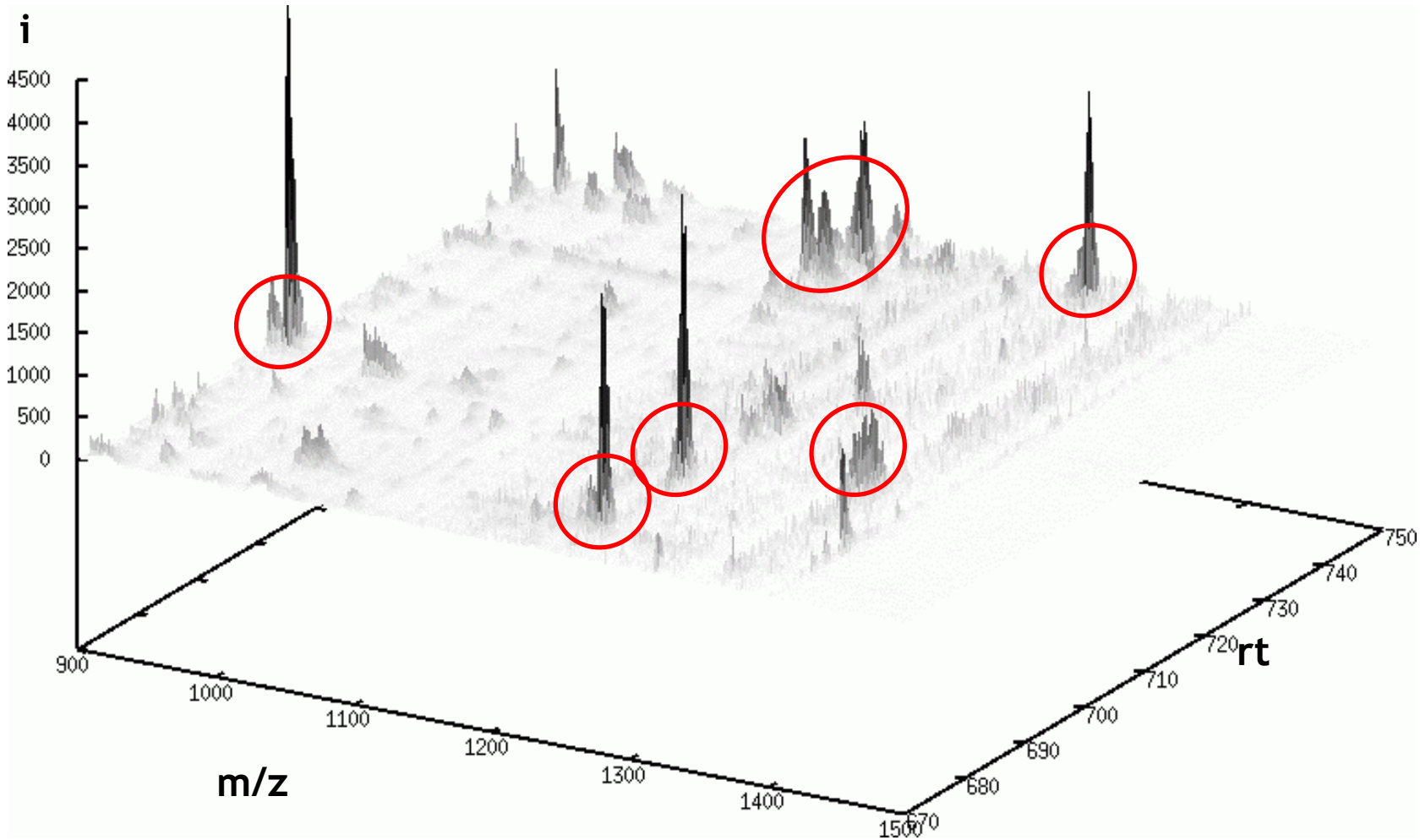
Ionization
Peptide receives z charge units.

Separation 2
Detector measures mass/charge (m/z).

LC-MS map



LC-MS data acquisition



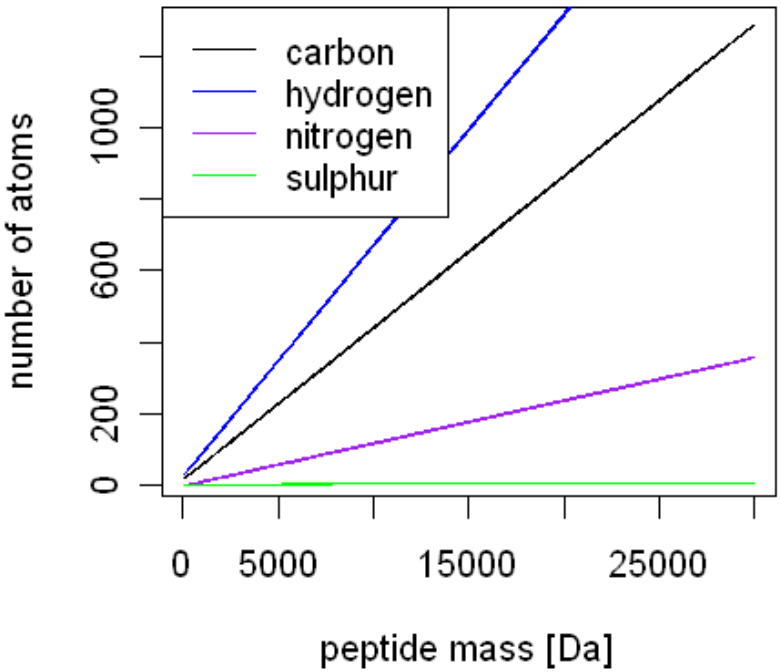
Isotopic pattern

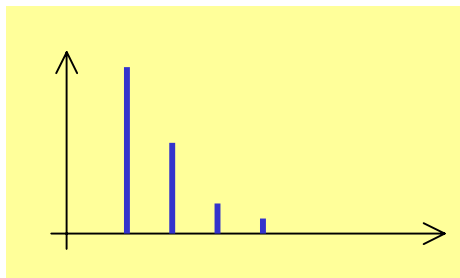
- Natural isotopes occur with well-known abundances.
- Can be modelled by a binomial distribution.
- Depend on molecular formula of peptide.

^{12}C	98.90%	^{13}C	1.10%		
^{14}N	99.63%	^{15}N	0.37%		
^{16}O	99.76%	^{17}O	0.04%	^{18}O	0.20%
^{1}H	99.98%	^{2}H	0.02%		

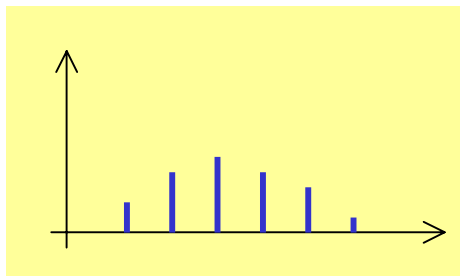
Modeling isotopic pattern

Collect peptides from protein database and compute average amino acid ("*averagines*"). Tabulate average isotope pattern for a range of peptide masses.





small peptide

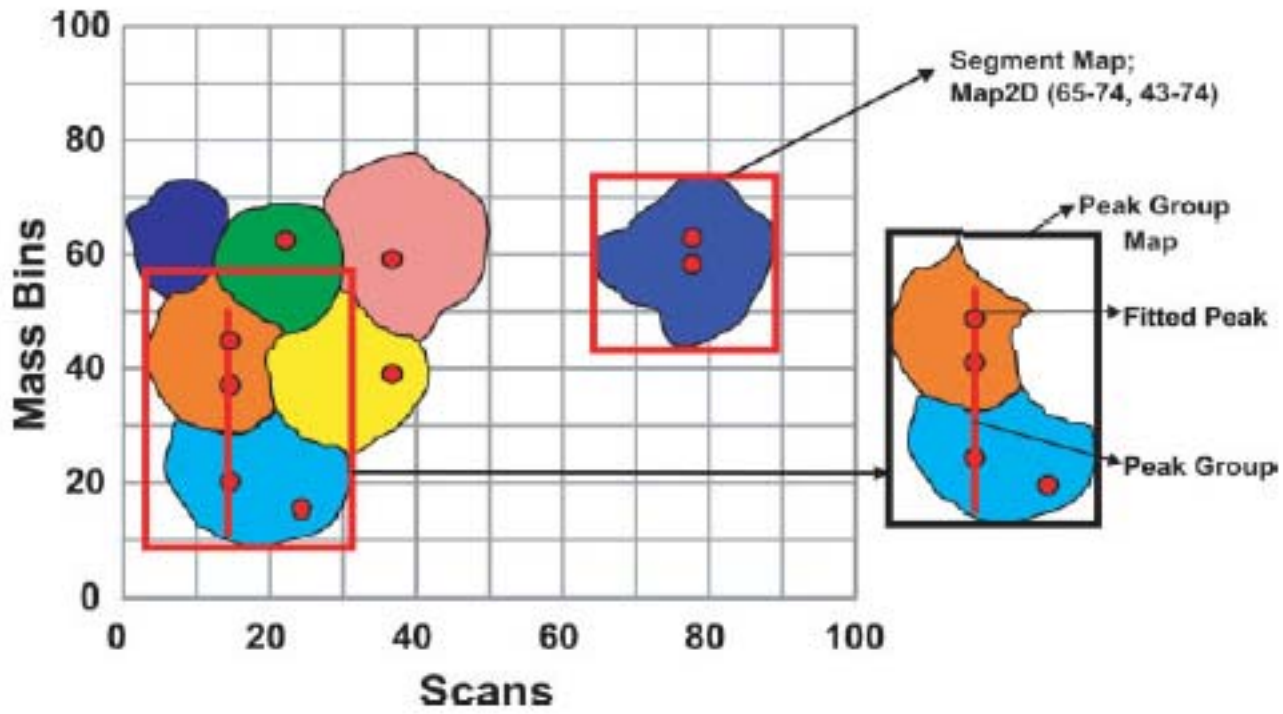


large peptide

Why bother ?

- Clinical studies but also basic research rely on an accurate quantification of peptides or proteins.
- All subsequent steps depend on its quality.
- Modern mass spectrometer generate thousands of spectra per day.
- Need for fast and accurate algorithms !

Previous work



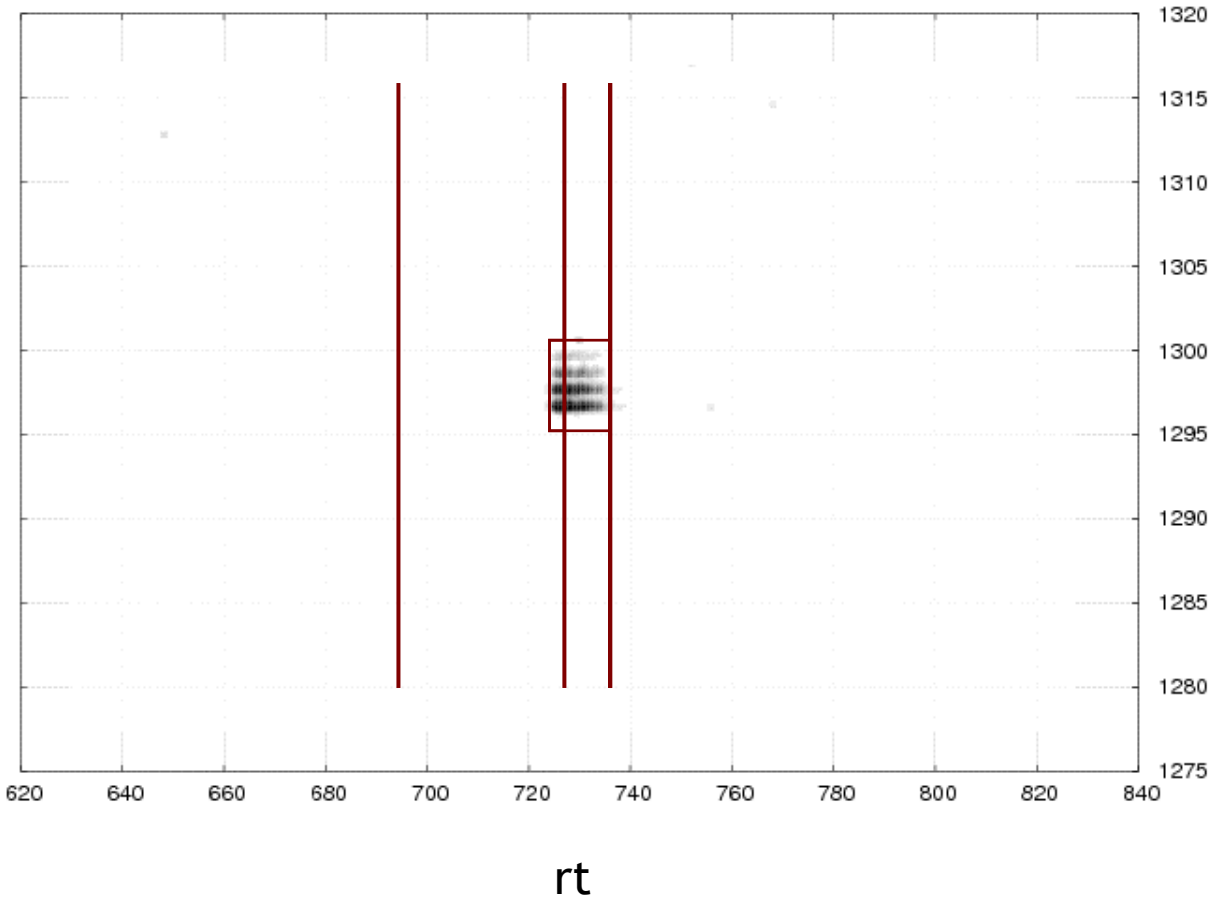
Other approaches based on image processing methods including a (global) segmentation of LC-MS map.

Our approach

Idea: We know what we are looking for and how it looks like. So why not use this knowledge?

1) Pre-process scans by local pseudo-alignment.
2) Sweep across the LC-MS map and scan for isotopic pattern using wavelets.
3) Combine isotopic pattern in subsequent scans.
4) Filter for false-positives by fitting a peptide template.

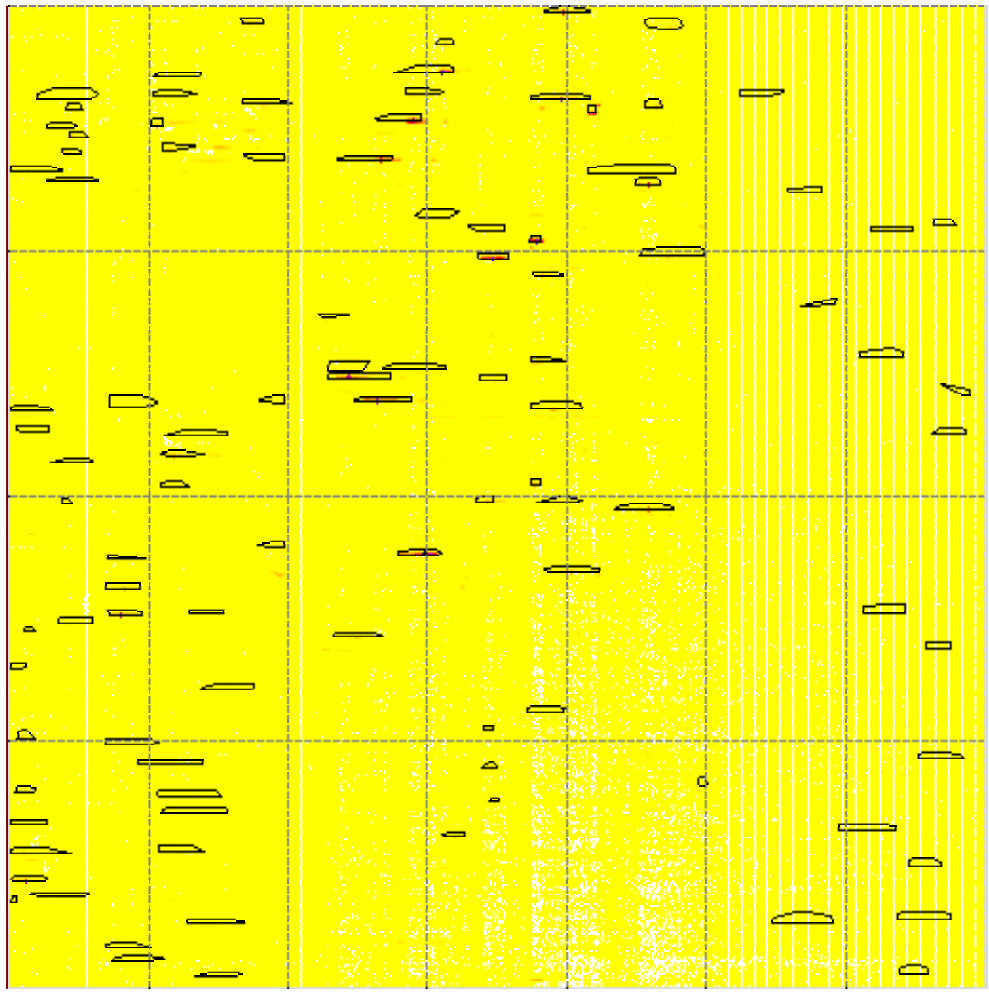
Sweep and combine



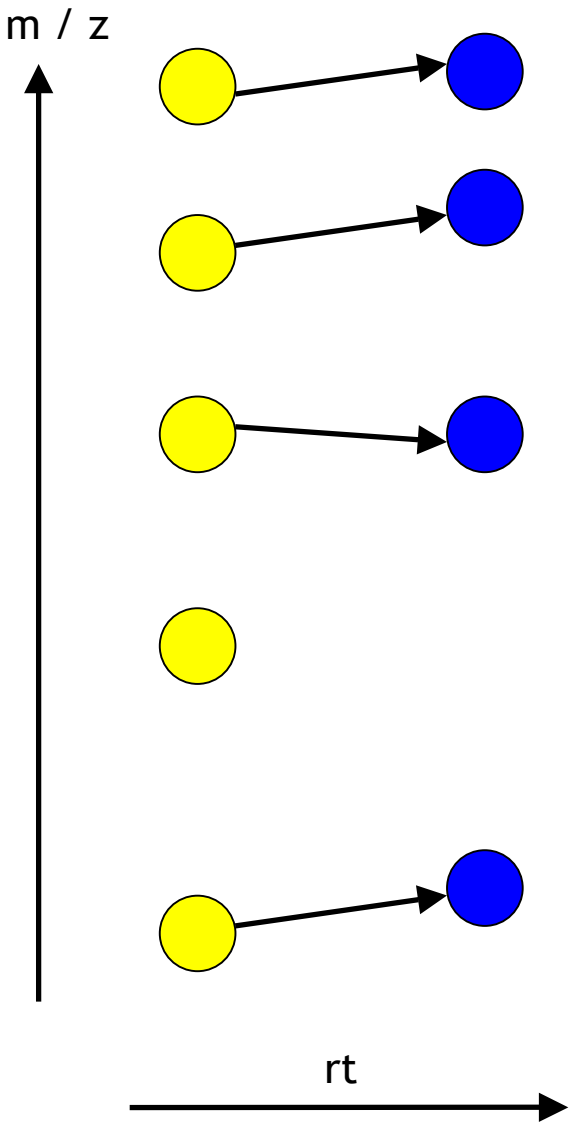
Determine most likely charge state and m/z by voting across all scans.

Hash m/z, charge estimate and extend match.

Sweep and combine



Pseudo-alignment



For each scan, look ahead in time (i.e. at the next scan) .

Add intensities of data points lying at similar positions in the next scan.

Aim: improve s/n ratio by raising isotopic pattern over noise level.

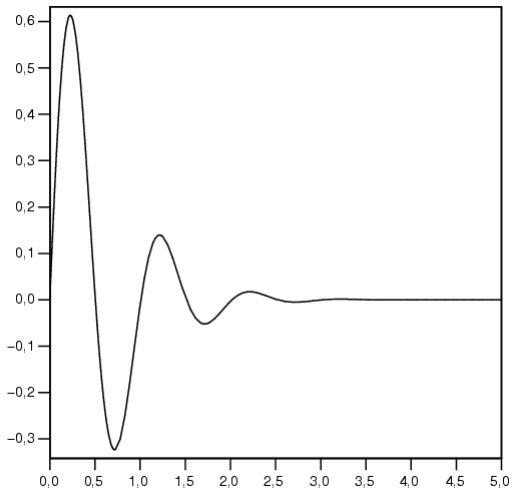
Sweep and combine

For each scan in LC-MS map:
do
    detect isotopic pattern using wavelets
    hash m/z and charge estimate for each pattern

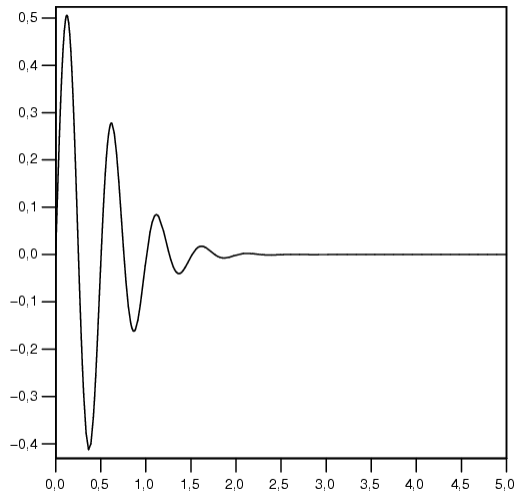
    if (isotopic pattern in previous scan(s) at similar position)
        then continue box
    else
        open new box surrounding the isotopic peaks.
    fi

done

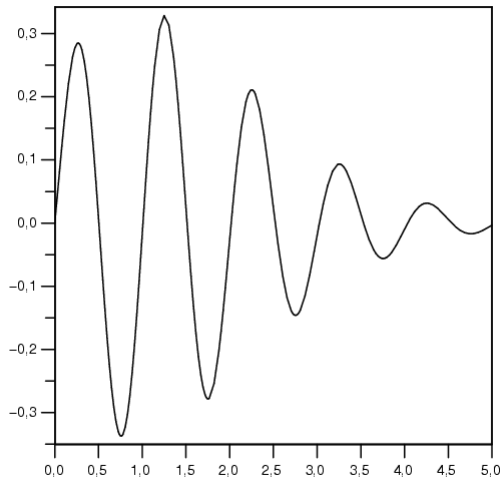
Wavelet-based pattern detection



mass 500, charge 1



mass 500, charge 2



mass 2000, charge 1

$$\mathrm{W}_{\varphi}\, s(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} s(x)\varphi'\left(\frac{x-b}{a}\right)dx,$$

transformed signal

signal

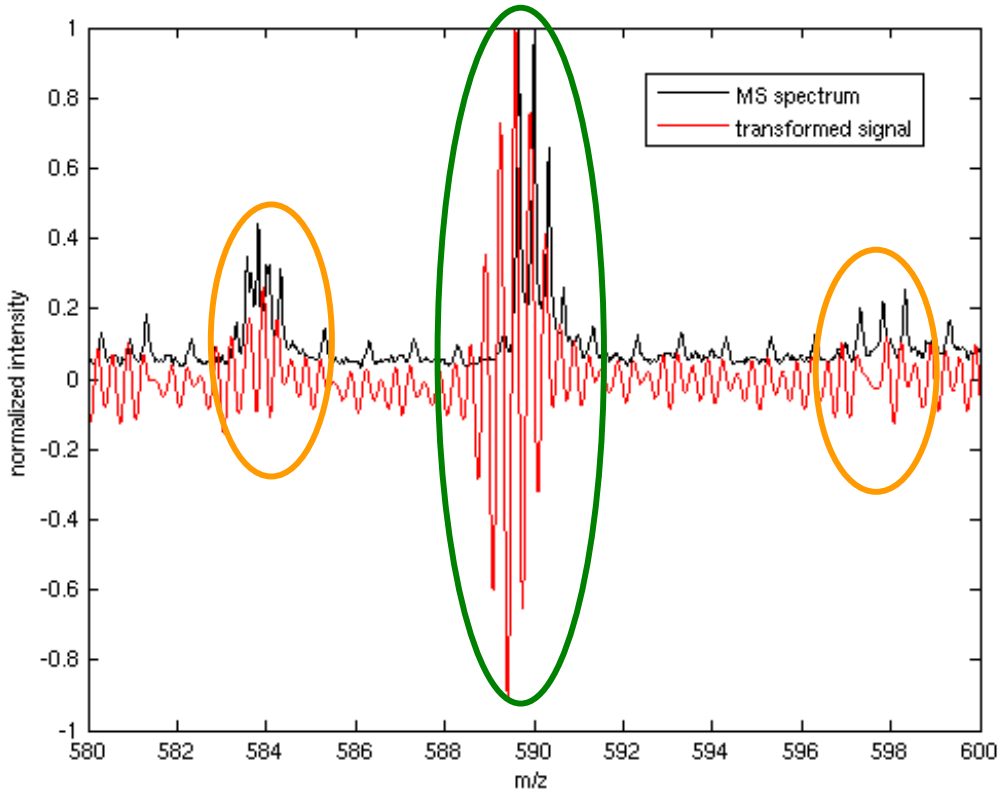
mother wavelet

Wavelet-based pattern detection

non-peptidic compound

peptide with charge 3

noise peaks

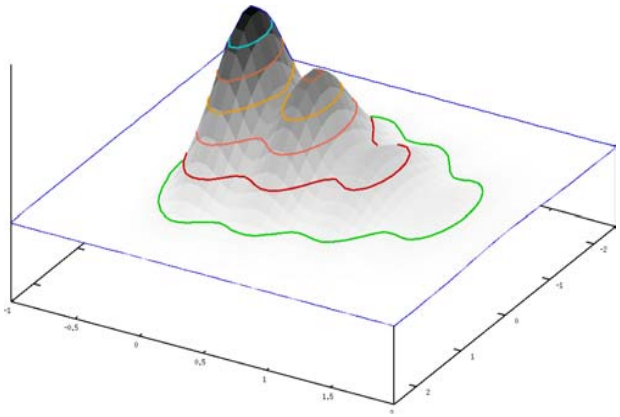


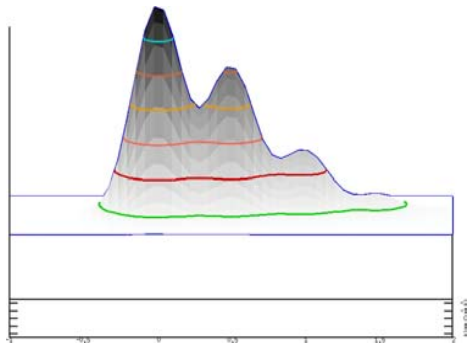
Scoring of intervals in wavelet transform based on mean intensity and (local) variance (F-statistic).

Filtering candidates

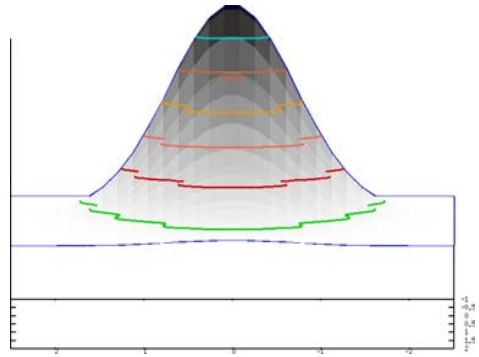
Filter for false positives using a peptide template.

Peptide template = isotope distribution + elution profile





m/z



RT

- Discard points with bad fit to temple.
- Discard regions with bad correlation to template.

Results

A test case: mix of standard peptides.

Stability analysis: can we detect distorted isotopic pattern, too ?

Add uniformly distributed noise with amplitude of 10%, 25%, 50% and 75% of the intensity of the monoisotopic peak.

Check mass, charge and bounding box the peptide sets extracted by our algorithm.

Results

Data set: mix of standard peptides.

Oxytocine, 1007.5 Th, Charge 1					
Amplitude	0%	10%	25%	50%	75%
scans	11/11	11/11	10/11	10/11	0/11
charge	Yes	Yes	Yes	Yes	No

Substance P, 674.5 Th, charge 2					
Amplitude	0%	10%	25%	50%	75%
scans	16/20	16/20	13/20	12/20	13/20
charge	Yes	Yes	Yes	Yes	No

What's next ?

- How "good" is my result (e.g. the set of peptides) ?

Sounds trivial, but it isn't !

Example: Using an addition experimental step (MS/MS ion fragmentation) we can get sequence information for several hundreds of peptide ions in a LC-MS map.

Feature extraction algorithms extract thousands of peptide signals from a typical map.

How good is my set ?

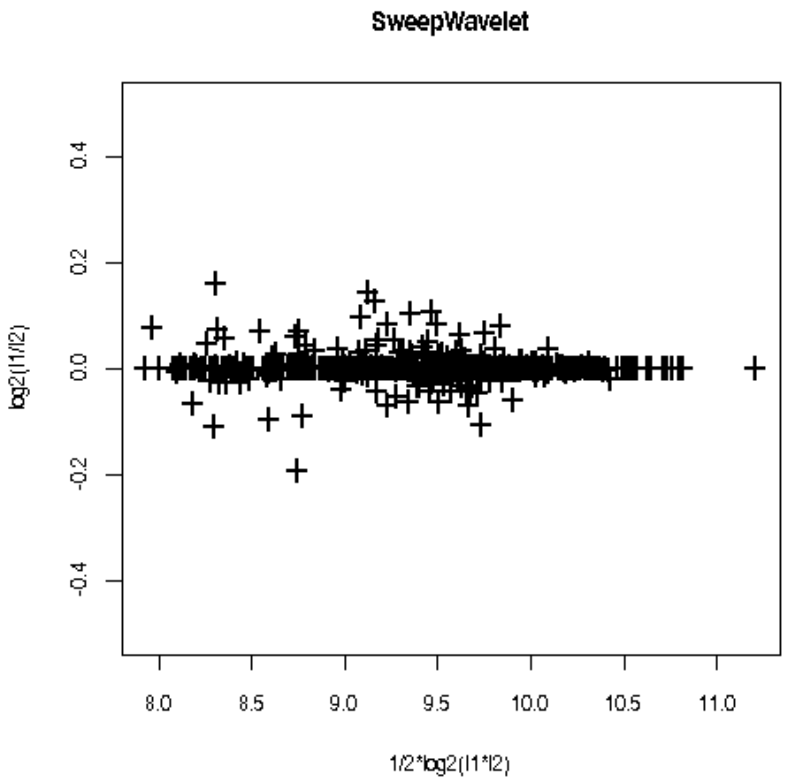
Many signals without sequence information. Too many to inspect them manually. Do they make sense?

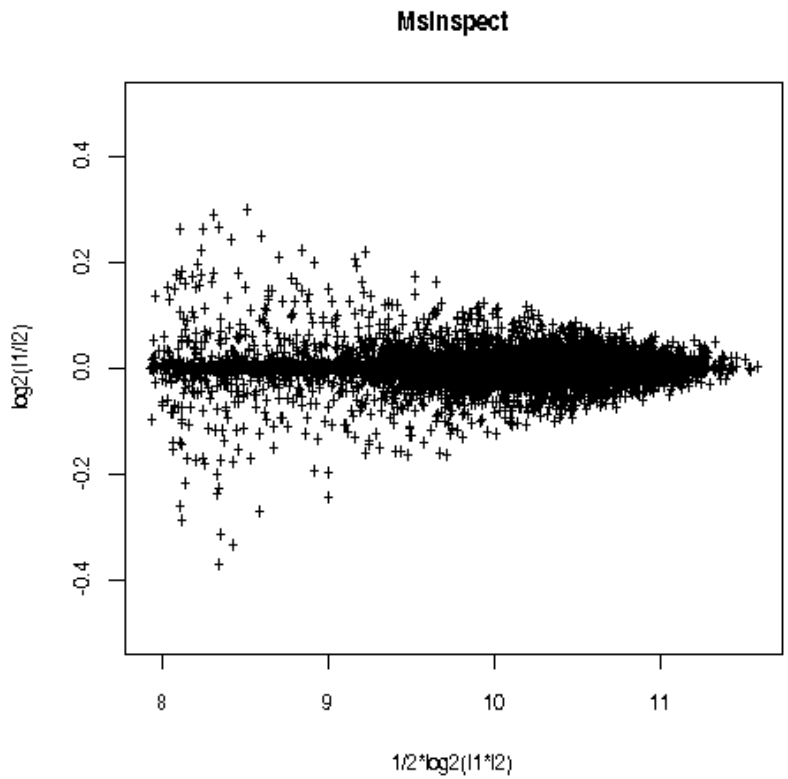
Two criteria: meaningful signals should

- have equal intensities between replicate samples (within some experimental error).

- their masses should be close to the masses of the peptides in this particular organism.

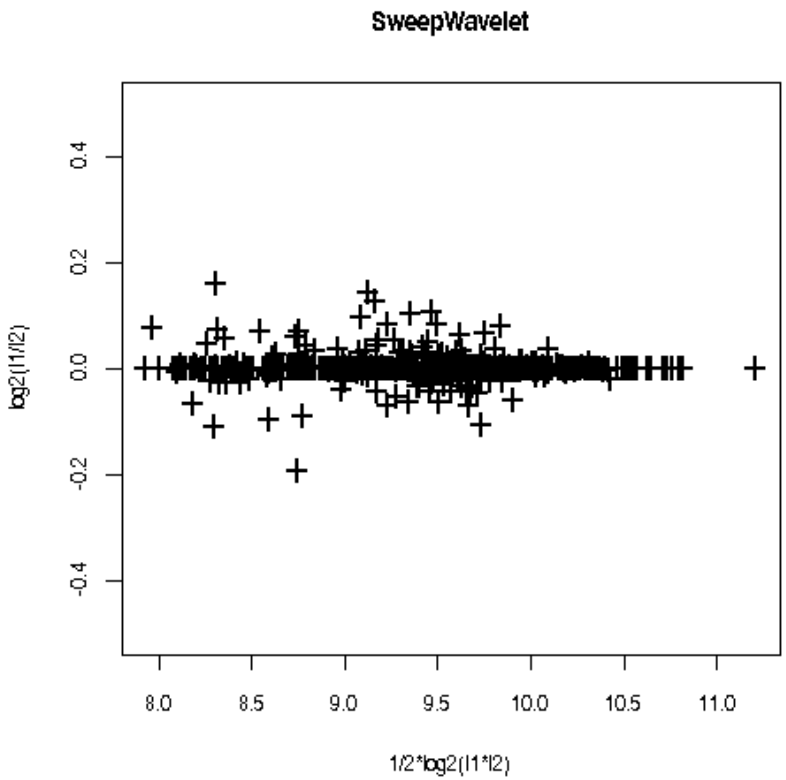
MA plot of replicate samples

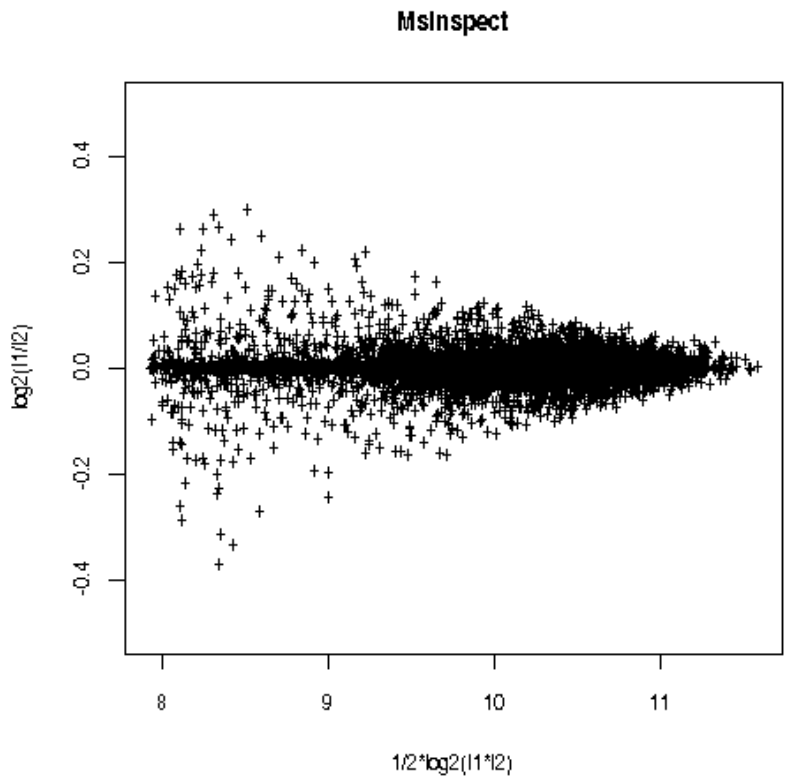




MA plot = average intensity of matching signals (x) vs. ratio of signal intensity (y), both on log-scale

MA plot of replicate samples

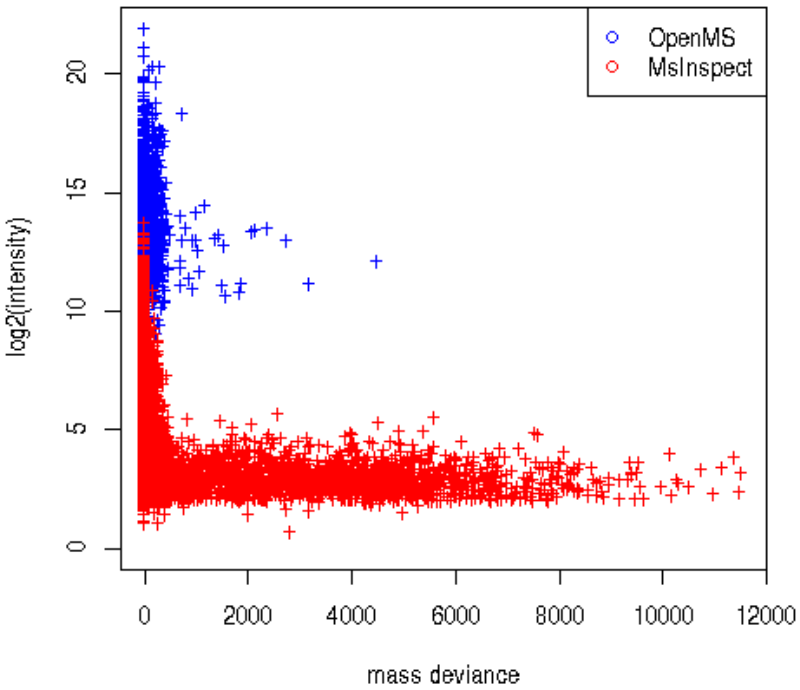




Quantification results of algorithm MsInspect show higher variation but MsInspect also extracts far more signals (10000 vs. 2000 for our approach).

Mass deviance

Do these additional signals make sense ?



Mass deviance = min. distance of a peptide signal mass to the mass of a theoretically obtained peptide feature.

MsInspect detects many features with high mass deviance. Either peptides not in sequence database (unlikely) or noise "picked up" by the algorithm.

How good is my set?

Conclusions:

- We can say a bit about the signals we detected (but not much).
- Different algorithms can give very different results.
- Lack of standard data sets impedes advancement of computational research.

Summary and future work

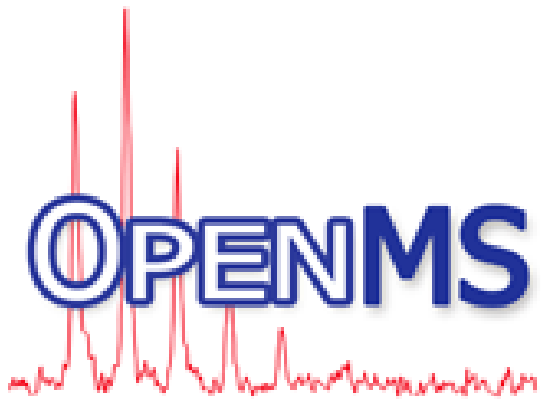
- Algorithm for an automated and accurate quantification of peptides from LC-MS data based on wavelet-based filtering.

- Available under the LGPL at www.openms.de.

Future work:

- Better quality control and more complex data.

- So far only quantification of peptides. How to infer the abundance of the corresponding protein ?



Thanks for your attention.

Any questions ?

trieglaf@inf.fu-berlin.de
www.openms.de