Informatik
Anwendungen in der Genom- und Proteomforschung

Knut Reinert

FU Berlin

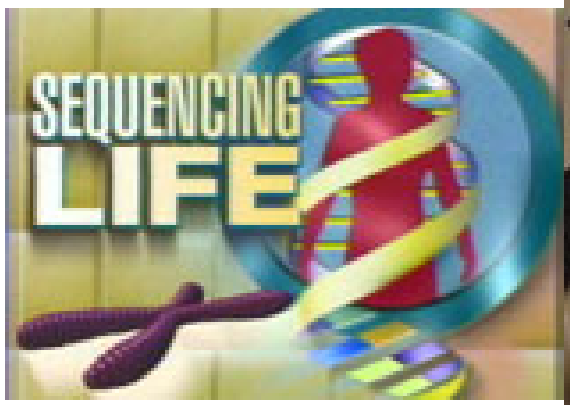
Mai 2006

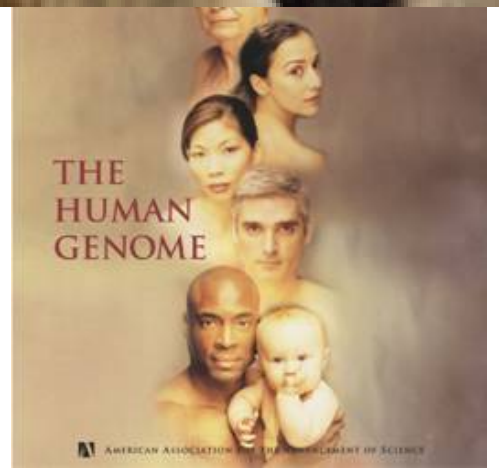
HU Berlin, Adlershof

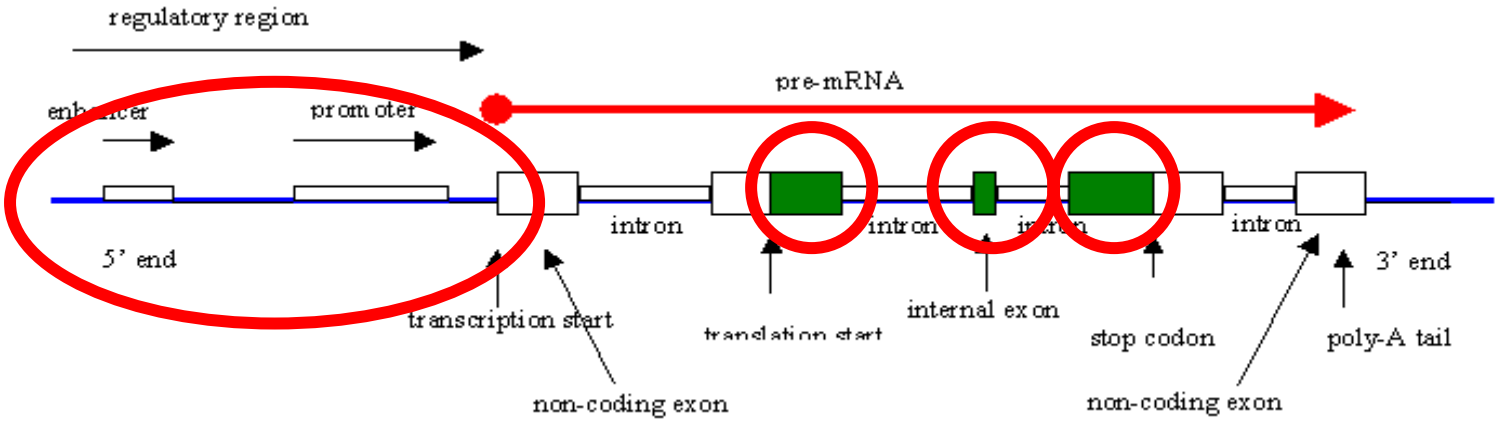
funded by the German Federal Ministry for Education and Research

Bioinformatics in the media

Freie Universität Berlin







Some pictures with the courtesy of the MPI für Informatik, Saarbrücken

Gene finding:

- Improve prediction of coding and regulatory regions
- Comparing multiple genomes is promising

Topics in Bioinformatics

Freie Universität Berlin

Molcular therapie of deseases:
protein-protein docking

Methotrexate bindet an dihydrofolate reductase

Find drugs that alter or inhibit the function of the target molecule.

Searching data bases helps to find suitable candidates and reduce side effects.

Topics in Bioinformatics

Freie Universität Berlin

Identifying SNPs (or other polymorphisms)

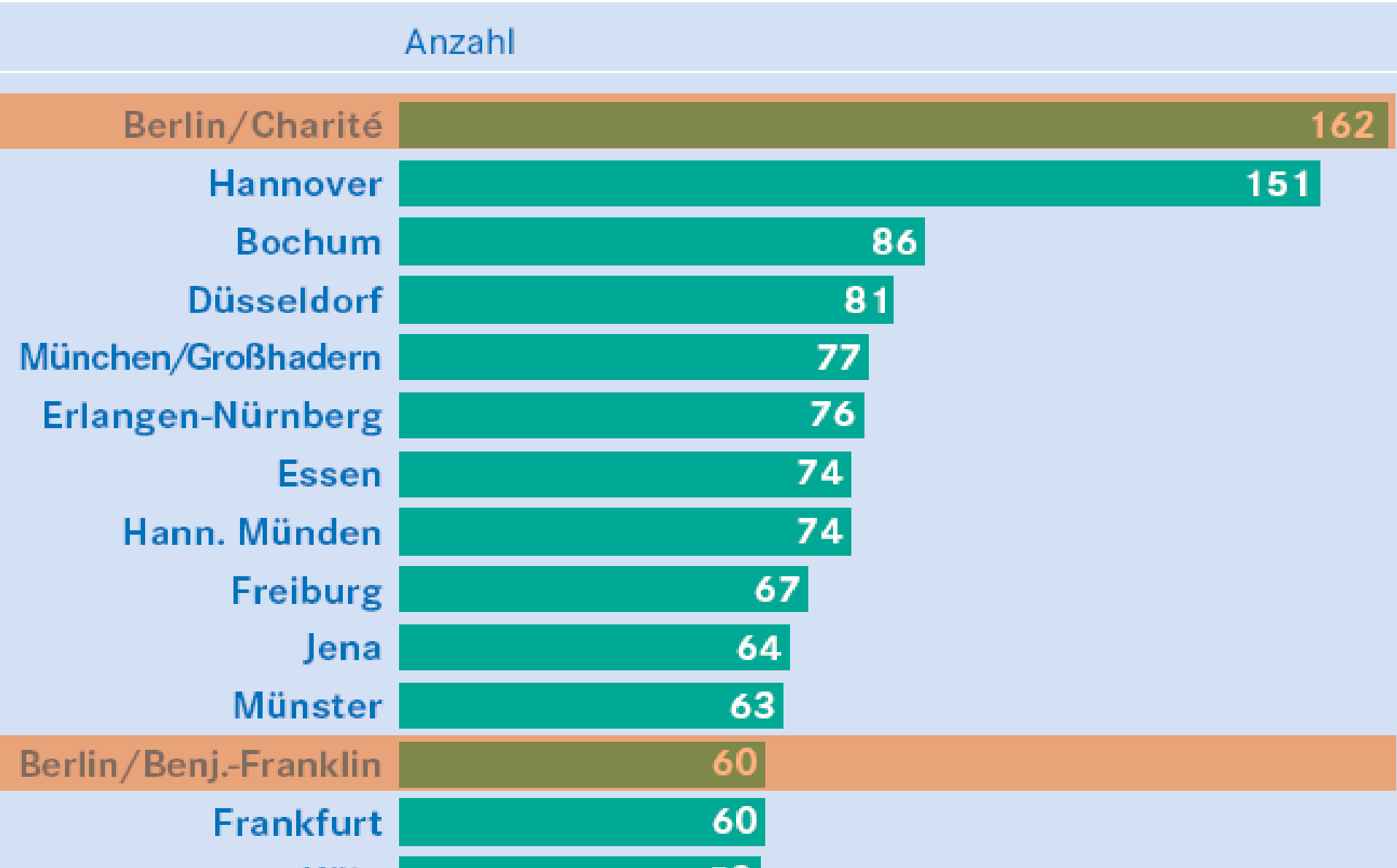
Identifying SNPs allows:

- the association of patterns of genetic diversity to diseases
- the association of genetic patterns to drug tolerance

Deutschland 2004

Nierentransplantationen (ohne Lebendspende)

aufgeschlüsselt nach Transplantationszentren



Topics in Bioinformatics

Freie Universität Berlin

Problem (Project together with Charité):

Within the first year 8-10% of the patients lose the transplant.
After 10 years about 50% have lost the transplant.
Diagnosis is invasiv and leads sometimes to loss of craft.

Goal:

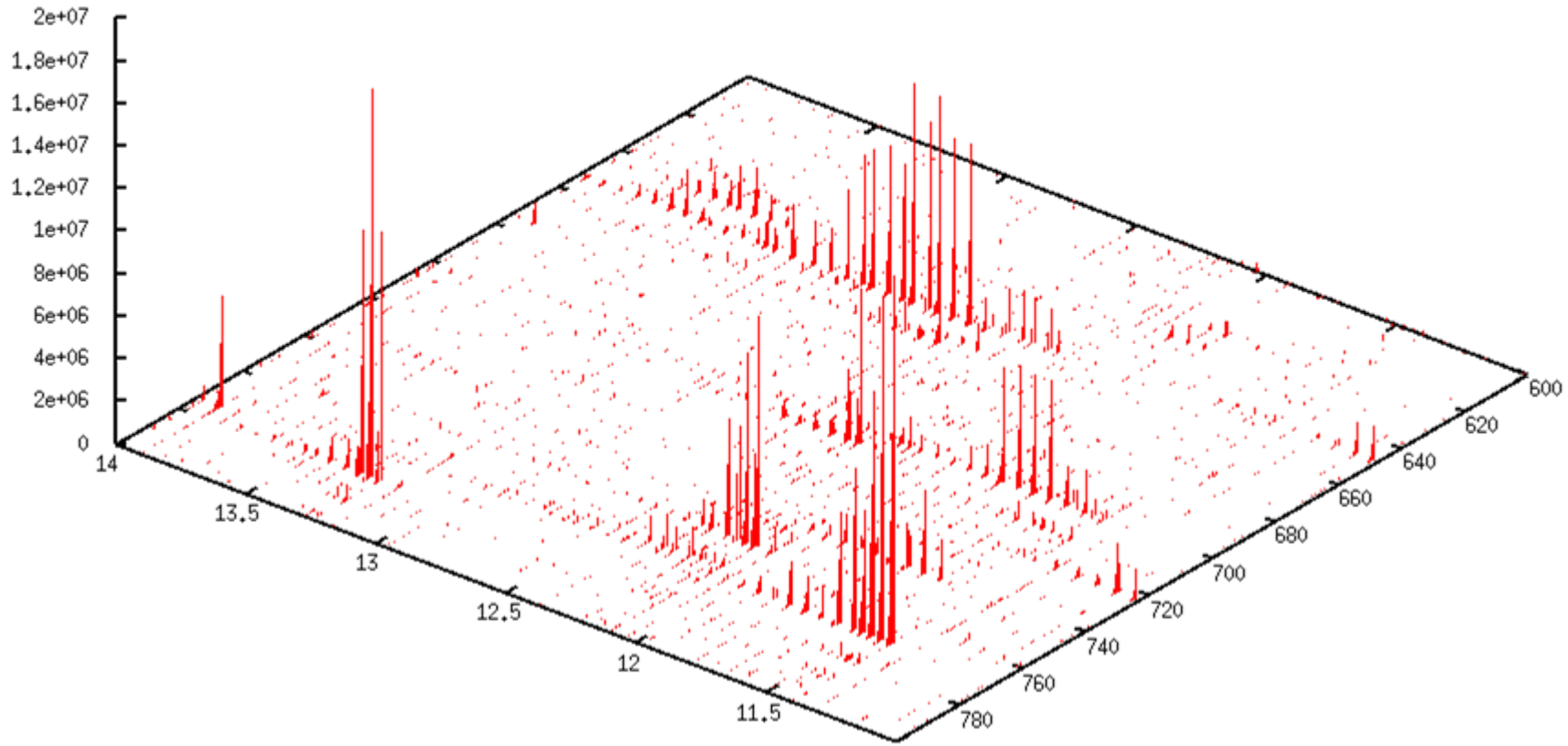
Analyse urin samples of patients and detect as early as possible diagnostic markers to counteract craft loss.

Topics in Bioinformatics

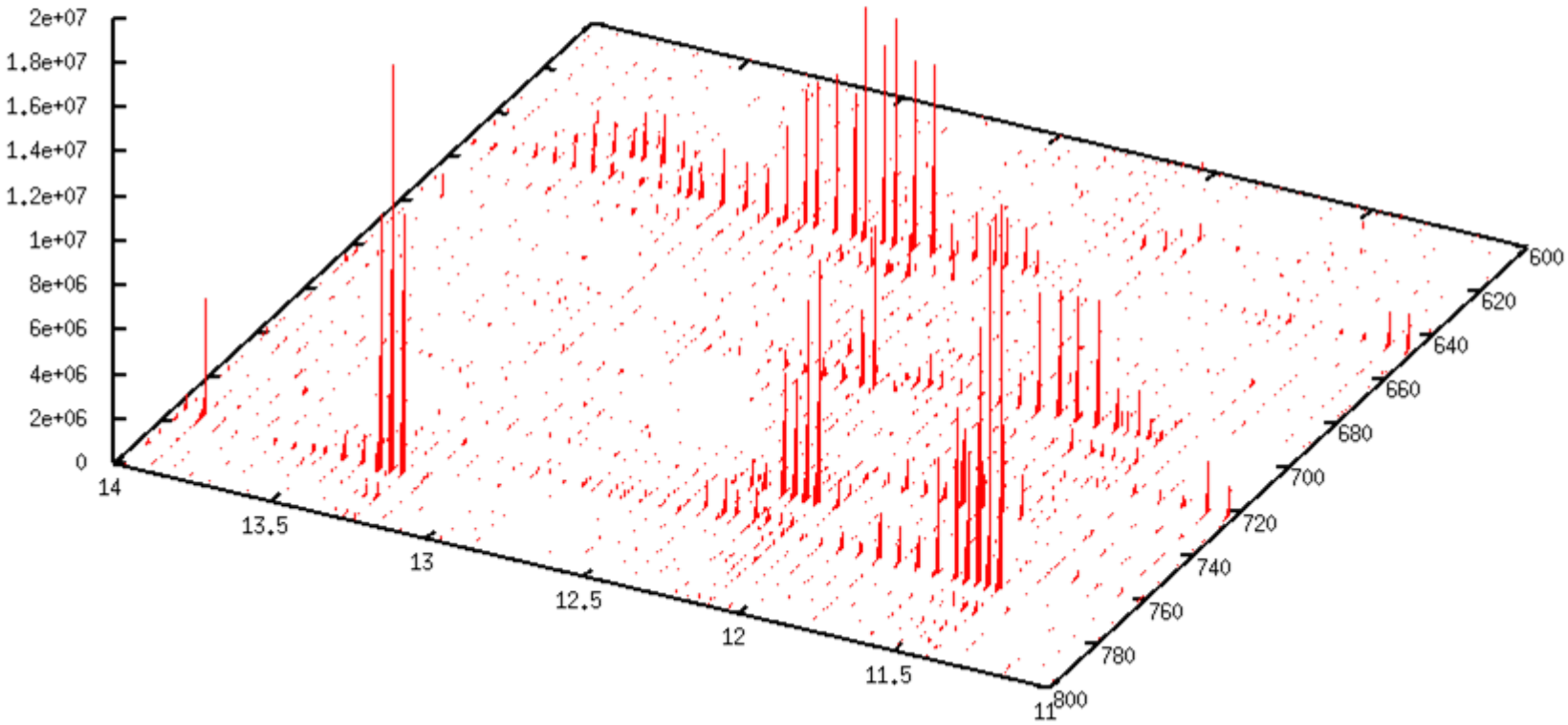
Freie Universität Berlin

Automated measurement methods lead to terabytes of data (Prof. Schlüter)

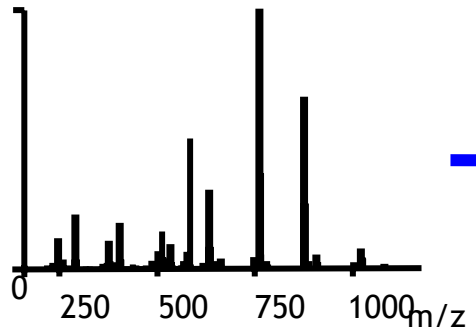
Mass spectrometry



Mass spectrometry

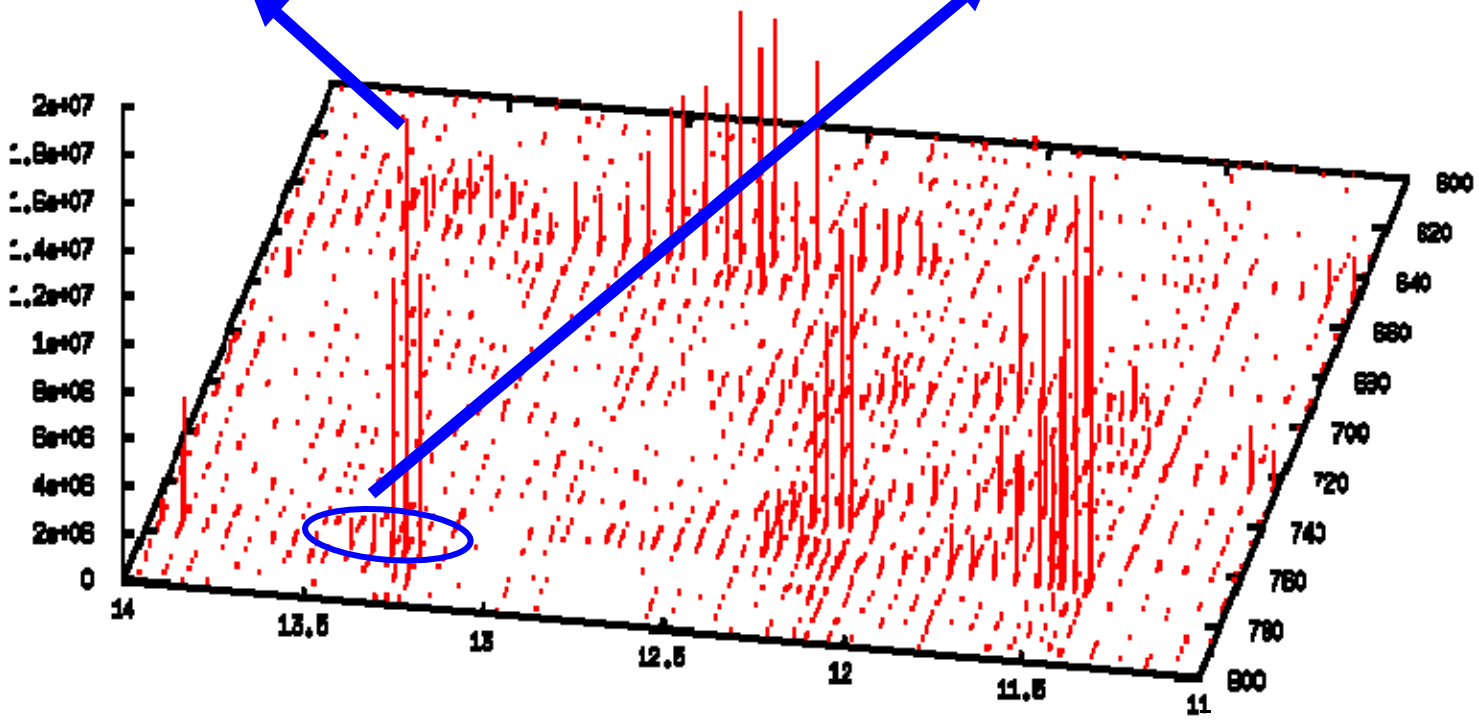


Mass spectrometry



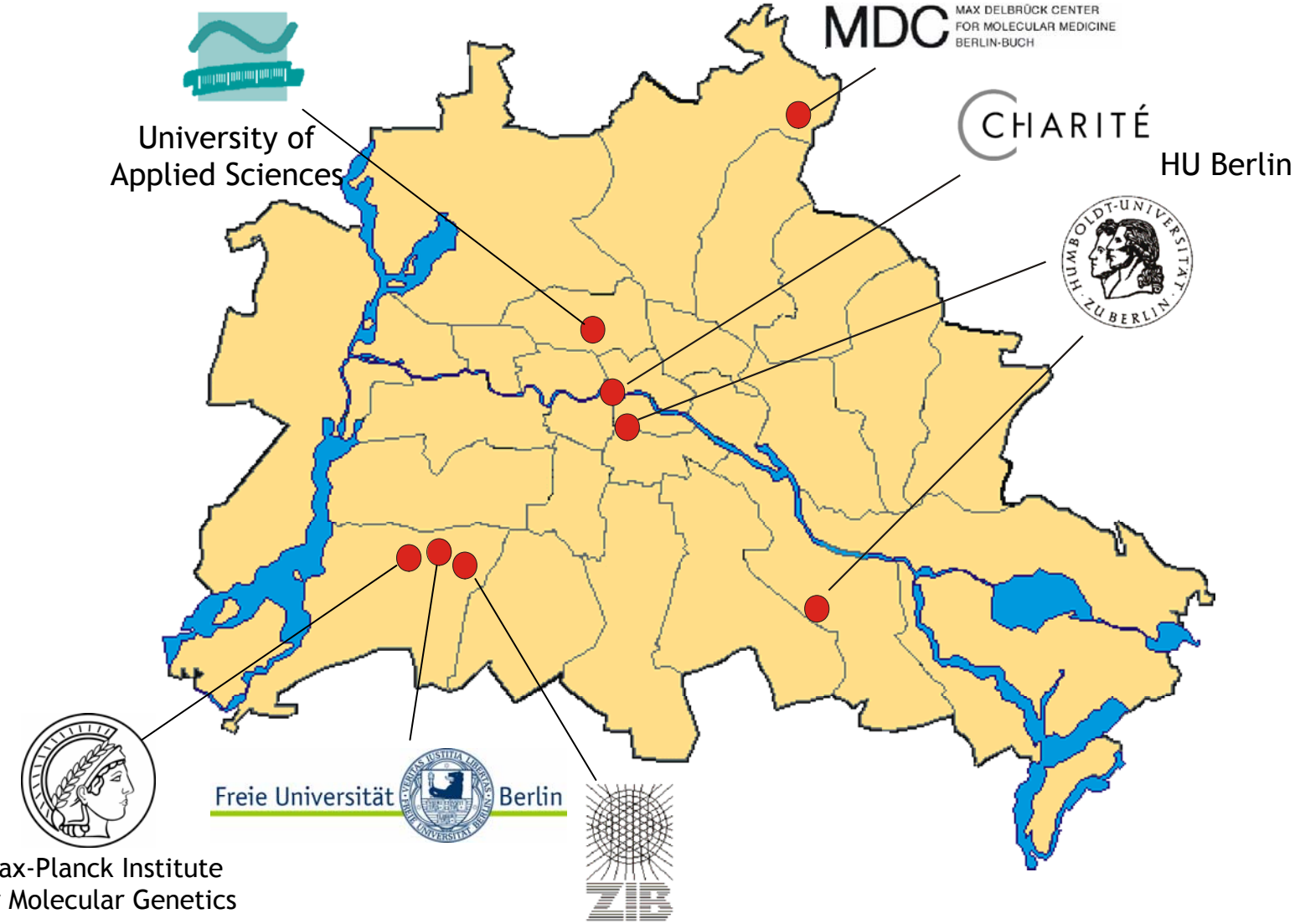
EVAAFAQFGSDLDASTK

3x increased expression



Bioinformatics in Berlin

Freie Universität Berlin

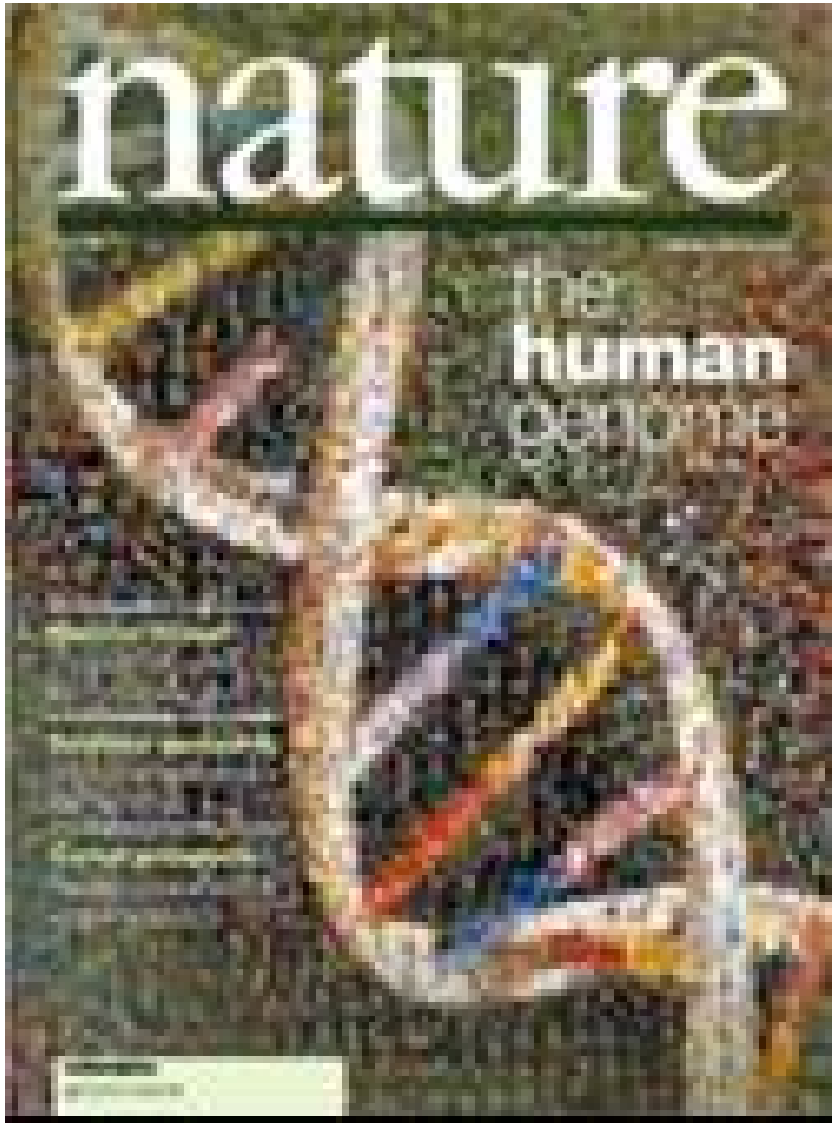


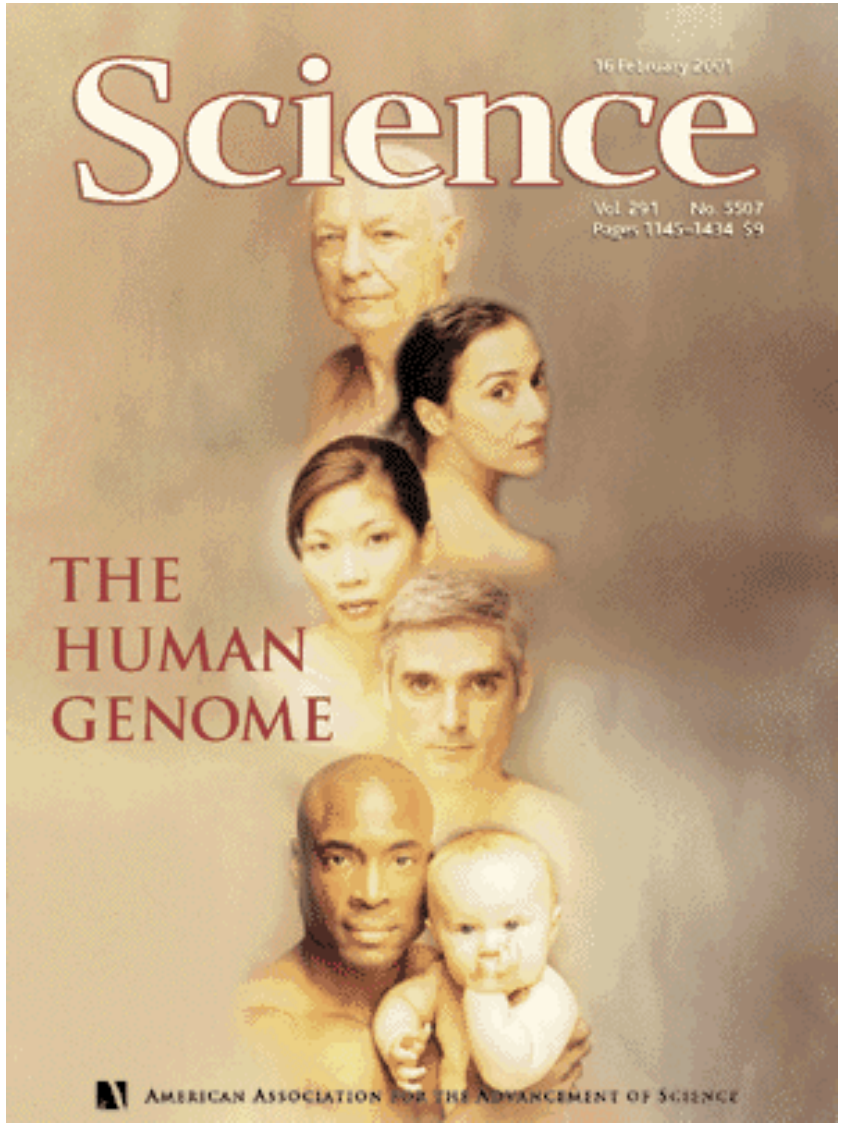
Overview

Freie Universität Berlin

- Motivation for OpenMS
- Bioinformatics Issues in quantitative Proteomics
  - Signal Processing
  - Feature Finding
  - Map Mapping
  - Differential Quantitation
  - Identification, Clustering
  - Software Engineering, Databases

Genomics


nature
the human genome


Science
THE HUMAN GENOME
AMERICAN ASSOCIATION FOR THE ADVANCEMENT OF SCIENCE

Genomics vs. Proteomics

Same **genome**…

…different **proteomes**

Genomics vs. Proteomics

Freie Universität Berlin

Genomics	Proteomics
Genome is rather **static**	Proteome is **dynamic** (age, tissue, what you had for lunch)
~ **30 k** genes	Up to **2000 k** Proteins
Established, fully automated technology (capillary sequencer)	**Emerging** technology (MS, HPLC/MS, protein chips)

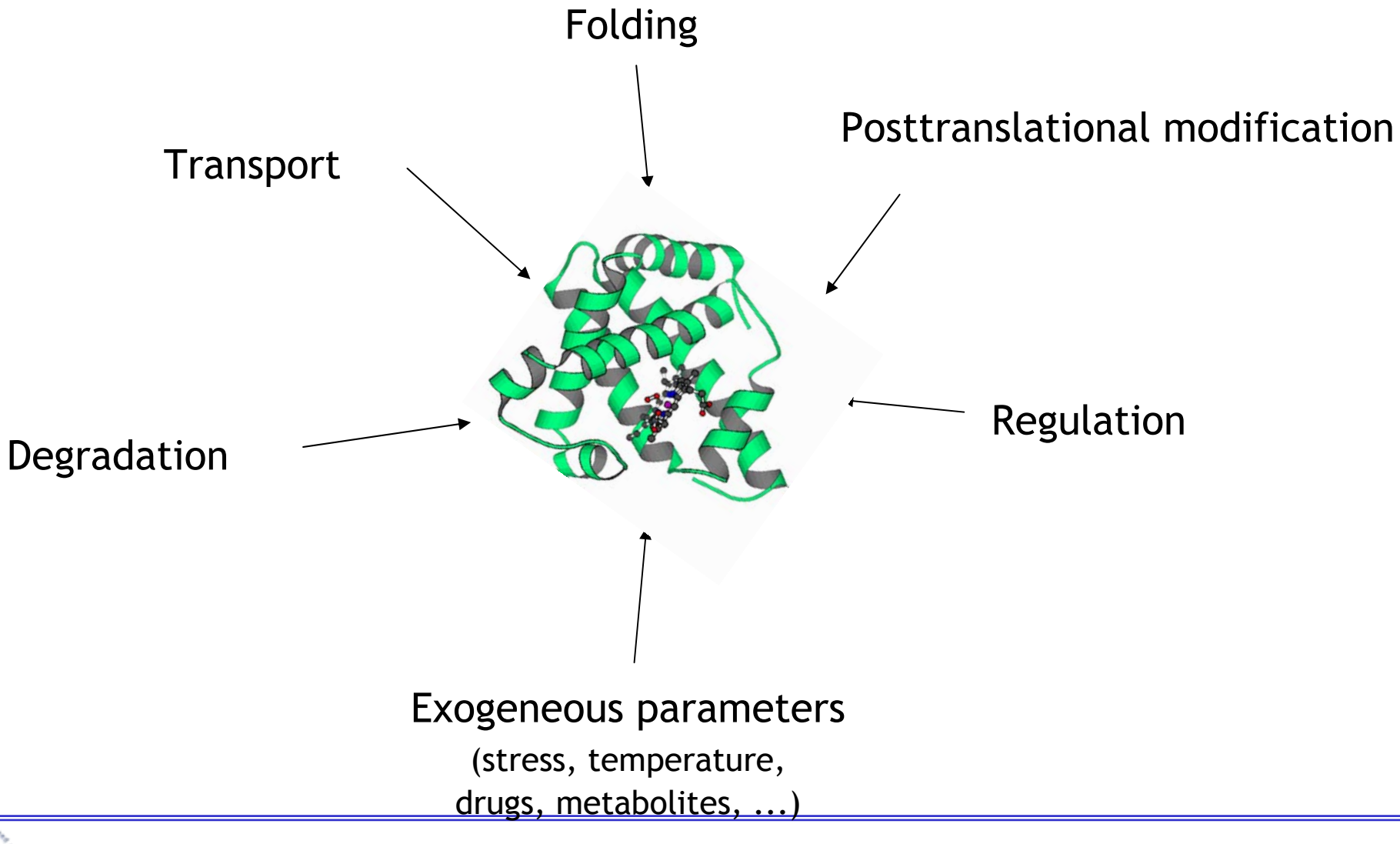
From Genes to Proteins

DNA —Transcription→ mRNA —Translation→ Protein

- Transcription and translation are heavily regulated
- Protein expression levels are not static
- mRNA levels and protein levels often not correlated
- Contradictory results from seemingly similar methods
  - RNA chips
  - DNA chips
  - gene disruption
  - knock out

Anderson et al., Electrophoresis (1998), **19**, 1853-61

Proteins = End Products?



Definition of proteomics

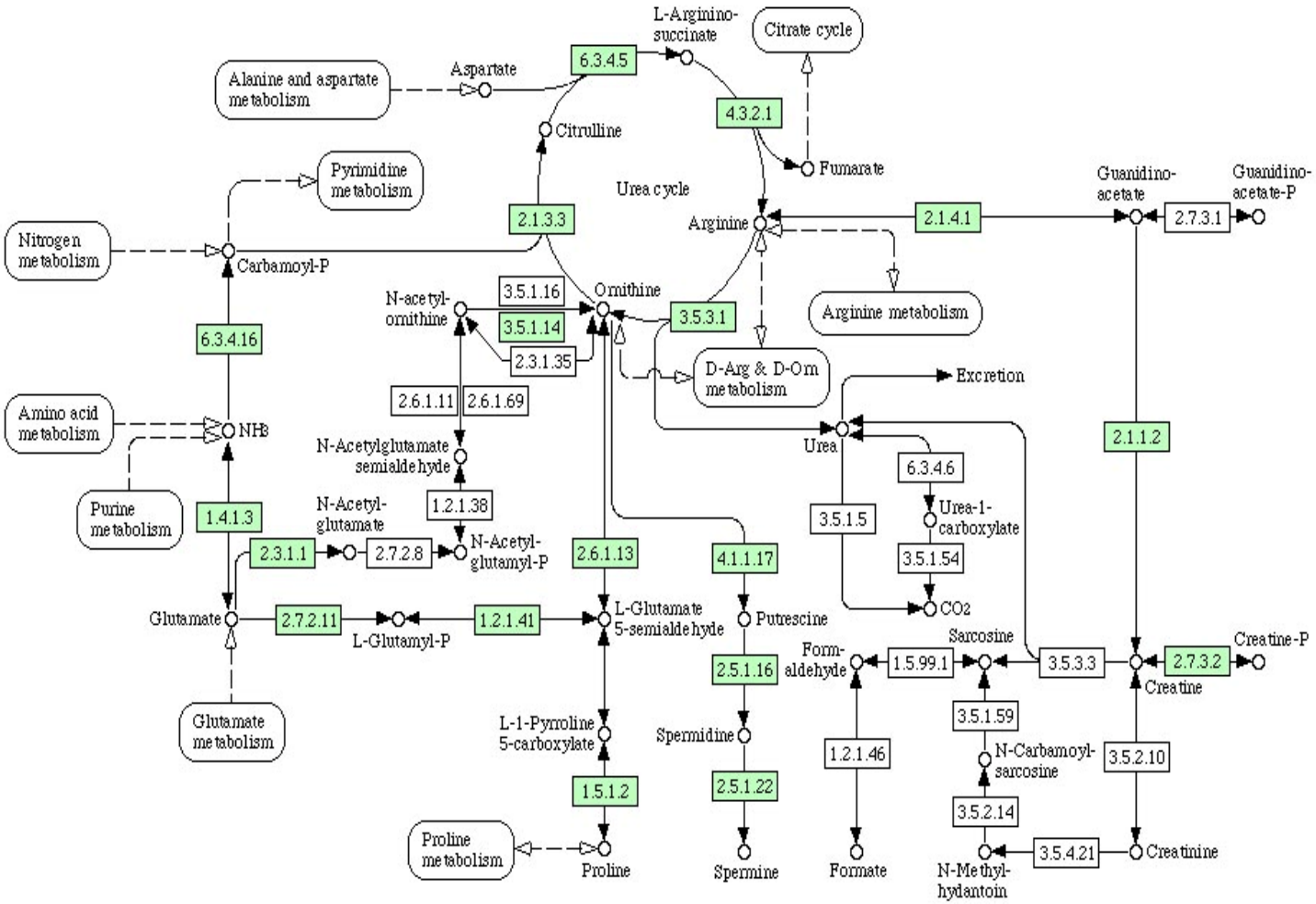
- **Proteomics** [n]: the branch of genetics that studies the full set of proteins encoded by a genome

(www.hyperdictionary.com)

- **Proteomics** can be defined as *the qualitative and quantitative comparison of proteomes under different conditions to further unravel biological processes.*

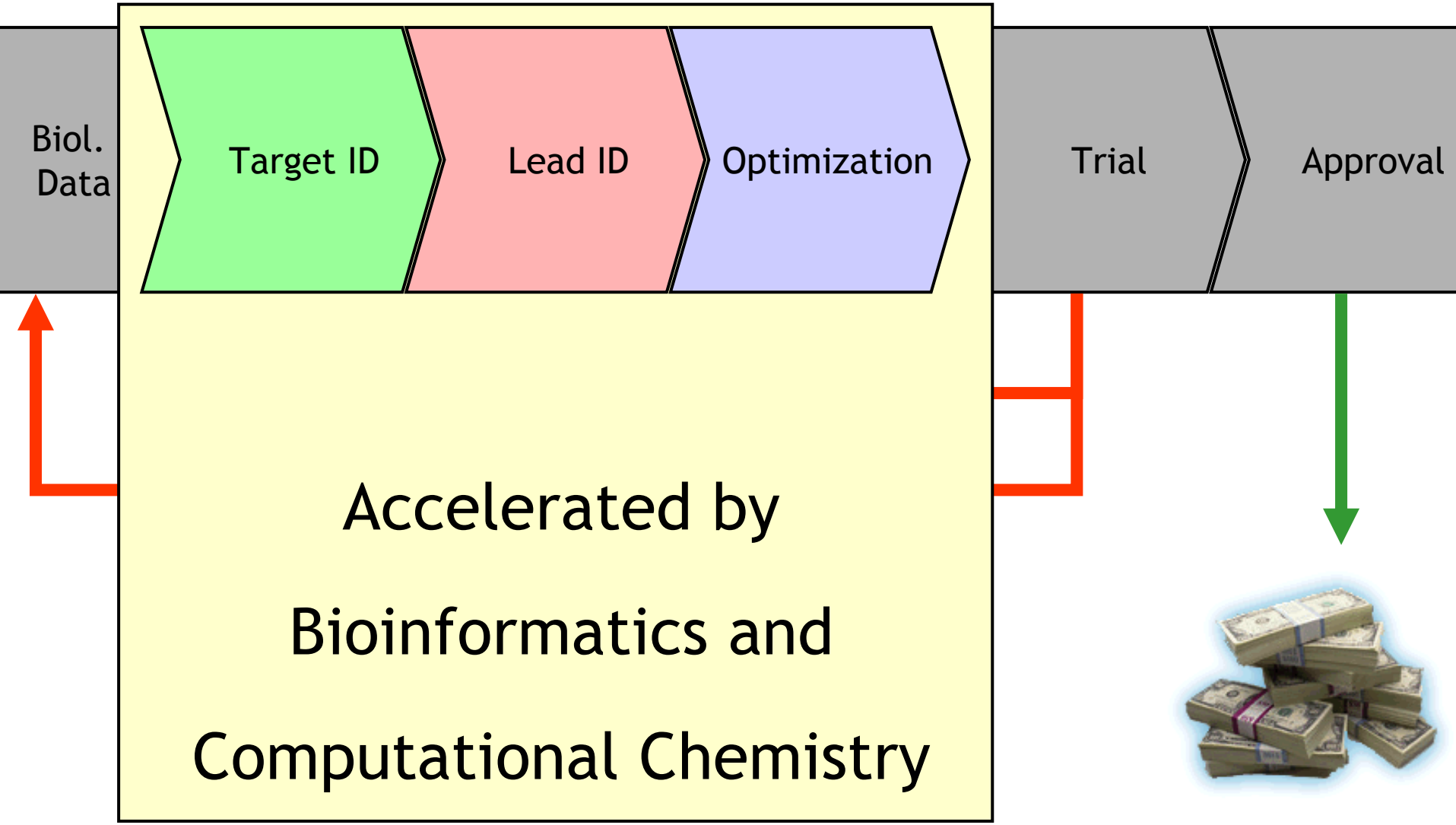
(www.expasy.org)

Target ID and biological networks



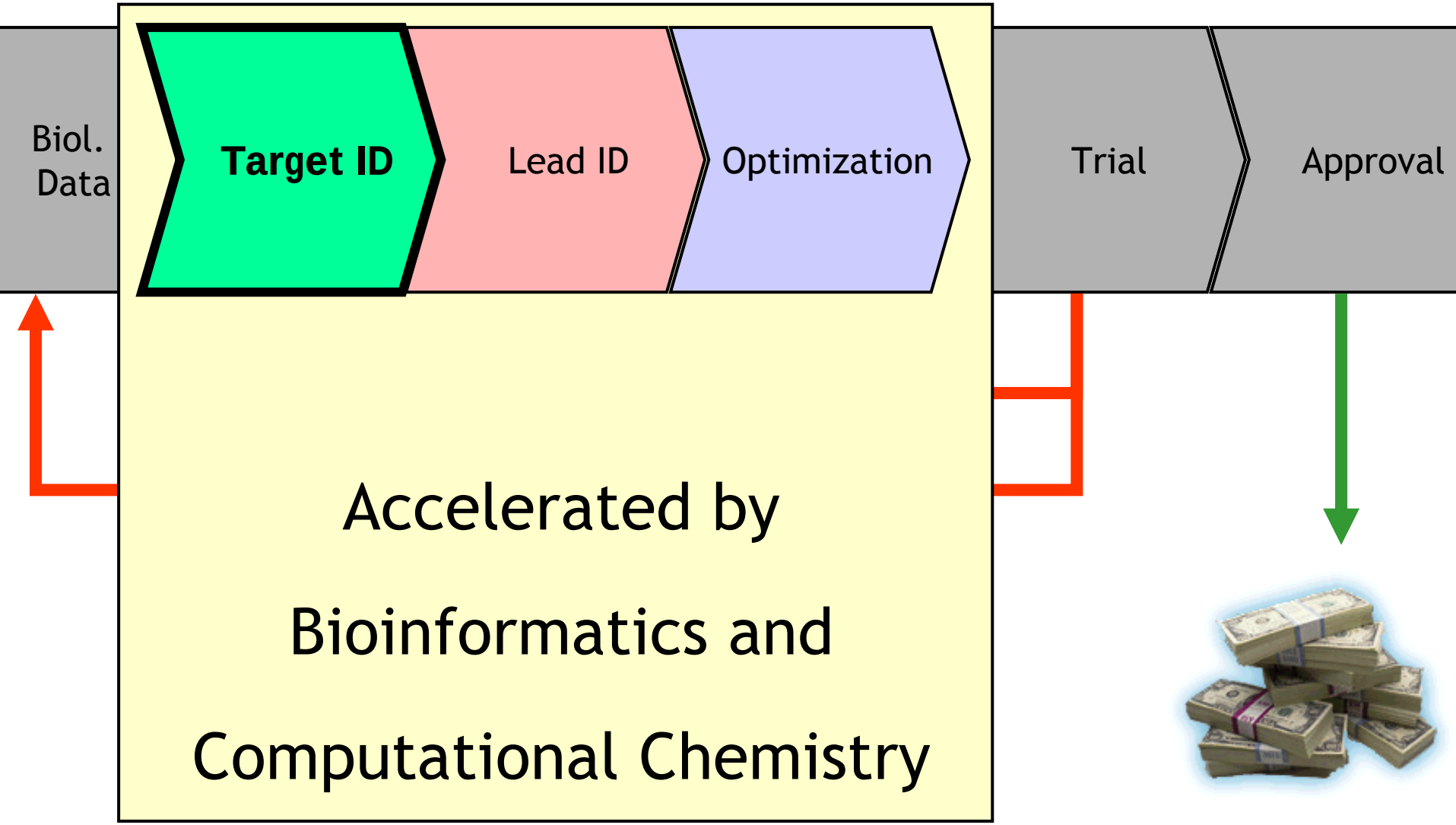
Drug Discovery Pipeline

Freie Universität Berlin



Drug Discovery Pipeline

Freie Universität Berlin



Application fields

Freie Universität Berlin

- **Diagnostics:** Find relevant patterns in one- or two-dimensional LC measurements
- **Time series:** Analyze the temporal behaviour in a time series experiment
- **Quantitative Measurements:** Determine absolute content of peptides using additive method (Myoglobin, Gliadin)

Differential Analysis in Proteomics

All techniques share the following steps:

1. Separation of proteins/peptides.

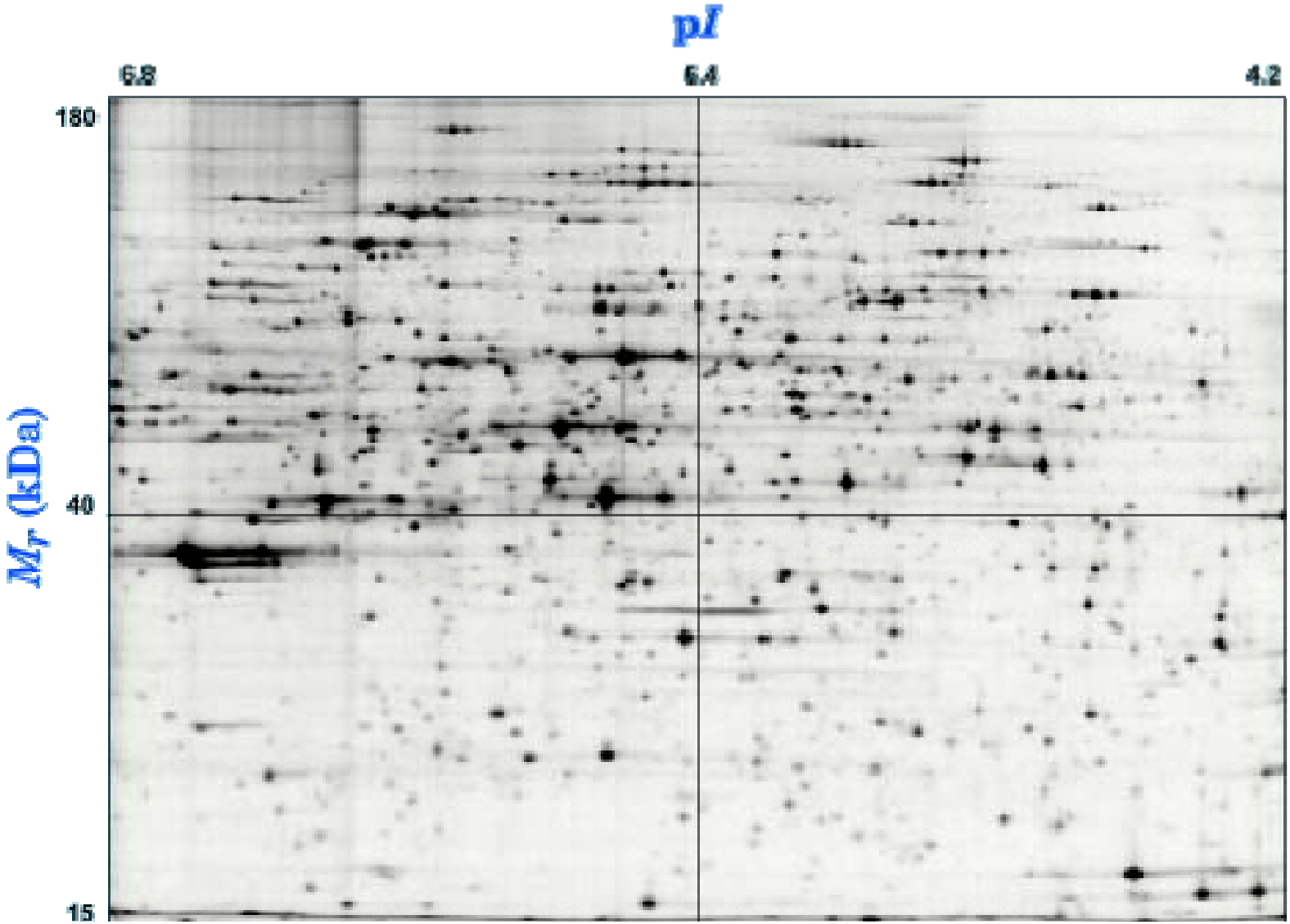
   Typical techniques: 2D PAGE, HPLC, DC, etc.

2. Assignment of proteins/peptides for *relative* quantitation

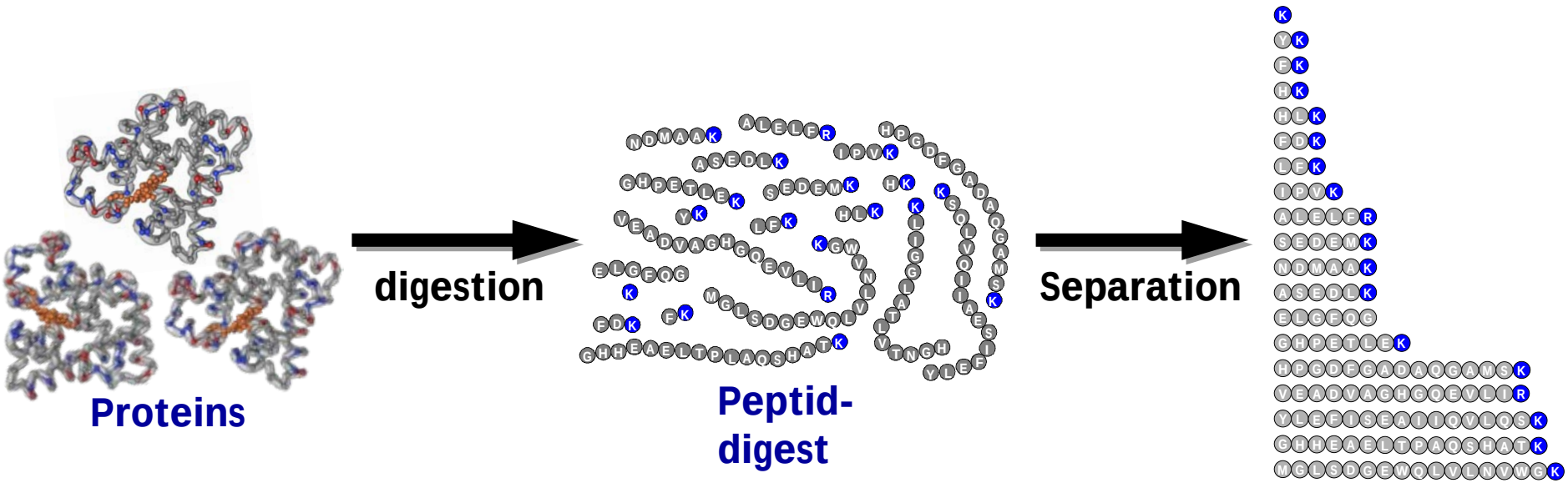
3. (Relative) Quantitation

4. Identification of peptides/proteins

Separation (2D-PAGE)

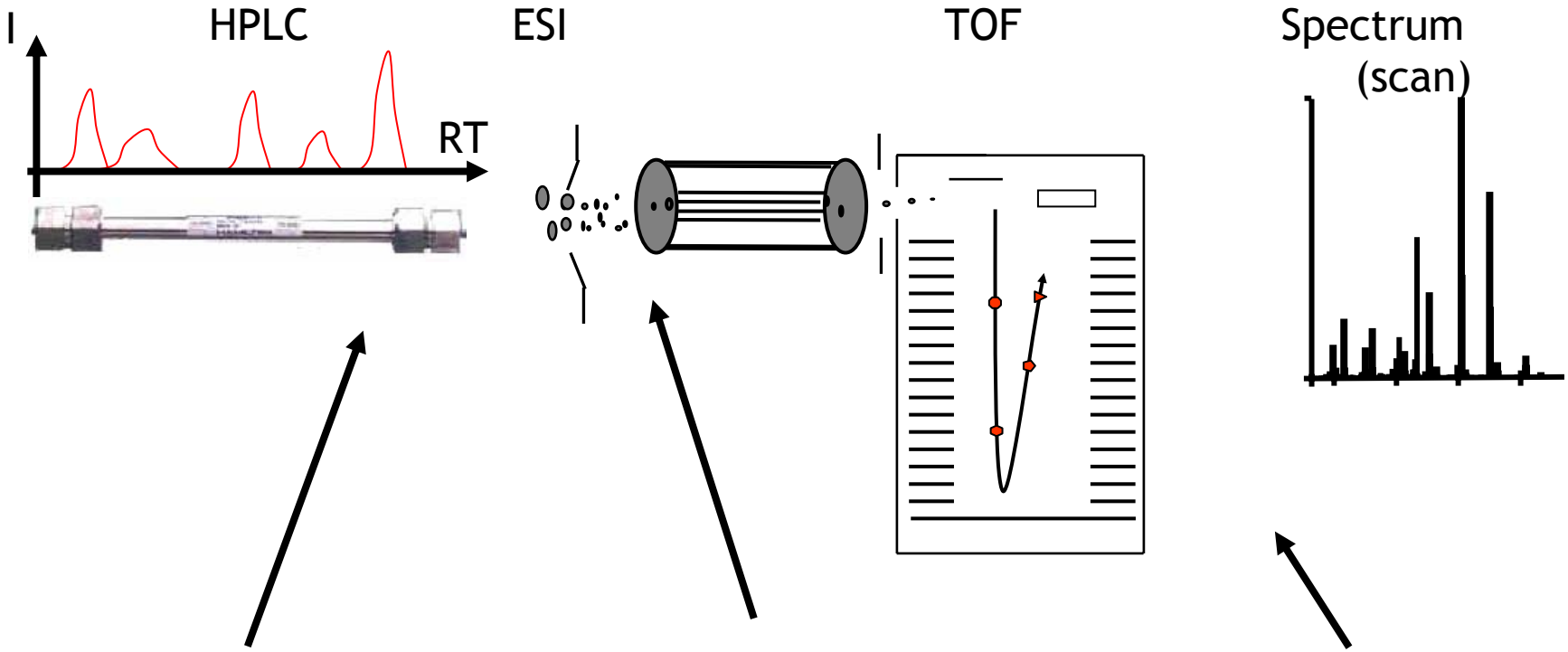


Separation (Shotgun Proteomics)



- **Key idea**
  - **Separation of whole proteins possible but difficult, hence digestion preferred**
  - **Separate peptides**
  - **Identify proteins through peptides**

HPLC-MS Analysis

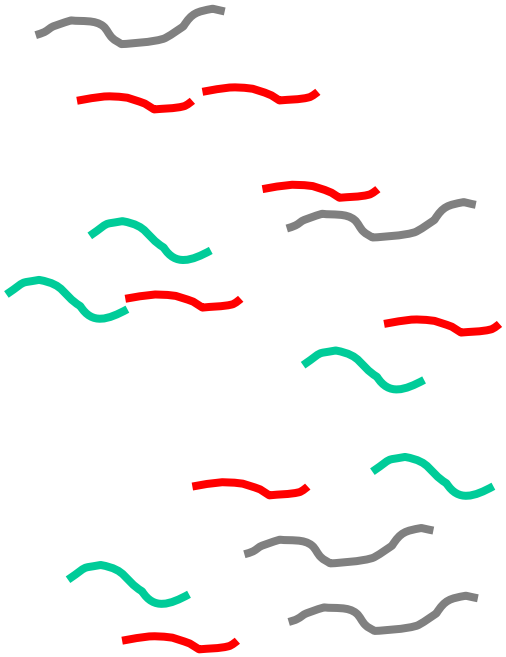


Separation 1
Different peptides have different retention time

Ionization
Peptide receives z charge units

Separation 2
Detector measures m/z

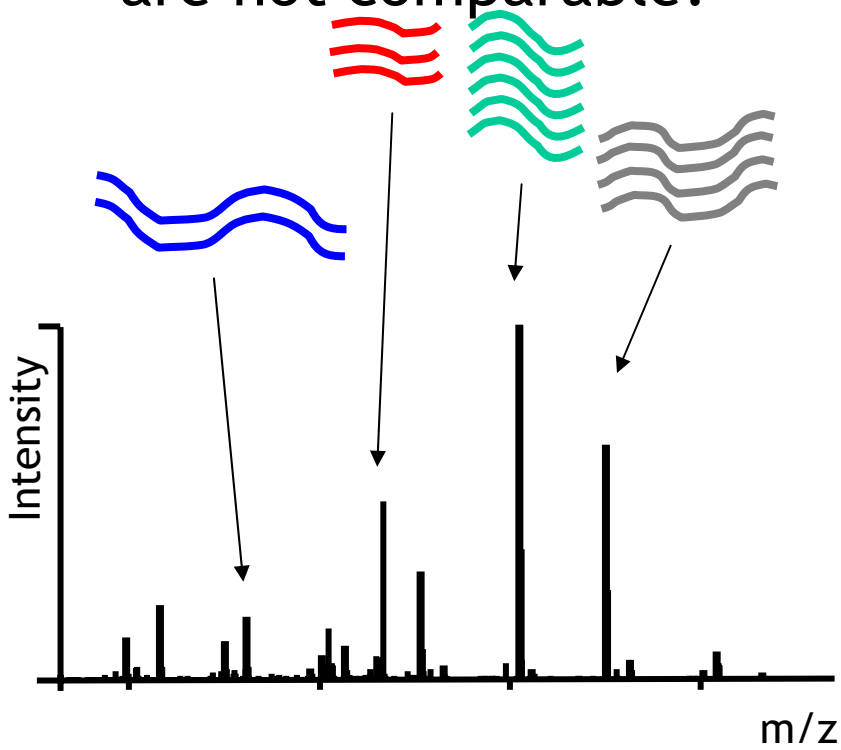
Single stage MS scan



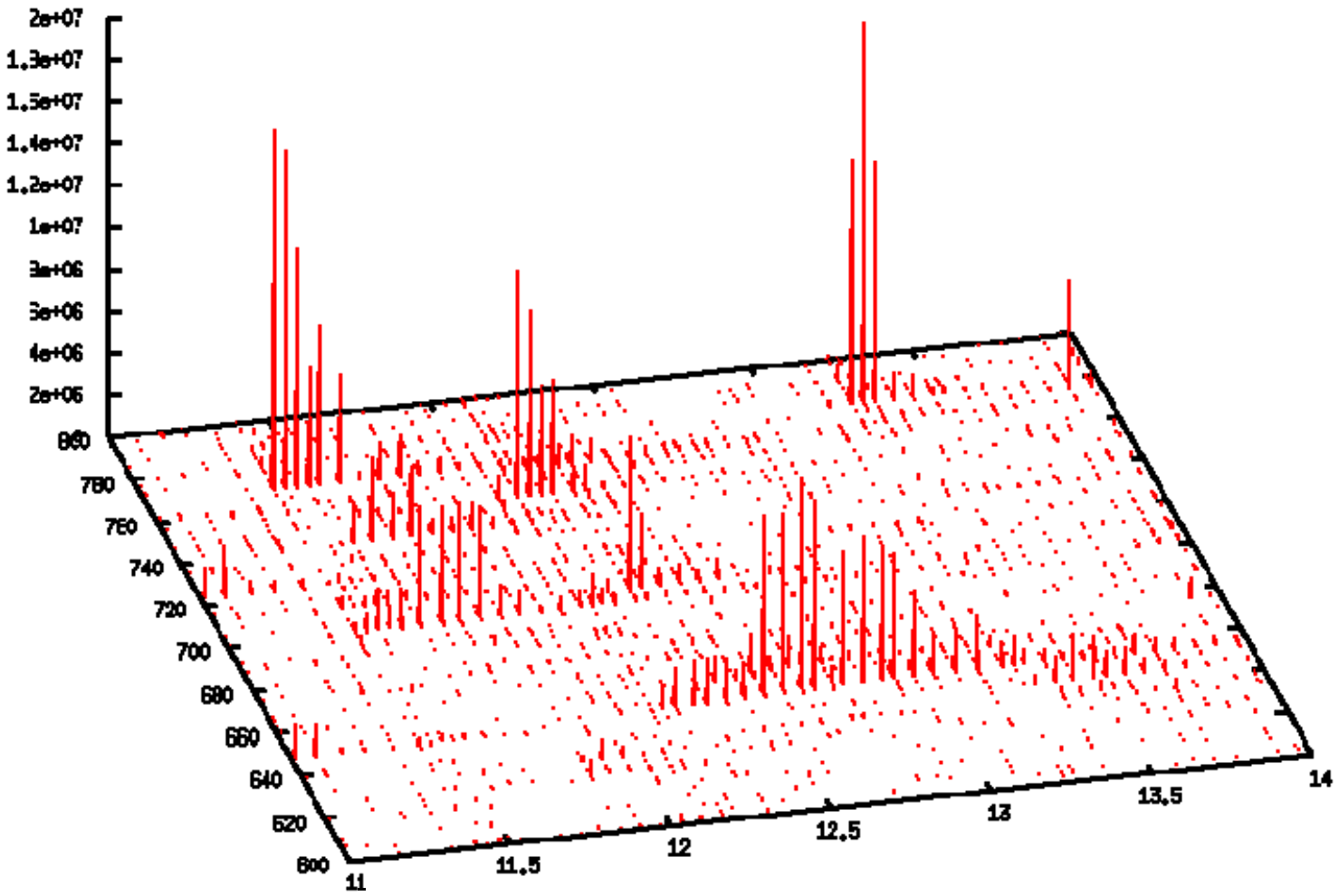
Mass spectrometry

Measures mass/charge ratio of ionized peptides for a short period of time

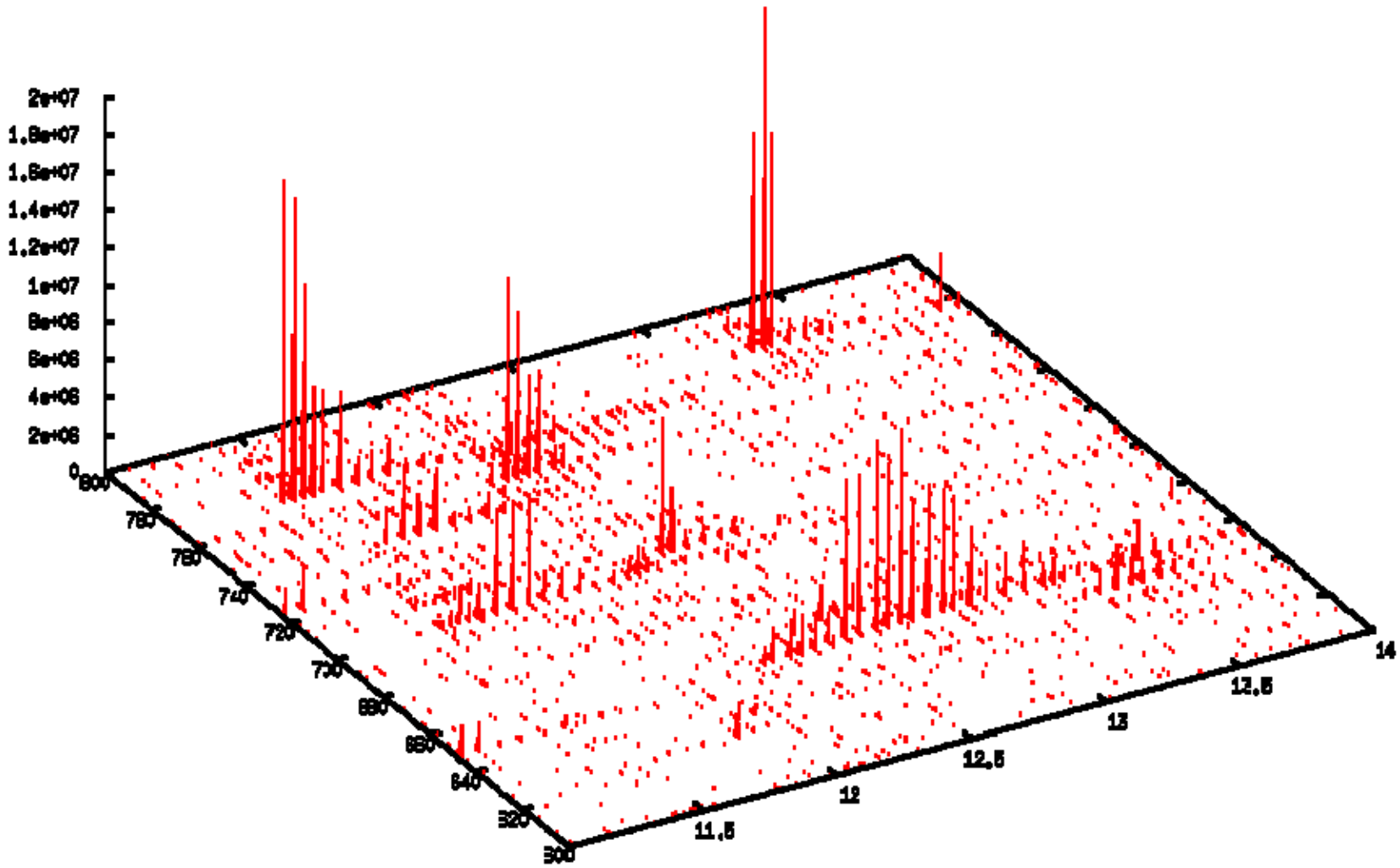
Peak intensity in scan corresponds to amount present, but intensities are not comparable!



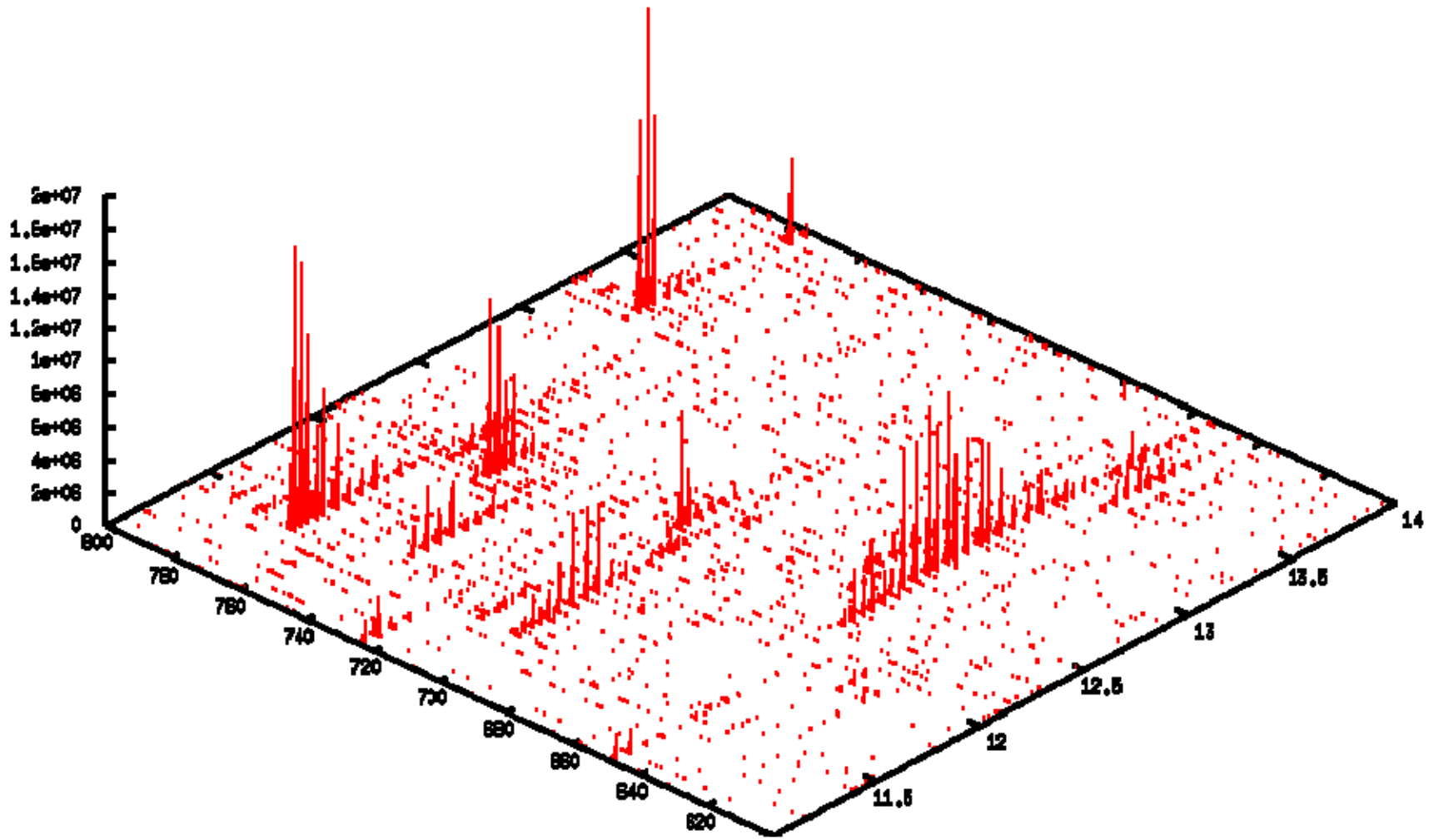
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
800
780
760
740
720
700
680
660
640
620
600
11
11.5
12
12.5
13
13.5
14

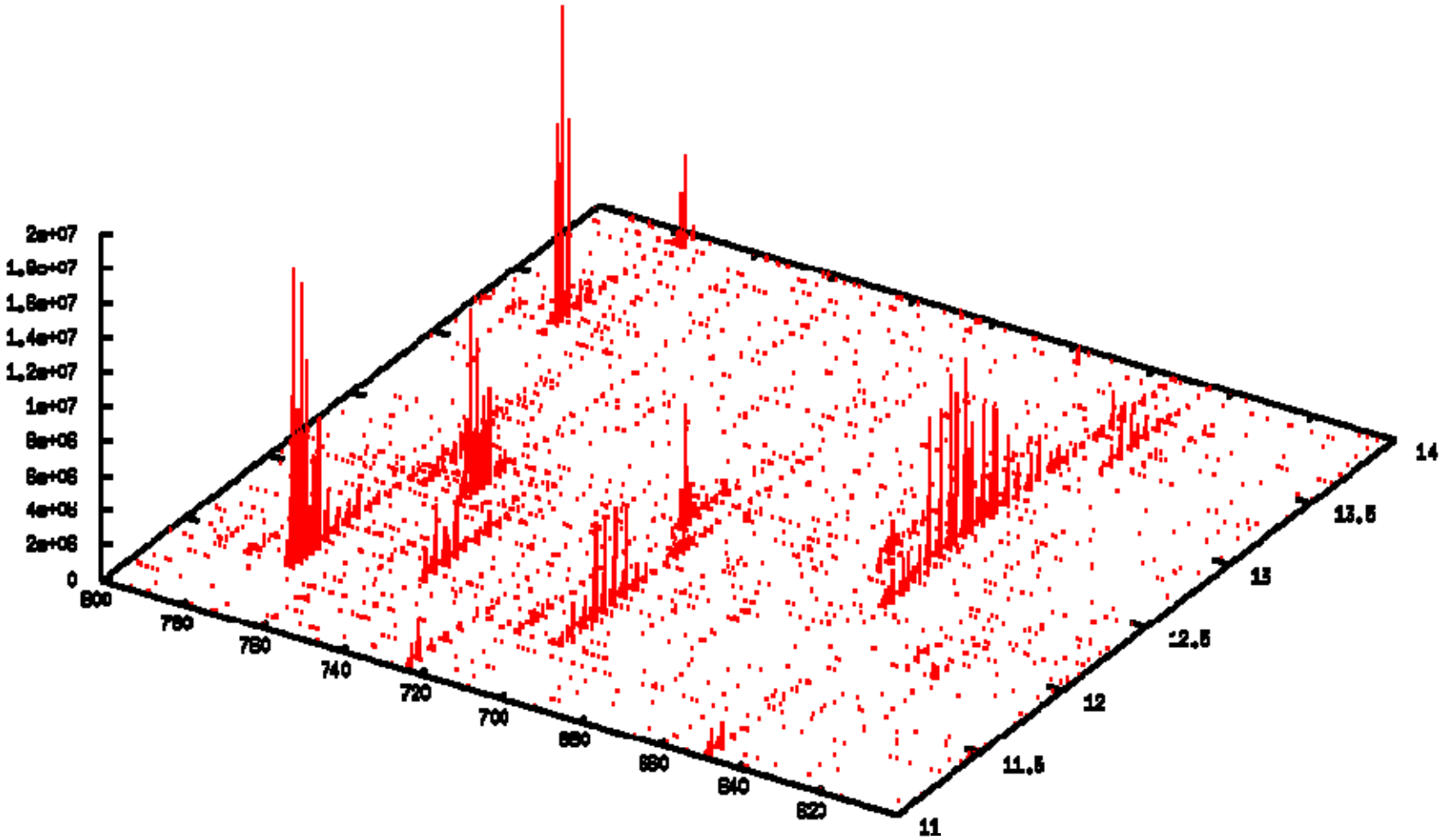
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
800
760
720
680
640
600
11.5
12
12.5
13
13.5
14

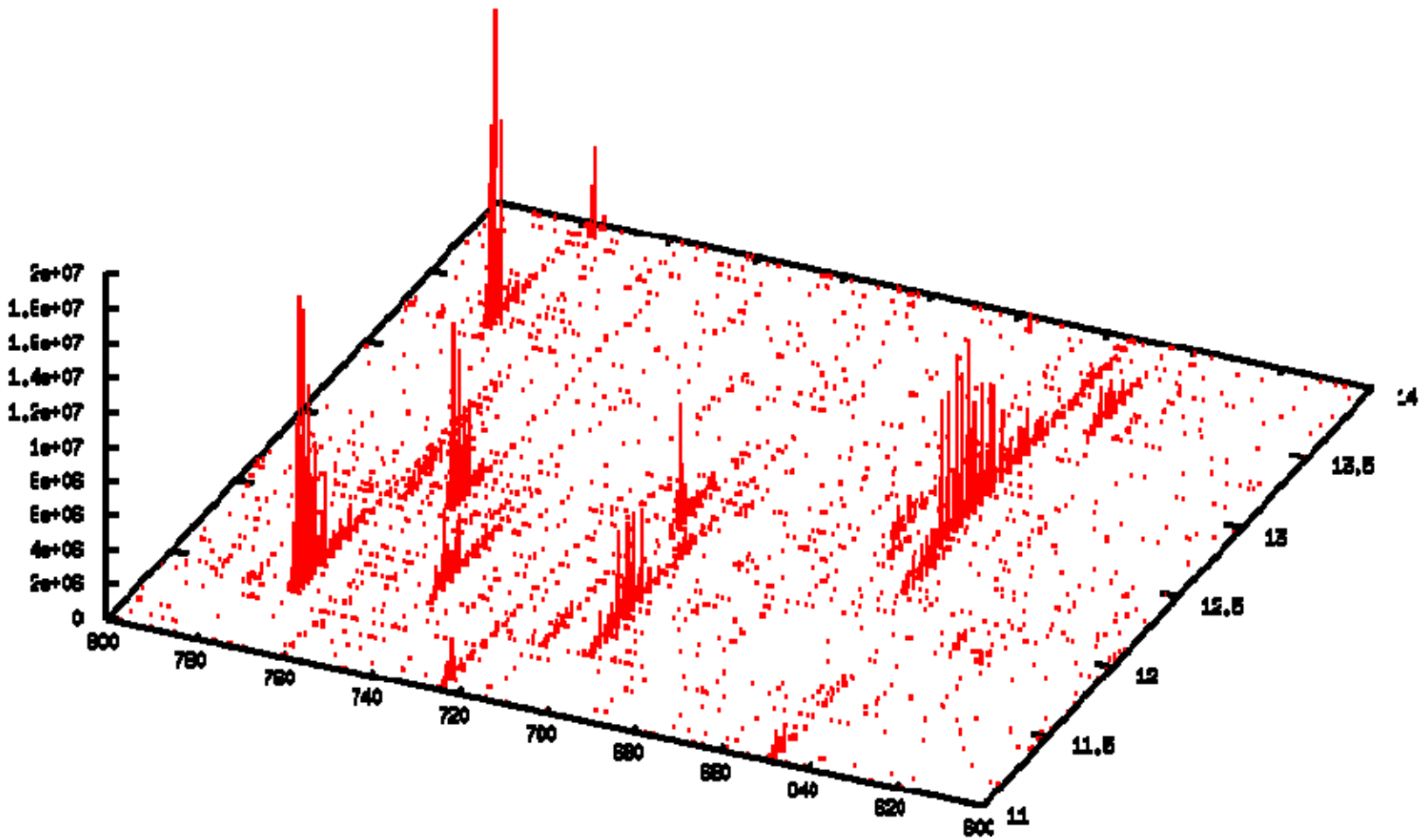
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
800
780
760
740
720
700
680
660
640
620
11.5
12
12.5
13
13.5
14

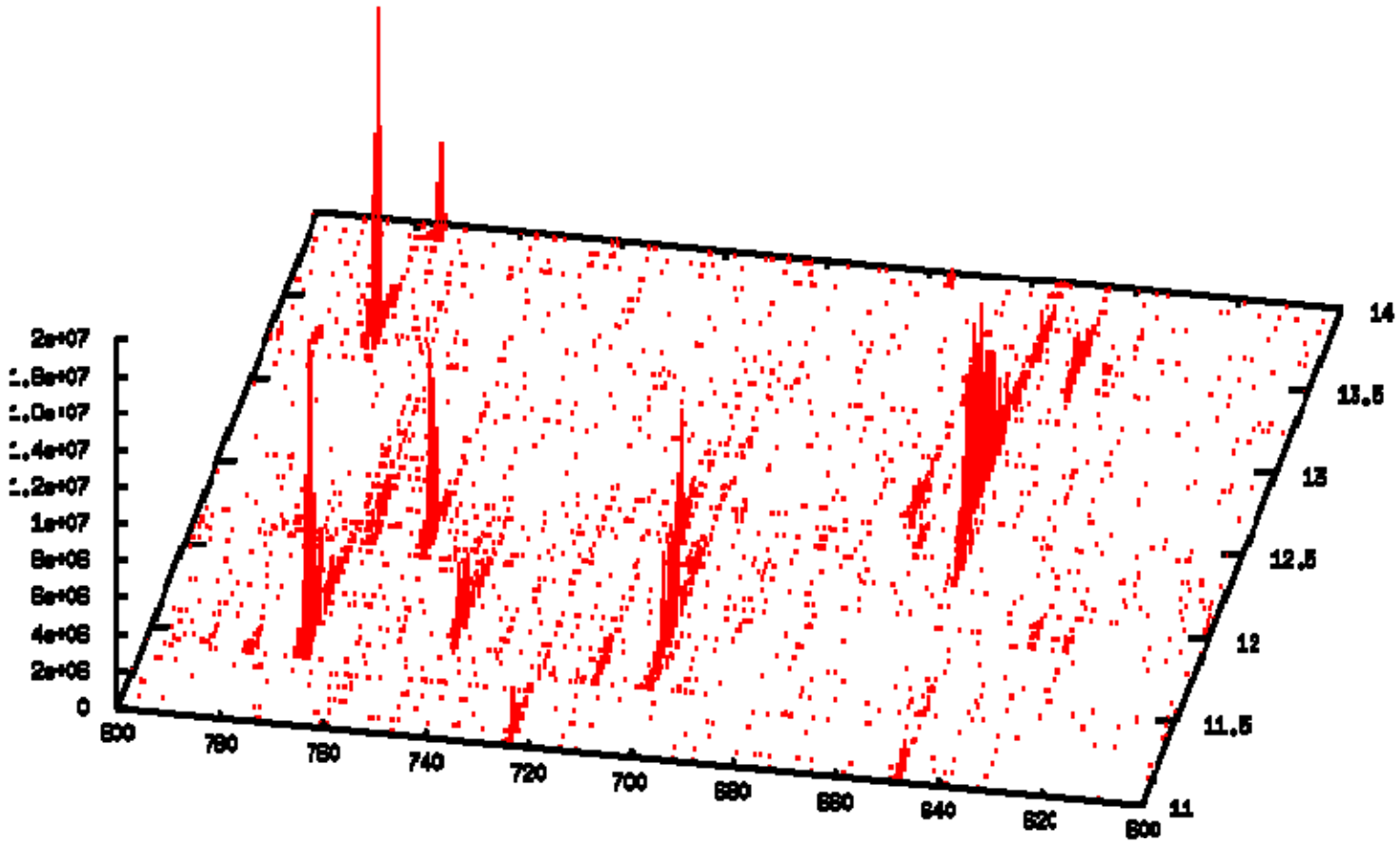
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
800
760
760
740
720
700
680
660
640
620
11
11.5
12
12.5
13
13.5
14

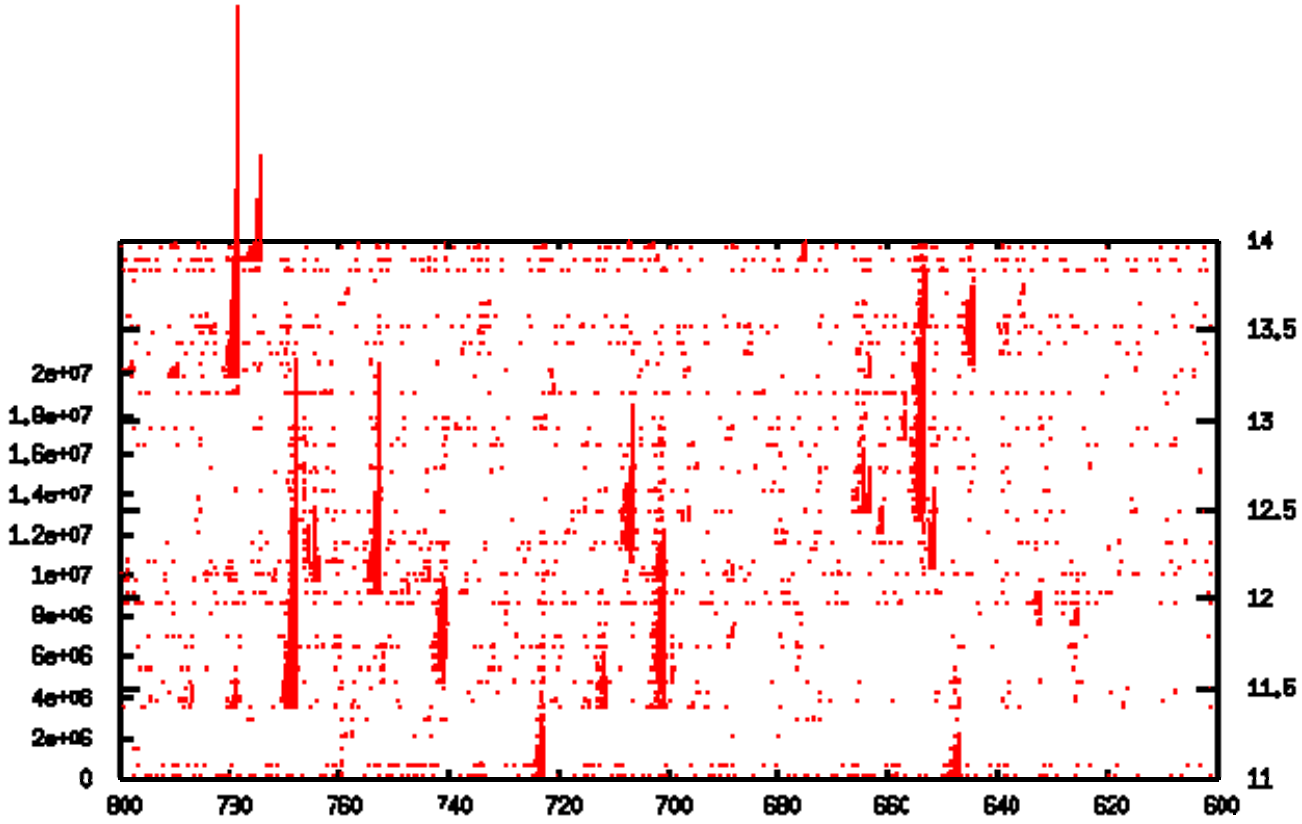
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
800
780
760
740
720
700
680
660
640
620
600
11
11.5
12
12.5
13
13.5
14

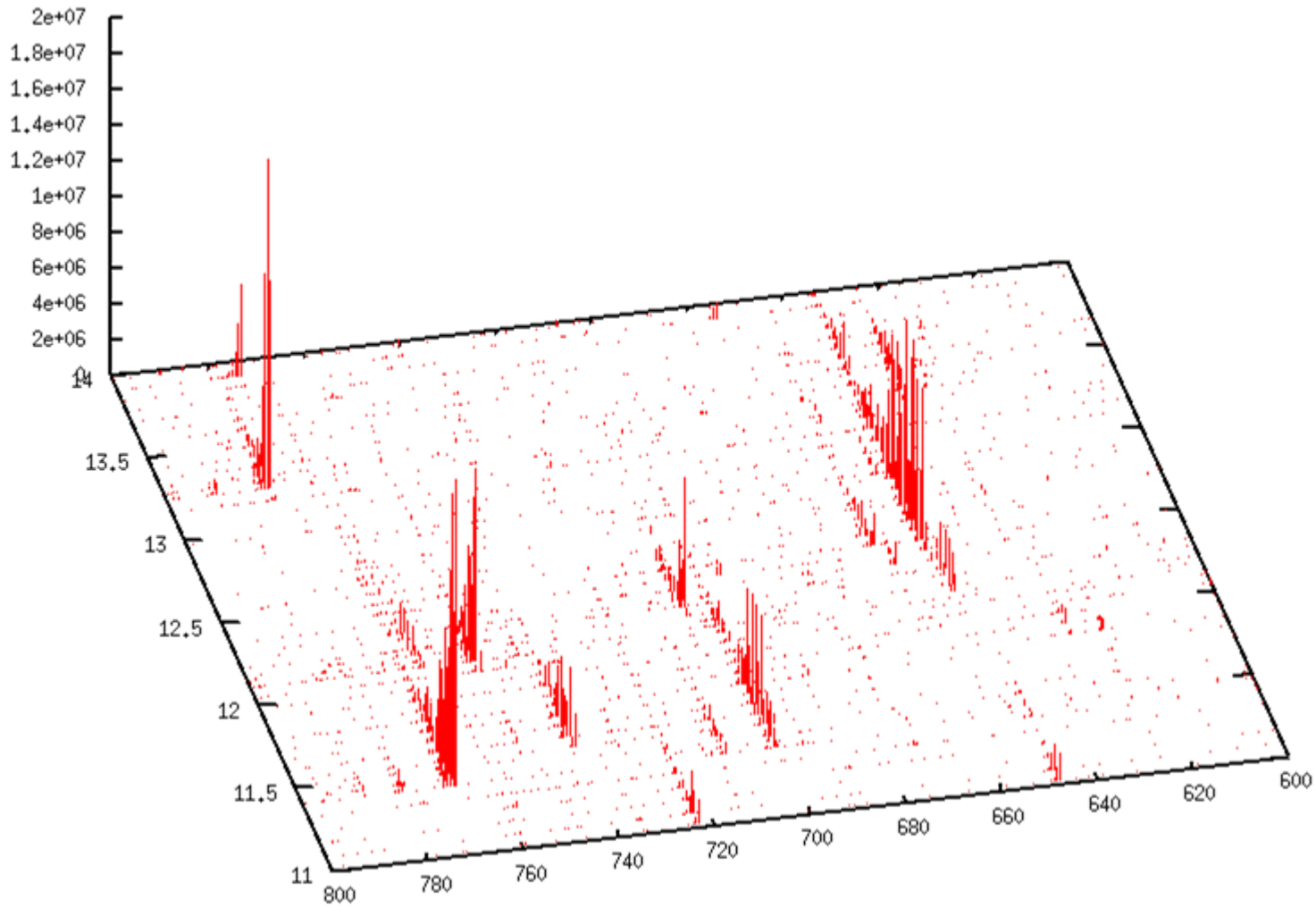
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
800
780
760
740
720
700
680
660
640
620
600
14
13.5
13
12.5
12
11.5
11

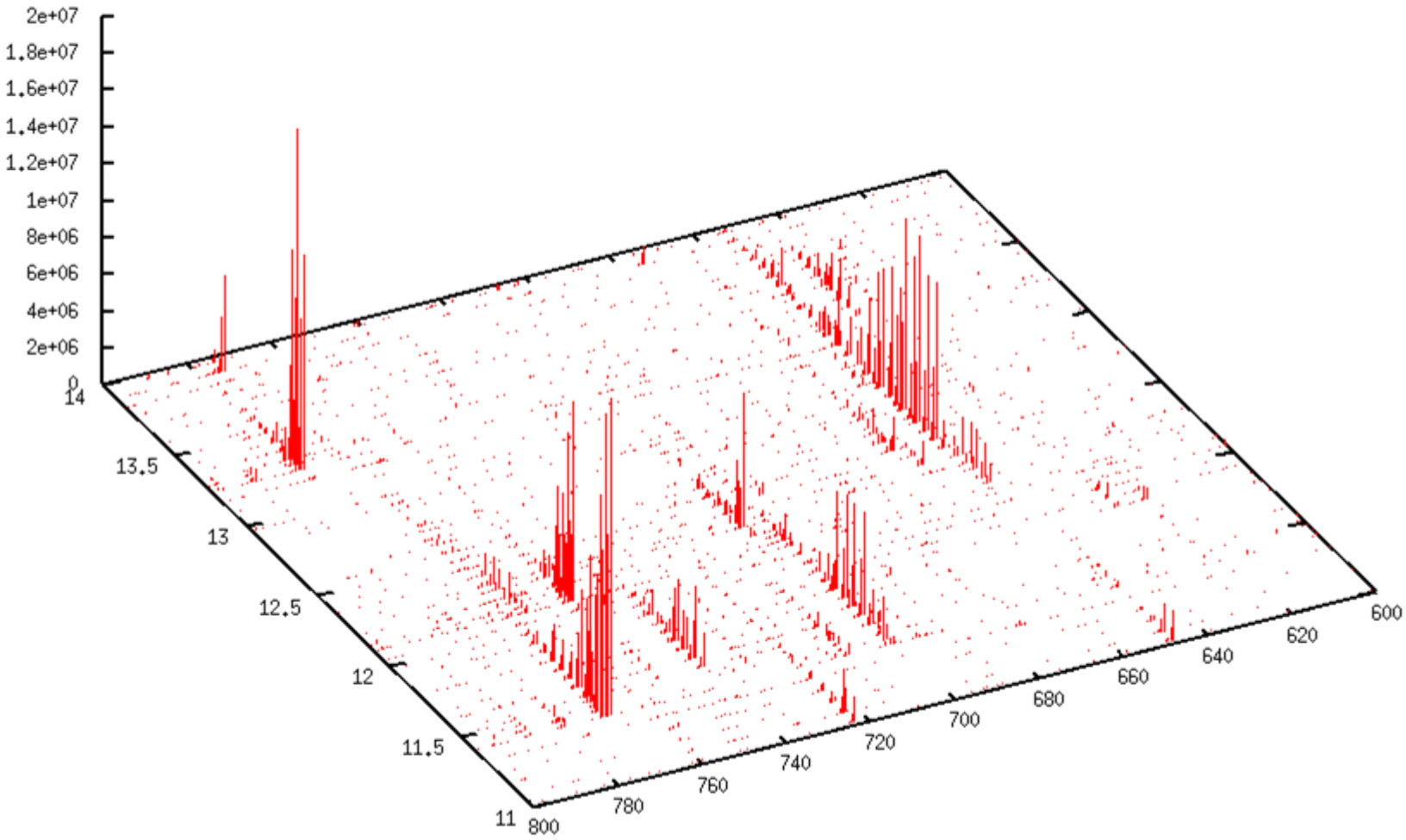
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
800
780
760
740
720
700
680
660
640
620
600
14
13.5
13
12.5
12
11.5
11

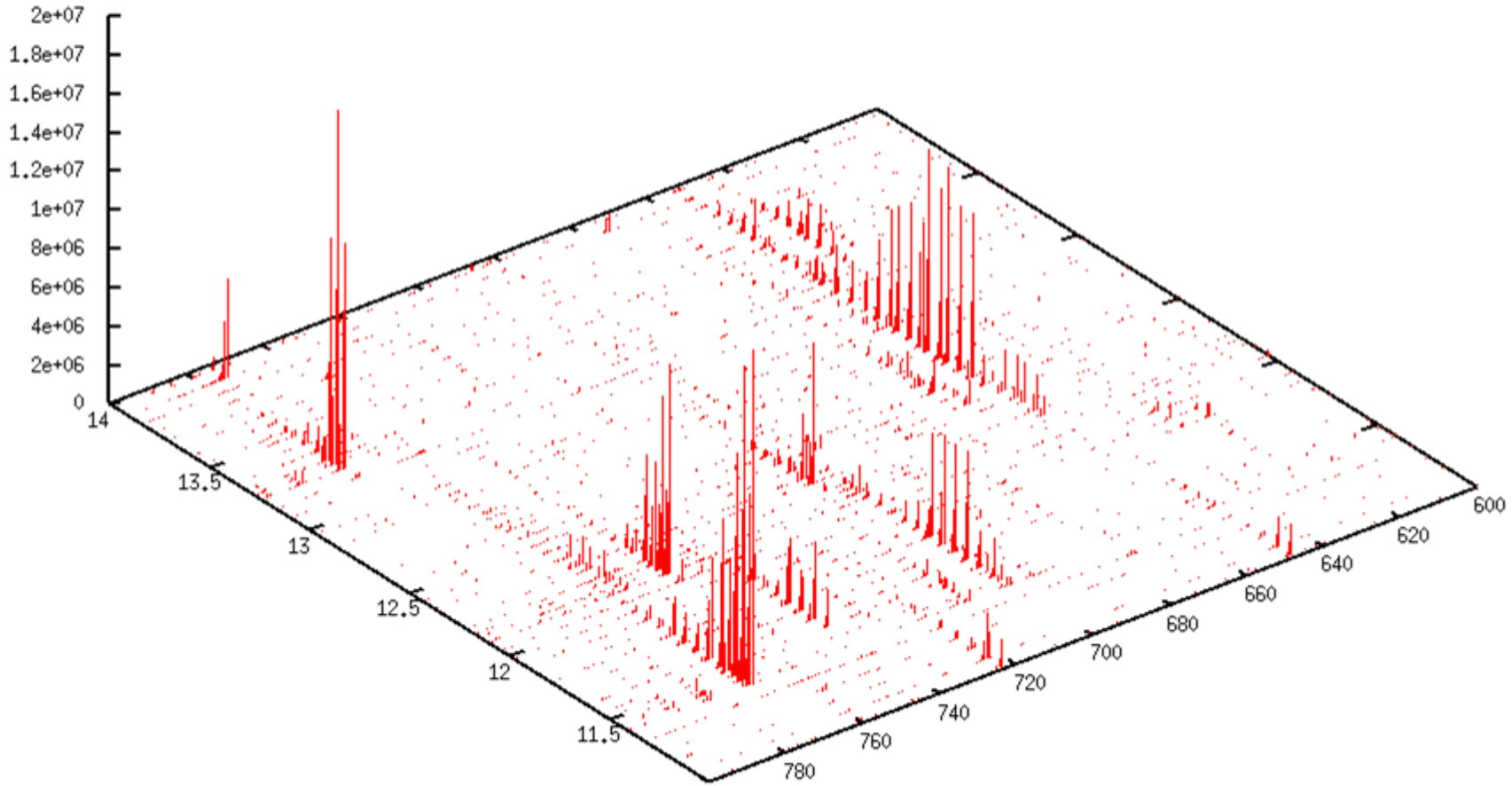
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
14
13.5
13
12.5
12
11.5
11
800
780
760
740
720
700
680
660
640
620
600

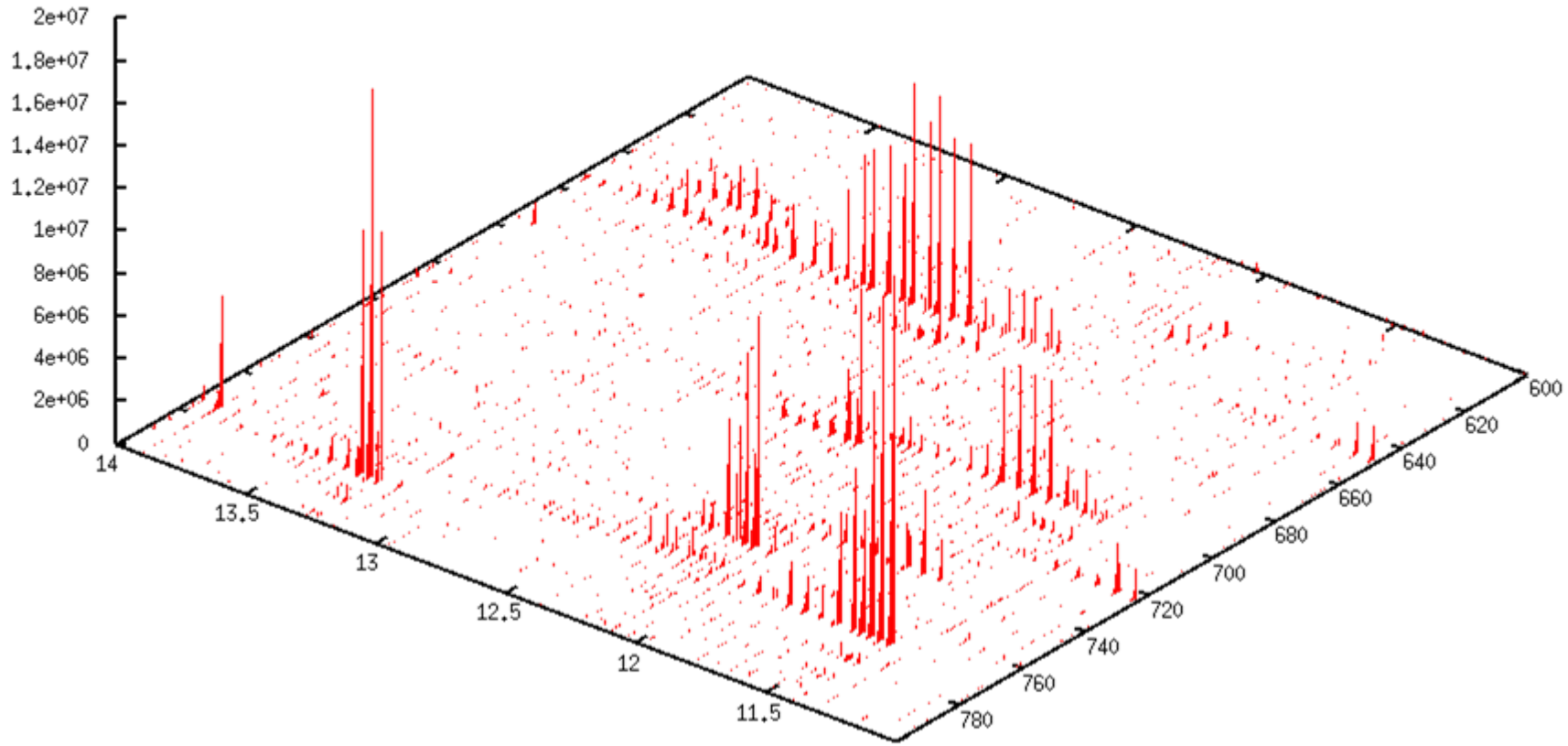
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
14
13.5
13
12.5
12
11.5
11
800
780
760
740
720
700
680
660
640
620
600

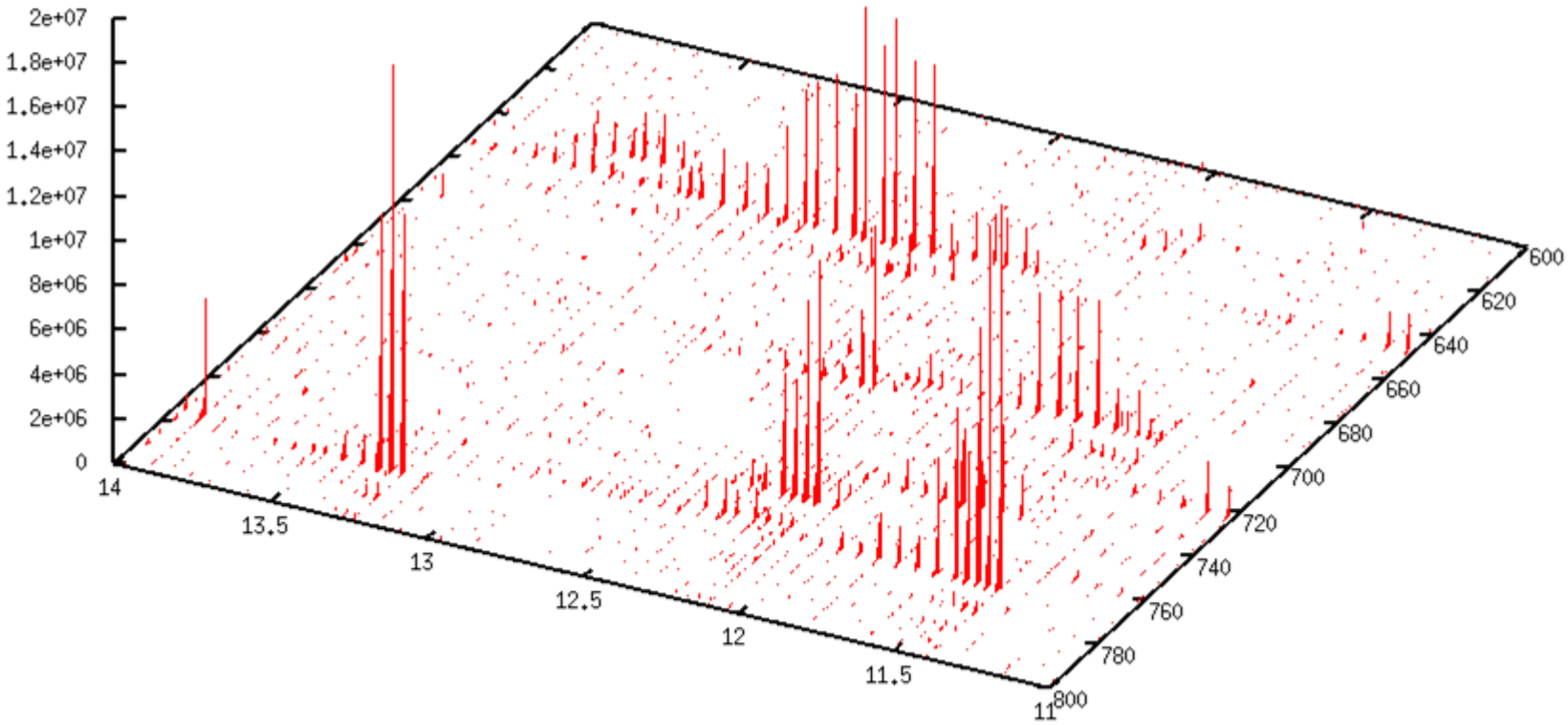
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
14
13.5
13
12.5
12
11.5
600
620
640
660
680
700
720
740
760
780

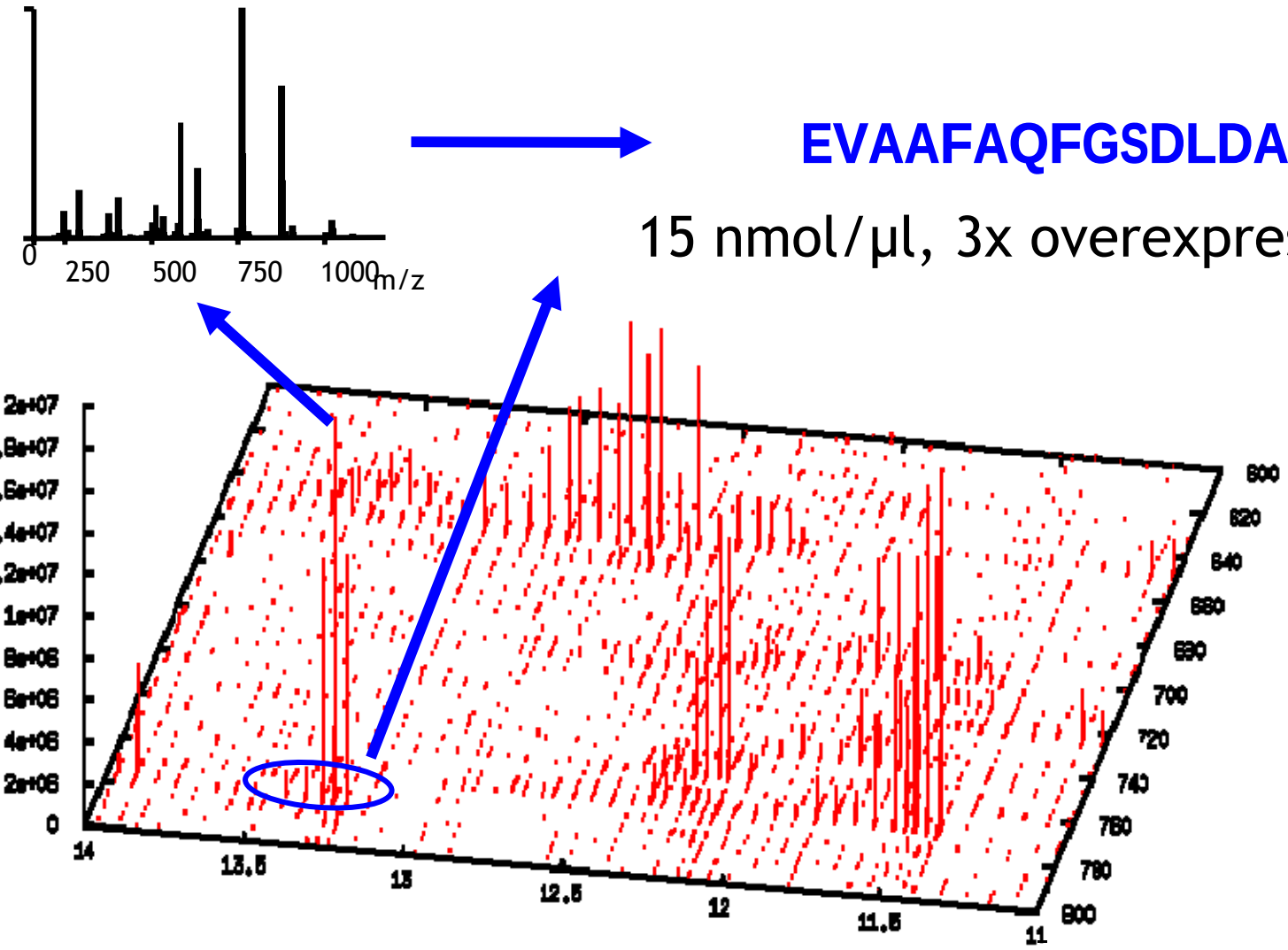
Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
14
13.5
13
12.5
12
11.5
600
620
640
660
680
700
720
740
760
780

Maps


2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
14
13.5
13
12.5
12
11.5
11
600
620
640
660
680
700
720
740
760
780
800

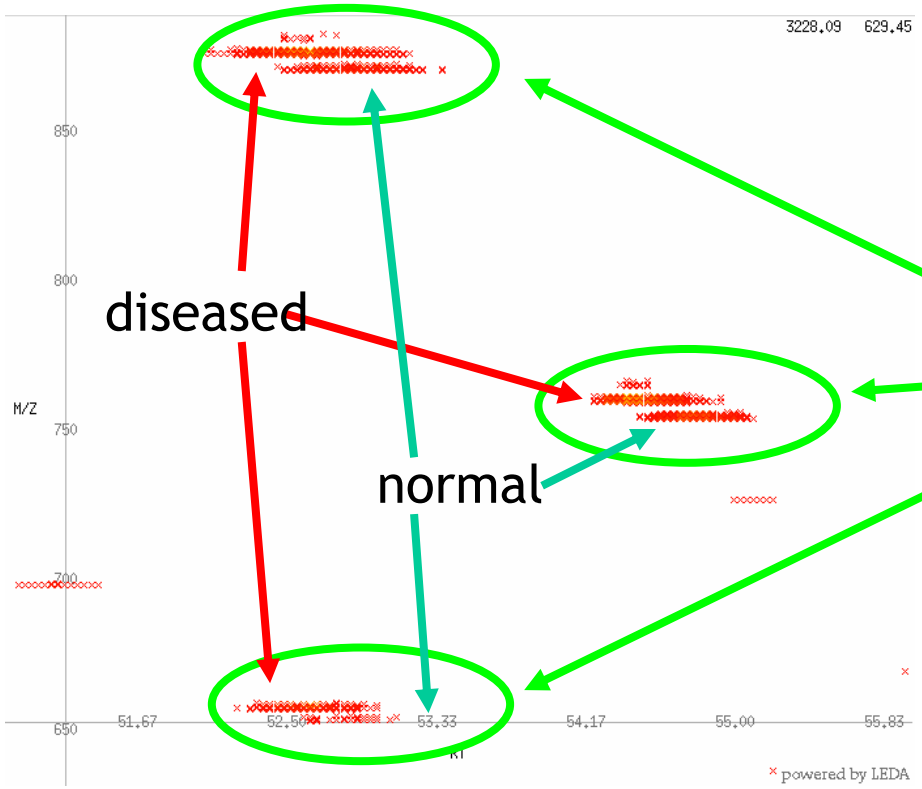
Maps


0
250
500
750
1000
m/z
EVAAFAQFGSDLDASTK
15 nmol/µl, 3x overexpressed, …
2e+07
1.8e+07
1.6e+07
1.4e+07
1.2e+07
1e+07
8e+06
6e+06
4e+06
2e+06
0
14
13.5
13
12.5
12
11.5
11
800
820
840
860
880
700
720
740
760
780
800

Differential Analysis

Two common basic approaches:

- Isotope labeling (e.g. ICAT, MeCAT, SILAC,…)
- Direct Differential Quantification (DDQ)



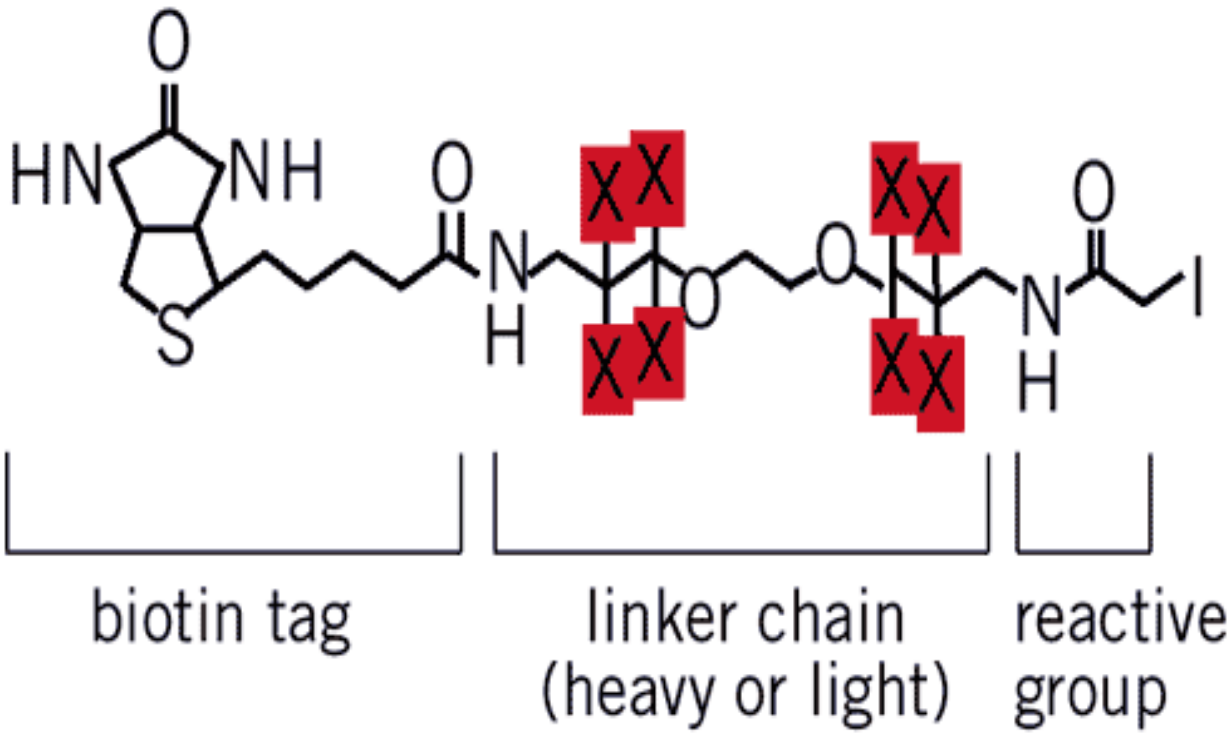
ICAT pairs at known distance

Isotope Labeling (ICAT)

Isotope-Coded Affinity Tags

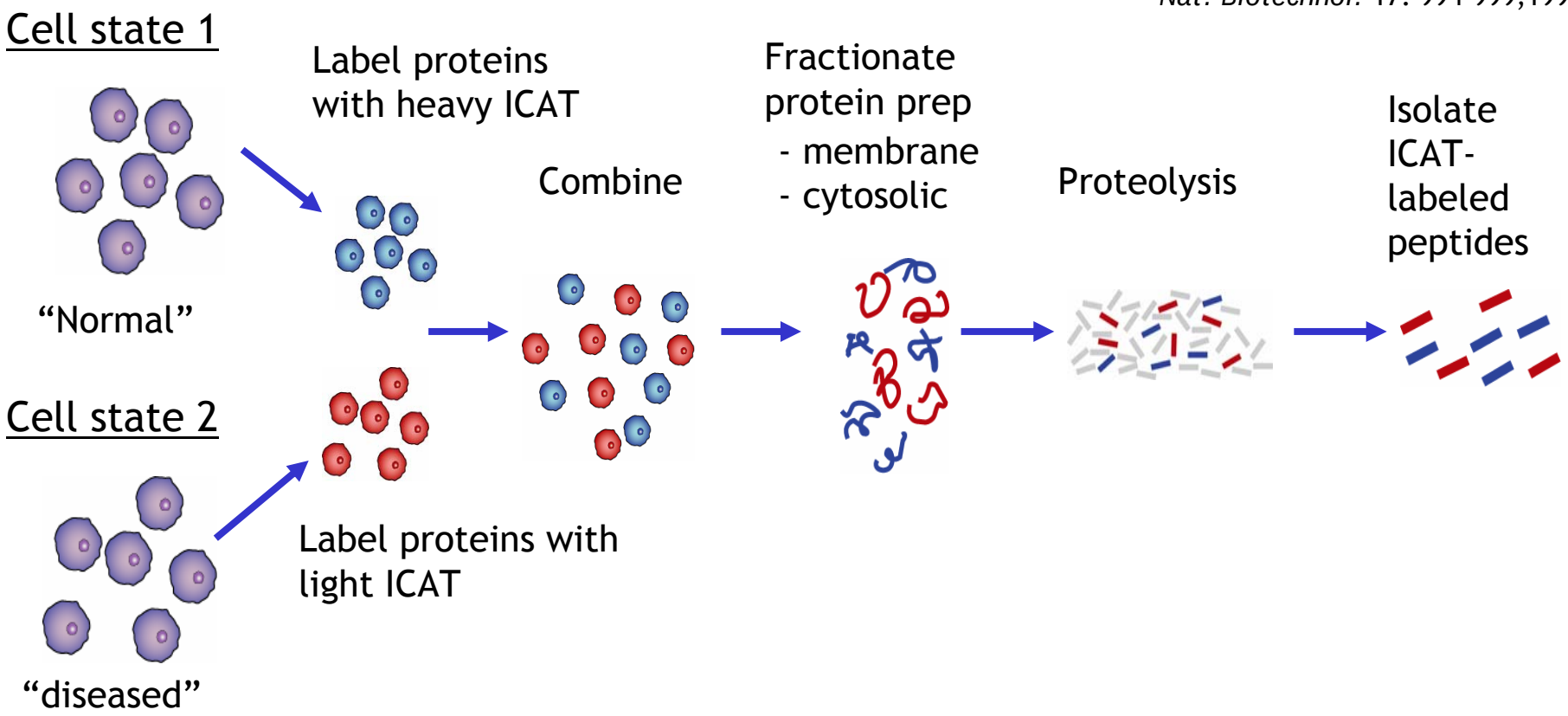
heavy reagent: D8-ICAT Reagent (X=deuterium)

light reagent: D0-ICAT Reagent (X=hydrogen)



Isotope Labeling (ICAT)

Nat. Biotechnol. 17: 994-999,1999

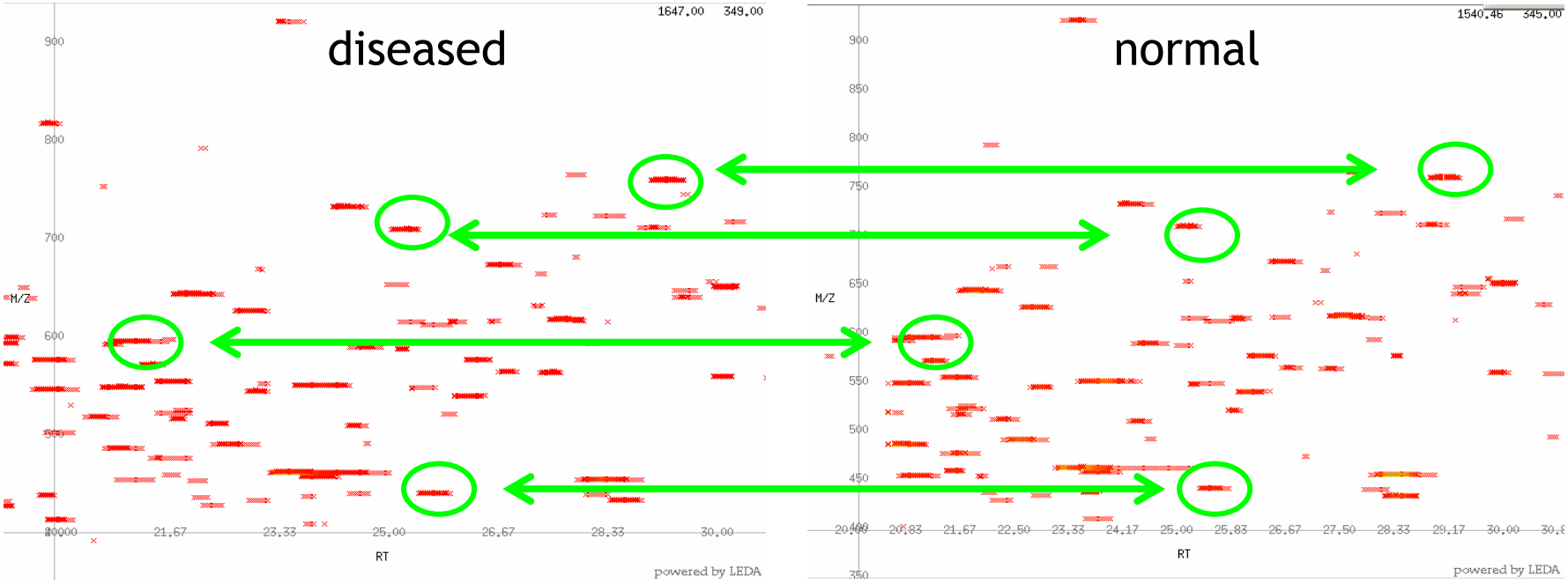


Heavy and light ICAT reagent 8 Dalton apart

Differential Analysis

Two common basic approaches:

- Isotope tagging (e.g. ICAT, MeCAT)
- Direct Differential Quantitation (DDQ)



What's in a Map?

Freie Universität Berlin

- **Retention time** (*RT*) for each scan
- **Peptide mass/charge ratio** (*m/z*)
  (usually within ~20 ppm)
- **Intensity** (*I*)

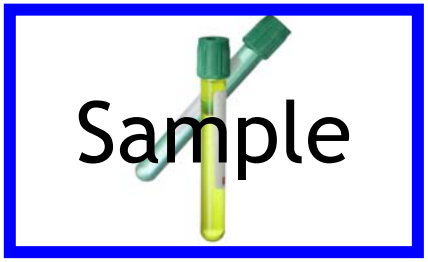
  ⇒ use *m/z*, *RT* to **identify** peptides

  ⇒ use I to **quantify** peptides

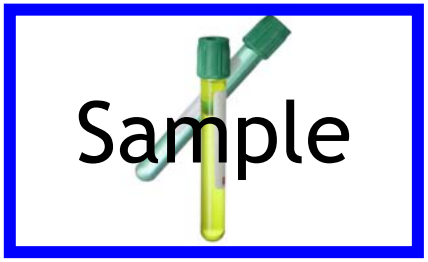
  (relative quantitation only!)

Maps become HUGE (10^8 Peaks)!

Proteomics Data Flow



Proteomics Data Flow



HPLC/MS

Raw Data

10 GB

Sig. Proc.

Filtered Raw Data

1 GB

Data Reduction

50 MB

Map

Diff. Anal.

Annotated Maps

50 MB

Identification

Differentially Expressed Proteins

1 kB

Advantages of HPLC/MS over 2D PAGE

1. Easier automation
   - no robots
   - no gel handling
2. Better separation power
   - Monolithic columns
   - Multi-dimensional chromatography
3. Simpler coupling to MS

HPLC/MS – Becoming the Standard?

Disadvantages of HPLC/MS

- Proteins are chopped up
- Quantitation difficult
- Huge amount of data (10 GiB/run)
- Data hard to manage/interpret

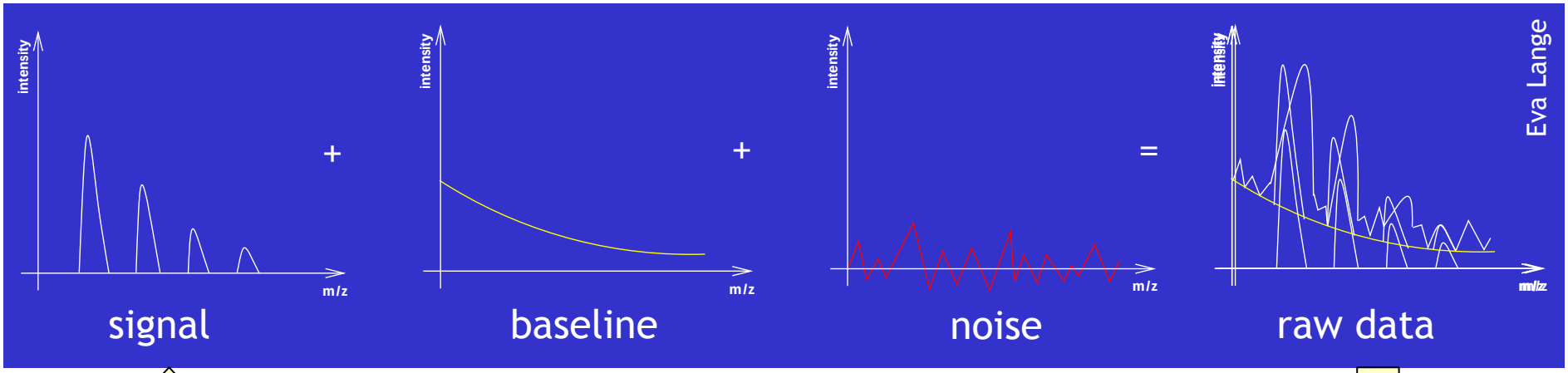
⇒ **Need for Informatics!**

Overview

Freie Universität Berlin

- Motivation for OpenMS
- Bioinformatics Issues in quantitative Proteomics
  - Signal Processing
  - Feature Finding
  - Map Mapping
  - Differential Quantitation
  - Identification, Clustering
  - Software Engineering, Databases

Signal Processing

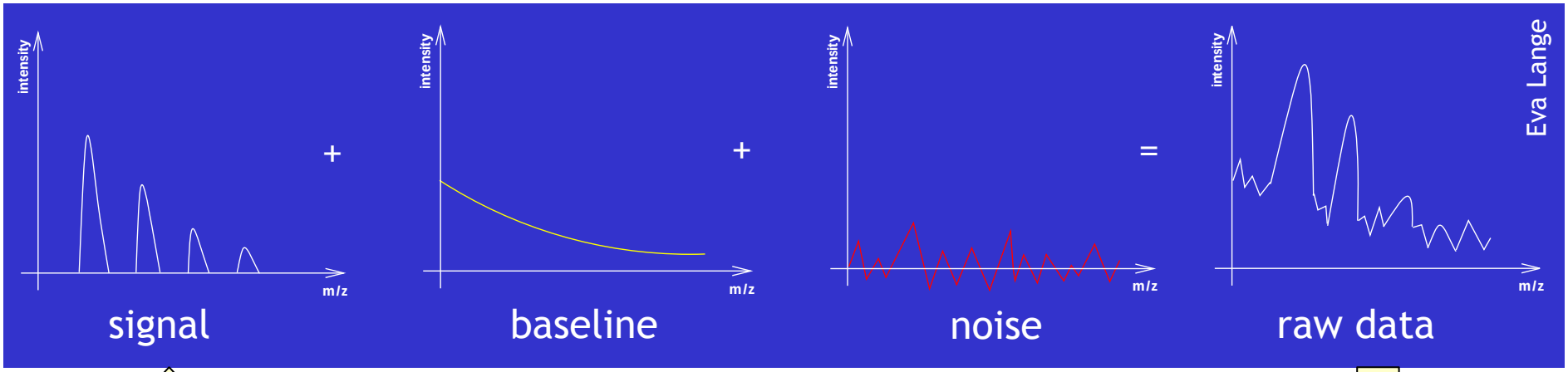


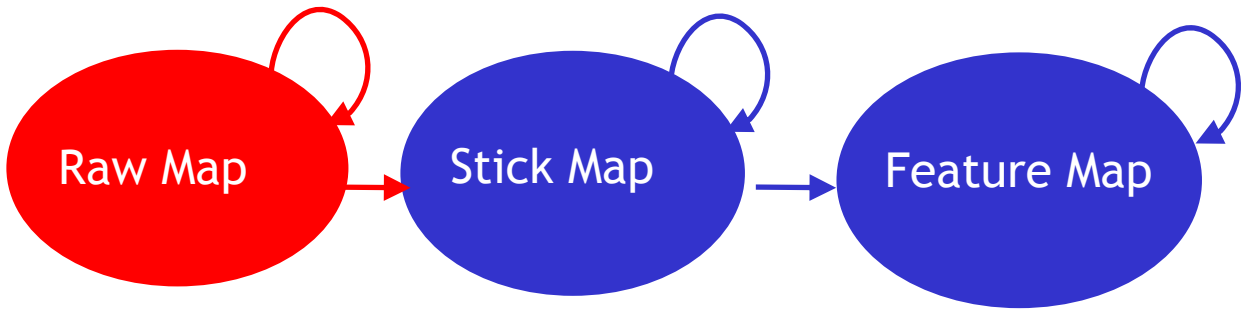
Issues

- Baseline reduction
- Noise filtering
- Calibration

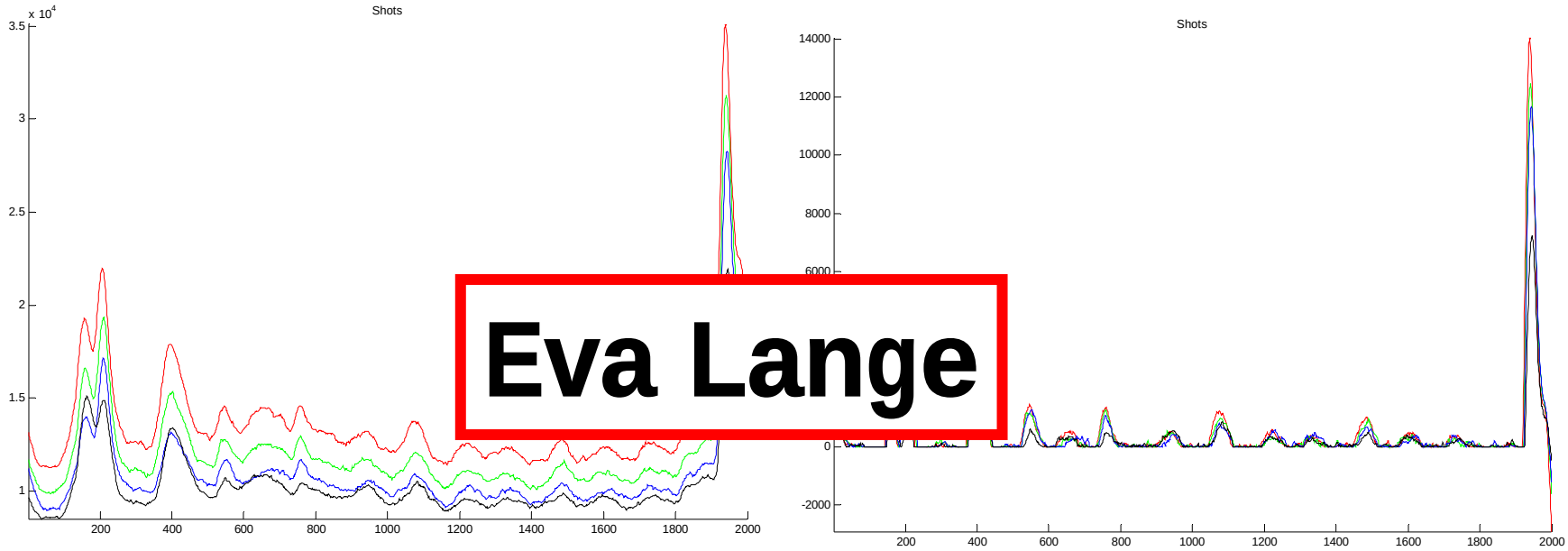
⇒ **Increase reliability of data**

Signal Processing



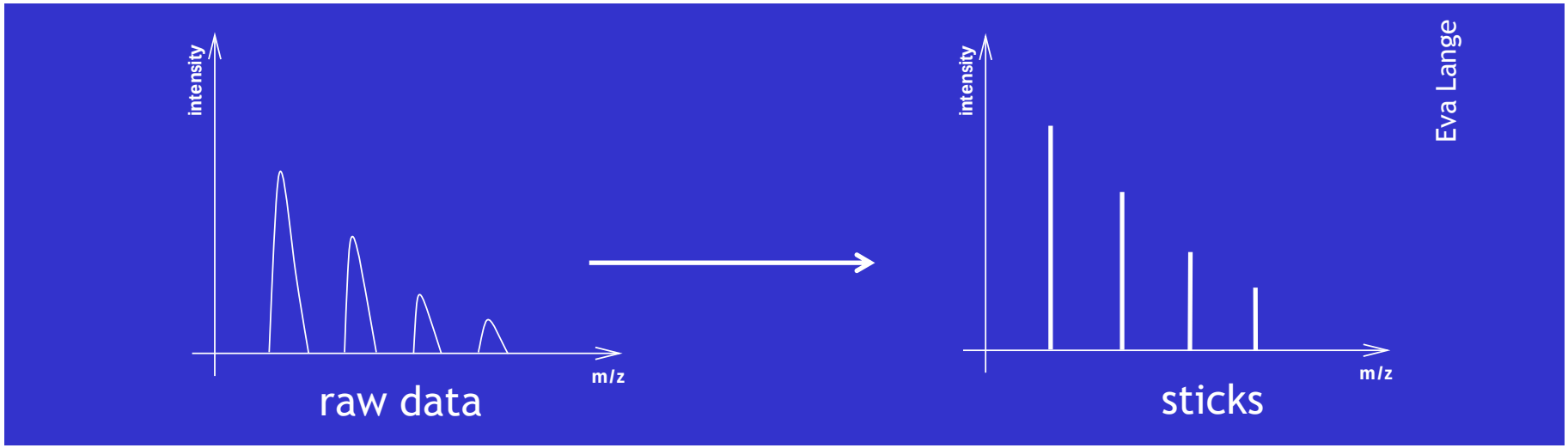


Baseline Filtering



TopHat filtering determines baseline

Stick conversion aka Peak picking

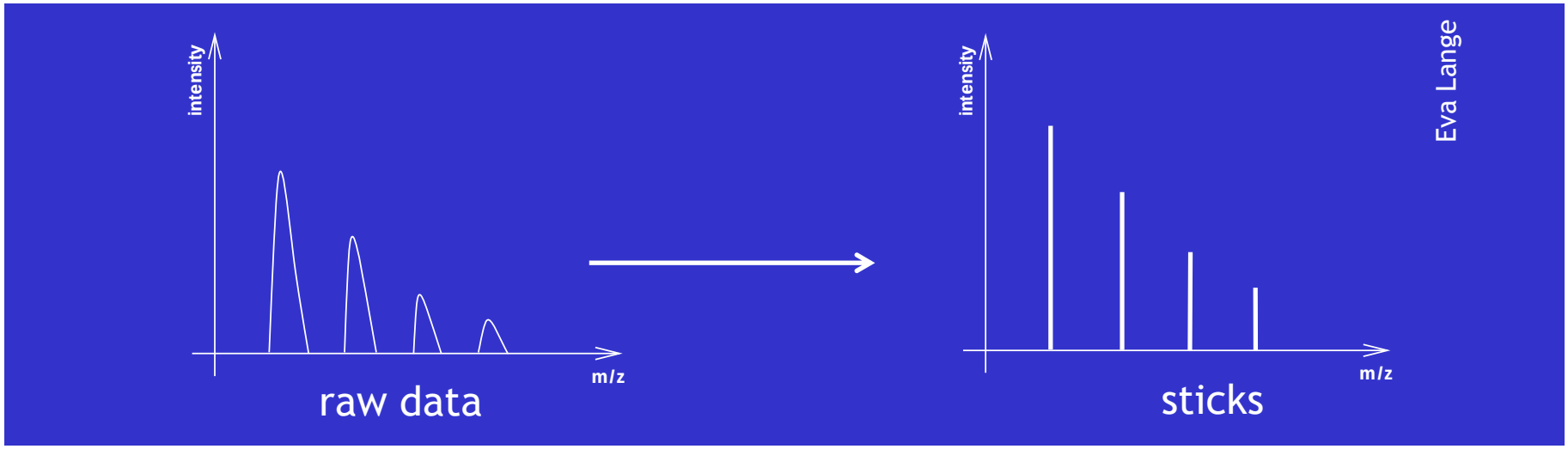


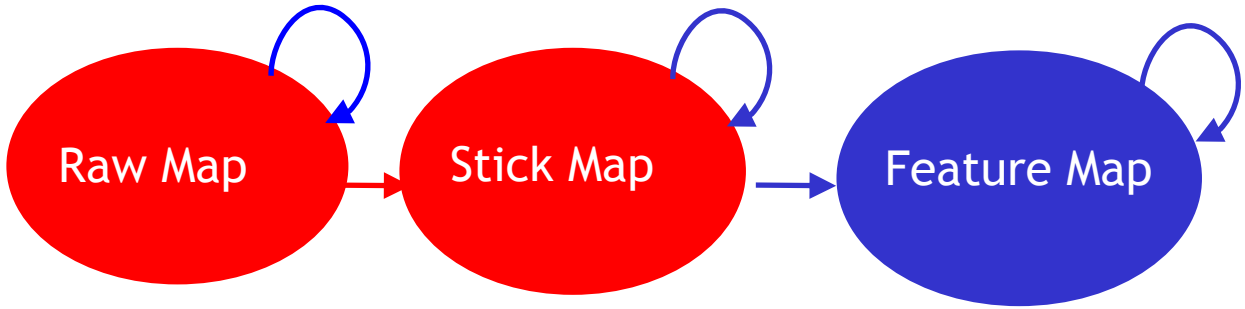
Issues

- „peak picking“ – Identify peaks
- „stick conversion“ – Integrate peaks to sticks

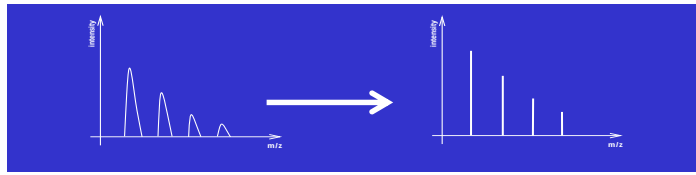
⇒ **Reduce amount of data** by factor of 10 – 100

Stick conversion aka Peak picking



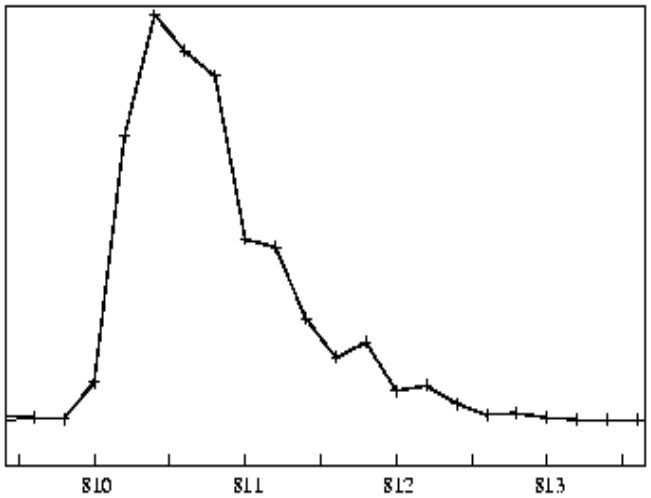


Peak picking in m/z dimension



Main objectives of a peak picking algorithm:

- precise peak positions
- run in real time

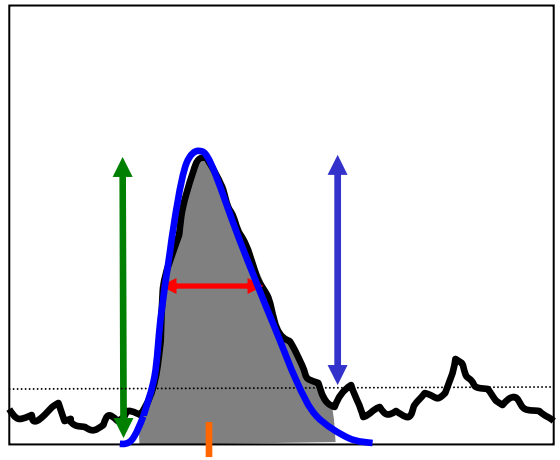


Main difficulties of a peak picking algorithm:

- considerable asymmetry of peaks
- convolution of isotopic peaks

Three step algorithm

1. Peak detection
2. Extract important peak parameter:
   - centroid
   - area
   - height
   - fwhm
   - asymmetric peak shape
   - signal/noise ratio

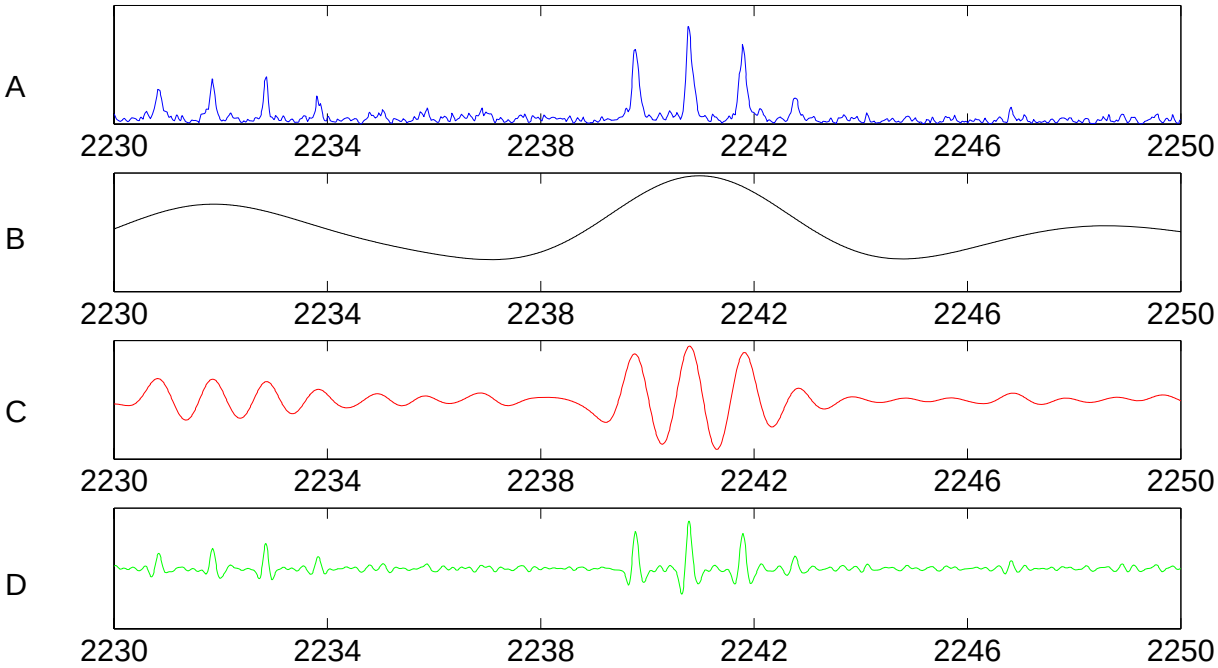


3. Optimize the peak parameter (optional)

1. Peak detection

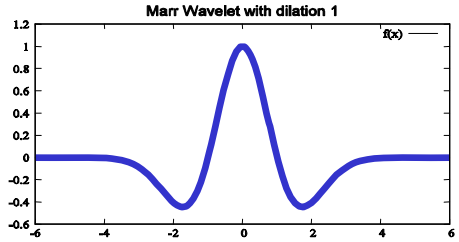
Idea: Split the signal into different frequency ranges (length scales)

Method: Continuous Wavelet Transformation (CWT)

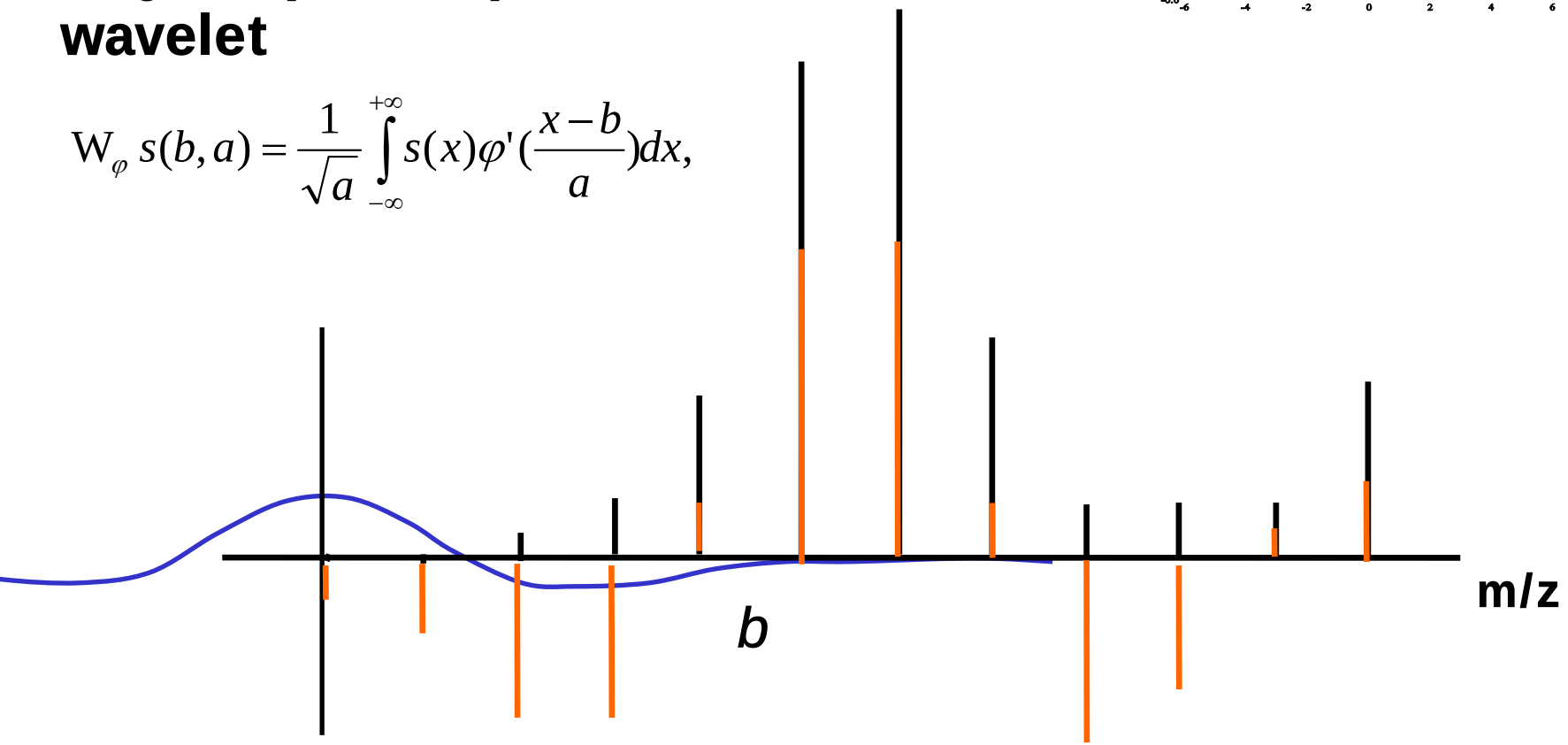


Continuous wavelet transformation

CWT is the sum over the signal multiplied by scaled, shifted versions of the wavelet



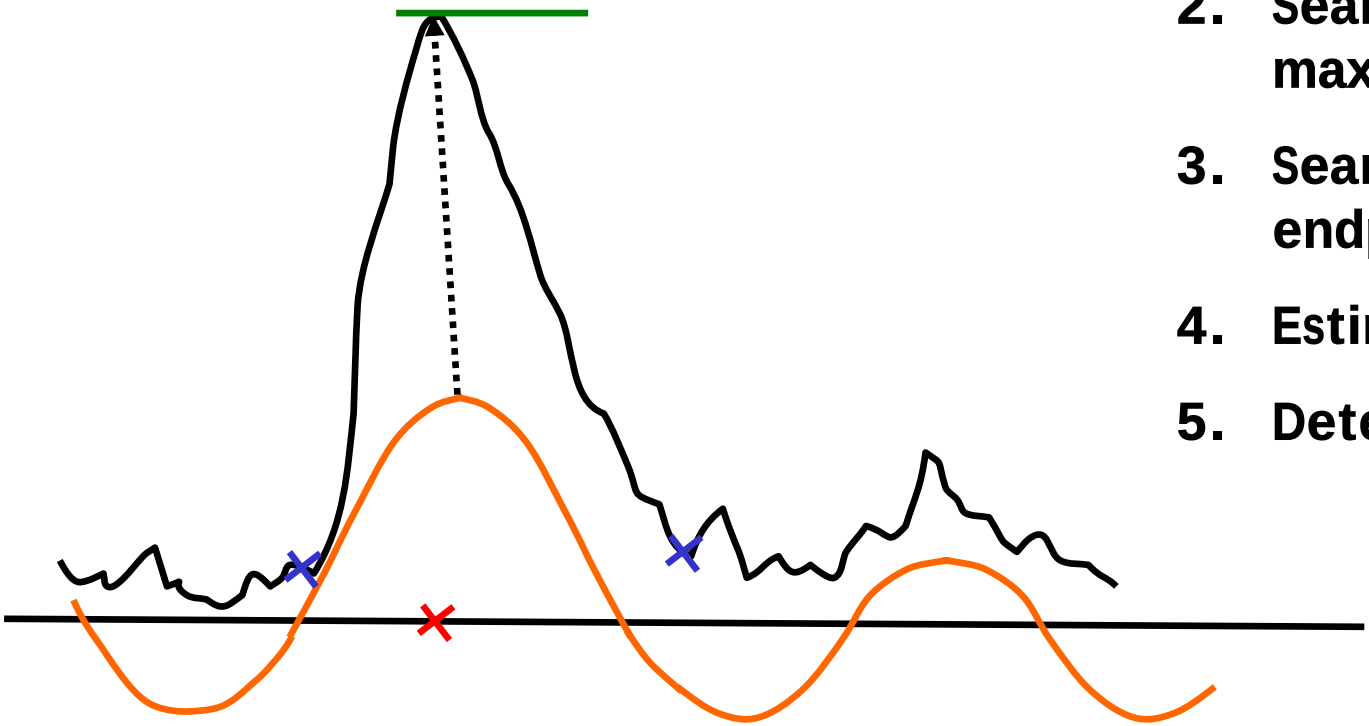
$$W_\varphi s(b,a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} s(x)\varphi'\left(\frac{x-b}{a}\right)dx,$$



1. Peak detection

Freie Universität Berlin

Workflow



1. Compute the wavelet transform
2. Search for a peak‘s maximum
3. Search for peak-endpoints
4. Estimate the centroid
5. Determine the height

1. Compute the peak positions

Compute the CWT:

- use the marr wavelet
- predefined scale a (should correspond to the typical peak width)

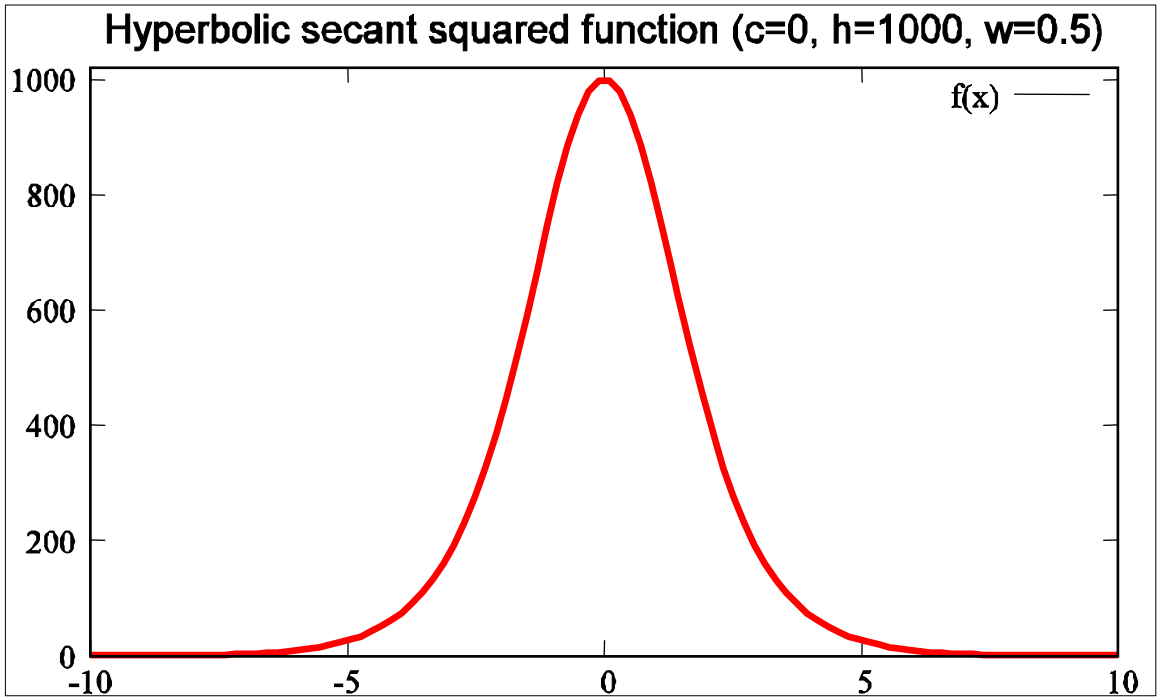
✓maximum position in the CWT is a good first estimate of the maximum in the data even for asymmetric peaks

✓convolution can be computed very efficiently with pre-tabulated values for the wavelet

2. Extract important peak parameter

Freie Universität Berlin

Fit functions similar to real raw mass peaks

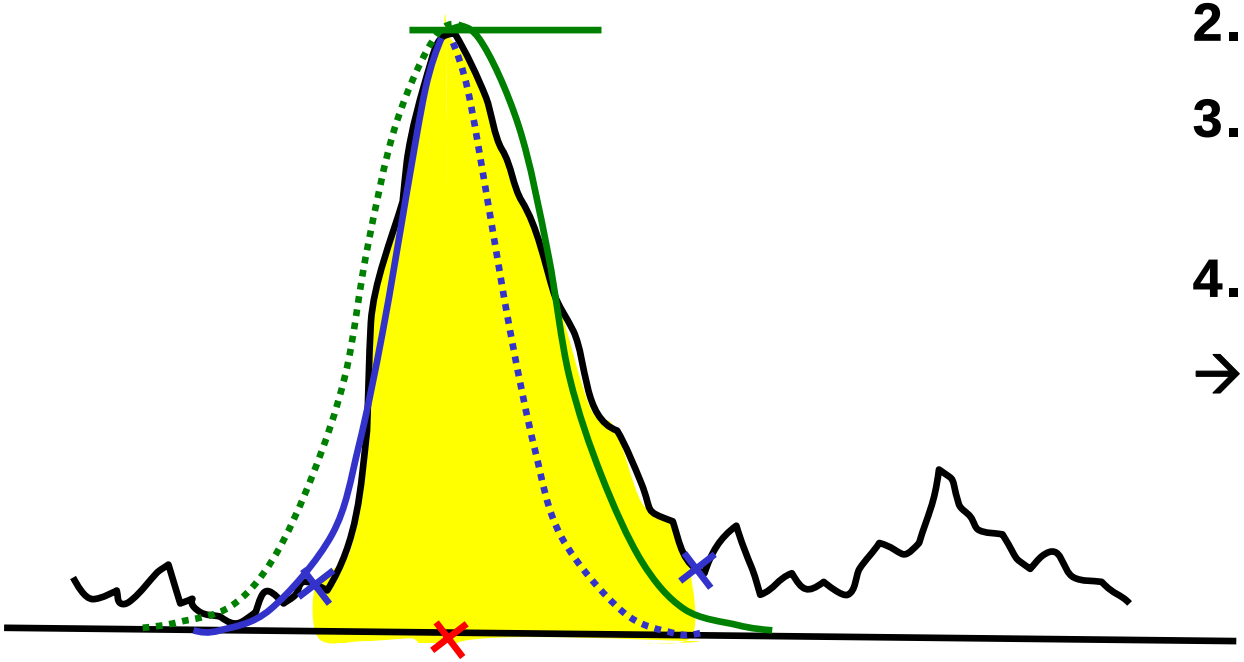


$$Sech^2{}_{h,w,x_0}(x) = \frac{h}{\cosh(w(x - x_0))^2}$$

2. Extract important peak parameter

Freie Universität Berlin

Workflow



1. Estimate the peak‘s left area
2. Fit a $sech^2$ to the left
3. Estimate the peak‘s right area
4. Fit a $sech^2$ to the right

→ Asymmetric peak shape

2. Extract important peak parameter

Fitting asymmetric lorentzian and $sech^2$ functions to a raw peak using the peak‘s :

- maximum position
- height
- area

✓introduces a smoothing effect

✓compute the peak‘s width from its area in constant time

✓yield very good approximations to the original peak shape

Peak picking algorithm

Freie Universität Berlin

for all mass spectra *s* **in** *experiment* **do**
    box_list **=** splitSpectrum(*s*)
    for all boxes *b* **in** *box_list* **do**
        peak_list = []
        w_b = continuousWaveletTransform(*b*)
        while getNextMaximumPosition(w_b,*b*,x_0) **do**
            (x_l,x_r) = searchForPeakEndPoints(*b*,x_0)
            c = estimateCentroid(x_l,x_r)
            h = intensity(x_0)
            (A_l,A_r) = integrateAreas(x_l,x_r)
            f = fitPeakShape(A_l,A_r,x_0,h)
            push(*f*,*peak_list*)
            removeRawData(x_l,x_r,*b*)
            w_b = continuousWaveletTransform(*b*)
        end while
        optimizePeakParameter(*peak_list*,*b*)
    end for
end for

Peak picking algorithm

Freie Universität Berlin

```
for all mass spectra s in experiment do
    box_list = splitSpectrum(s)
    for all boxes b in box_list do
        peak_list = []
        w_b = continuousWaveletTransform(b)
        while getNextMaximumPosition(w_b,b,x_0) do
            (x_l,x_r)  = searchForPeakEndPoints(b,x_0)
                c      = estimateCentroid(x_l,x_r)
                h      = intensity(x_0)
            (A_l,A_r)  = integrateAreas(x_l,x_r)
                f      = fitPeakShape(A_l,A_r,x_0,h)
            push(f,peak_list)
            removeRawData(x_l,x_r,b)
            w_b = continuousWaveletTransform(b)
        end while
        optimizePeakParameter(peak_list,b)
    end for
end for
```

Peak picking algorithm

Freie Universität Berlin

```
for all mass spectra s in experiment do
    box_list = splitSpectrum(s)
    for all boxes b in box_list do
        peak_list = []
        w_b = continuousWaveletTransform(b)
        while getNextMaximumPosition(w_b,b,x_0) do
            (x_l,x_r)  = searchForPeakEndPoints(b,x_0)
                c      = estimateCentroid(x_l,x_r)
                h      = intensity(x_0)
            (A_l,A_r)  = integrateAreas(x_l,x_r)
                f      = fitPeakShape(A_l,A_r,x_0,h)
            push(f,peak_list)
            removeRawData(x_l,x_r,b)
            w_b = continuousWaveletTransform(b)
        end while
        optimizePeakParameter(peak_list,b)
    end for
end for
```

Peak picking algorithm

Freie Universität Berlin

for all mass spectra *s* **in** *experiment* **do**
  box_list **=** splitSpectrum(*s*)
  for all boxes *b* **in** *box_list* **do**
    peak_list = []
    w_b = continuousWaveletTransform(*b*)
    while getNextMaximumPosition(w_b,*b*,x_0) **do**
      (x_l,x_r) = searchForPeakEndPoints(*b*,x_0)
      c = estimateCentroid(x_l,x_r)
      h = intensity(x_0)
      (A_l,A_r) = integrateAreas(x_l,x_r)
      f = fitPeakShape(A_l,A_r,x_0,h)
      push(*f*,*peak_list*)
      removeRawData(x_l,x_r,*b*)
      w_b = continuousWaveletTransform(*b*)
    end while
    optimizePeakParameter(*peak_list*,*b*)
  end for
end for

Peak picking algorithm

Freie Universität Berlin

for all mass spectra *s* **in** *experiment* **do**
    box_list **=** splitSpectrum(*s*)
    for all boxes *b* **in** *box_list* **do**
        peak_list = []
        w_b = continuousWaveletTransform(*b*)
        while getNextMaximumPosition(w_b,*b*,x_0) **do**
            (x_l,x_r) = searchForPeakEndPoints(*b*,x_0)
            c = estimateCentroid(x_l,x_r)
            h = intensity(x_0)
            (A_l,A_r) = integrateAreas(x_l,x_r)
            f = fitPeakShape(A_l,A_r,x_0,h)
            push(*f*,*peak_list*)
            removeRawData(x_l,x_r,*b*)
            w_b = continuousWaveletTransform(*b*)
        end while
        optimizePeakParameter(*peak_list*,*b*)
    end for
end for

Evaluation of a peak picking method

Freie Universität Berlin

Compute accurate:

- centroid
- area
- height

→ test against a set of known composition

Results are heavily affected by:

- quality of experimental data
- calibration method

Evaluation of our peak picking method

Freie Universität Berlin

Data set I: peptide mix (peptide standards mix #P2693 from Sigma Aldrich) of nine known peptides.

Measuring method:

HPLC-MS using a quadrupole ion trap mass spectrometer equipped with an electrospray ion source in full scan mode (m/z 500-1500)

→ low resolution ($\Delta m=0.2$) and several overlapping isotope patterns

Evaluation of our peak picking method

Freie Universität Berlin

Evaluation:

- compare with vendor supplied Bruker Data-Analysis software (3.2) on the same spectra
- no sophisticated calibration, we only allow a constant mass offset to keep the number of fit parameters as small as possible

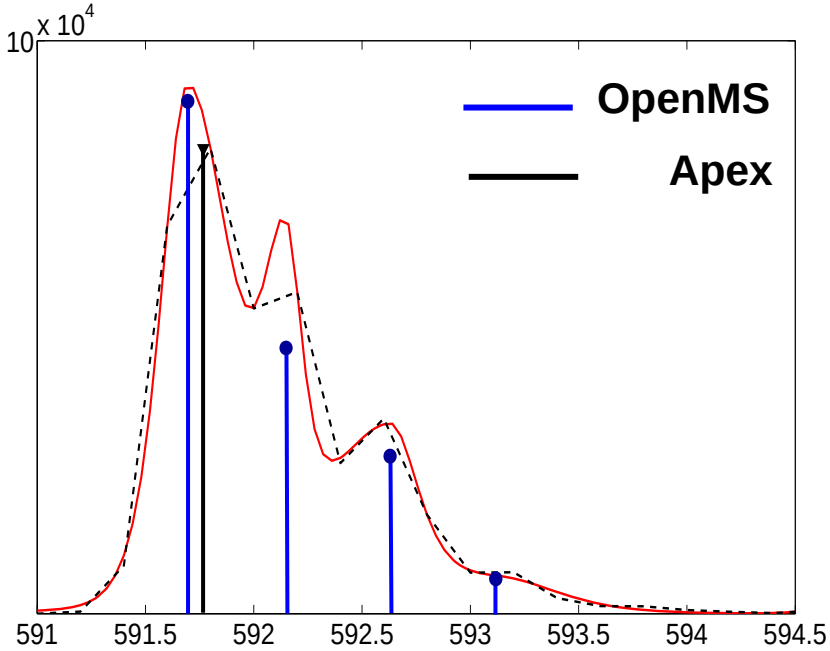
Performance criteria:

1. number of the discovered and separated isotopic patterns (given by at least three consecutive peaks) of each peptide in an expected retention time interval
2. average relative error of the monoisotopic mass

Results: LC-ESI-ion trap data
(Full scan mode (m/z 500-1500))

Freie Universität Berlin

Data set I (Peptide mix of nine known peptides)



Relative error of centroids:

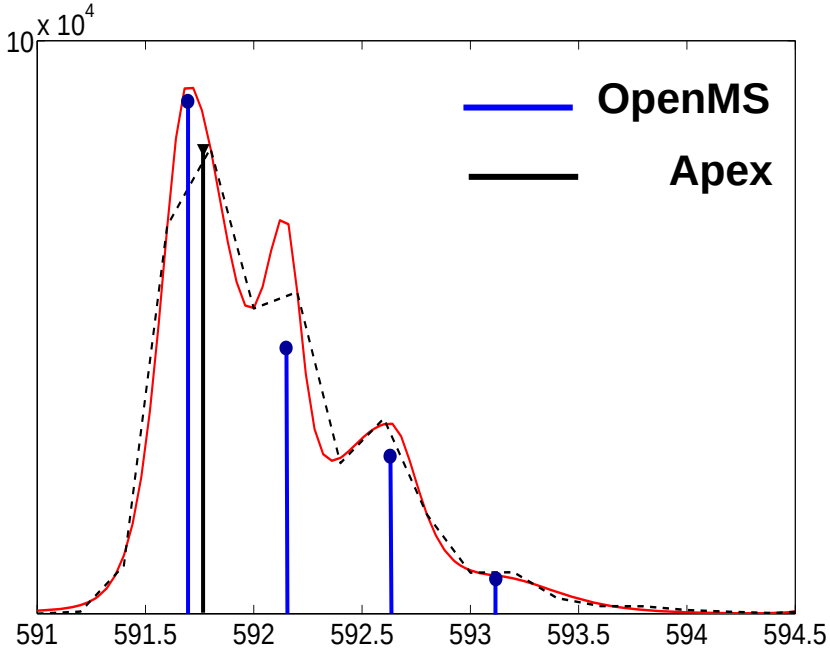
21ppm, 1.2ppm, 35ppm, 16ppm

	$mass_{theo}$	z	#occ OpenMS	#occ Apex
A	556.2693	1	22	19
B	573.3071	1	29	29
C	574.2257	1	19	15
D	1007.4365	1	8	5
E	1084.4379	1	3	2
F	1061.5614	2	3	2
G	1183.5730	2	7	0
H	1349.7360	2	8	1
I	1620.8151	2	13	0

Results: LC-ESI-ion trap data
(Full scan mode (m/z 500-1500))

Freie Universität Berlin

Data set I (Peptide mix of nine known peptides)



Relative error of centroids:

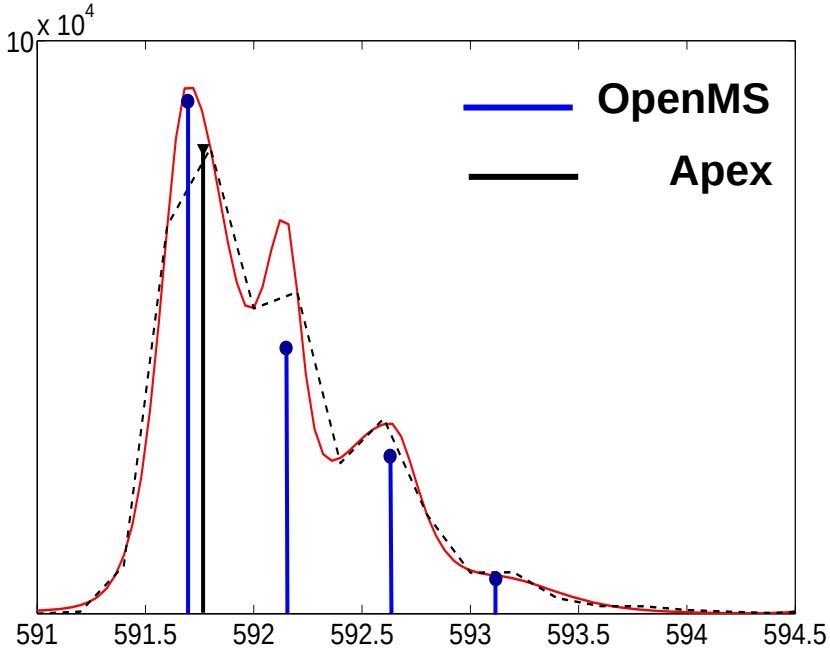
21ppm, 1.2ppm, 35ppm, 16ppm

	$mass_{theo}$	z	#occ OpenMS	#occ Apex
A	556.2693	1	22	19
B	573.3071	1	29	29
C	574.2257	1	19	15
D	1007.4365	1	8	5
E	1084.4379	1	3	2
F	1061.5614	2	3	2
G	1183.5730	2	7	0
H	1349.7360	2	8	1
I	1620.8151	2	13	0

Results: LC-ESI-ion trap data
(Full scan mode (m/z 500-1500))

Freie Universität Berlin

Data set I (Peptide mix of nine known peptides)



Relative error of centroids:

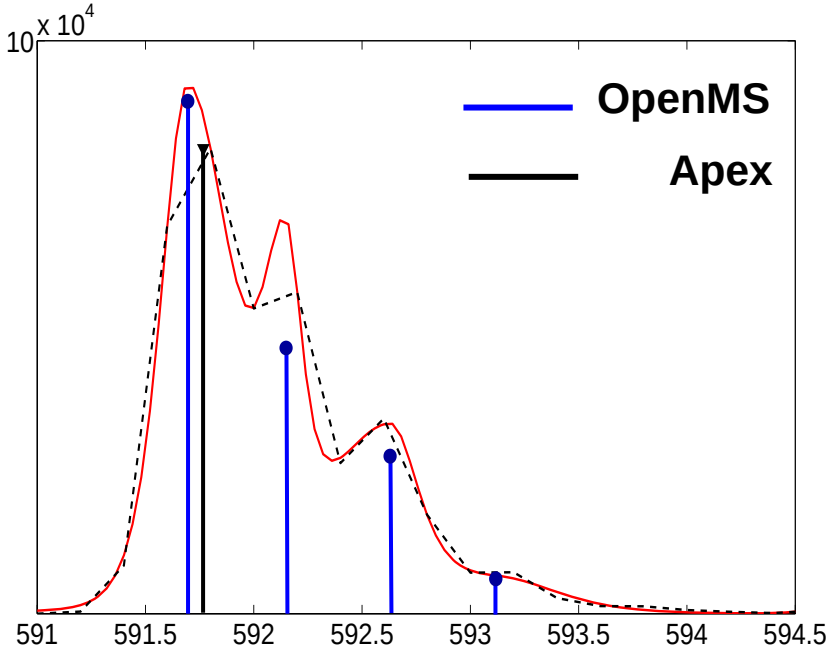
21ppm, 1.2ppm, 35ppm, 16ppm

	$mass_{theo}$	z	#occ OpenMS	#occ Apex
A	556.2693	1	22	19
B	573.3071	1	29	29
C	574.2257	1	19	15
D	1007.4365	1	8	5
E	1084.4379	1	3	2
F	1061.5614	2	3	2
G	1183.5730	2	7	0
H	1349.7360	2	8	1
I	1620.8151	2	13	0

Results: LC-ESI-ion trap data
(Full scan mode (m/z 500-1500))

Freie Universität Berlin

Data set I (Peptide mix of nine known peptides)



Relative error of centroids:

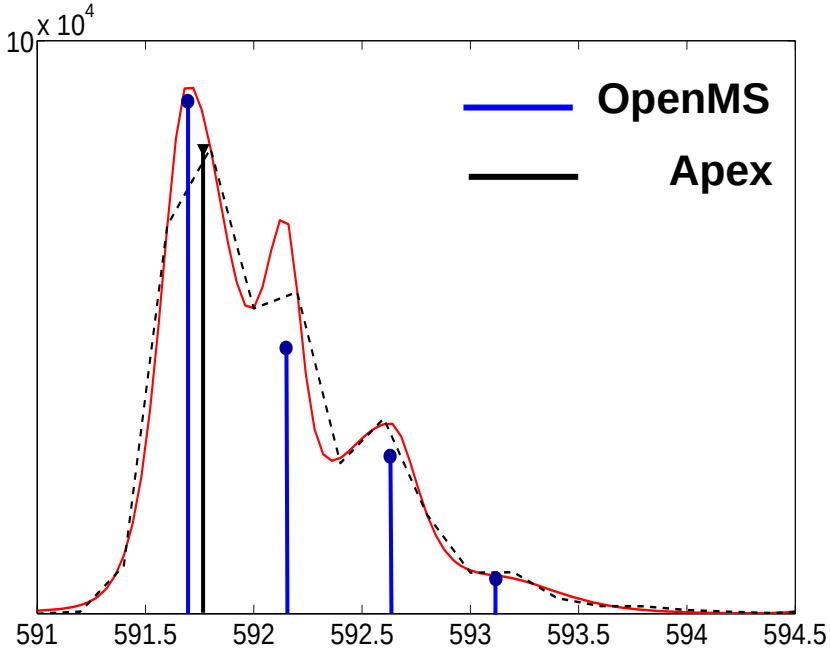
21ppm, 1.2ppm, 35ppm, 16ppm

	$mass_{theo}$	z	rel. err. [ppm] OpenMS	Apex
A	556.2693	1	31	39
B	573.3071	1	16	16
C	574.2257	1	44	21
D	1007.4365	1	25	94
E	1084.4379	1	18	12
F	1061.5614	2	56	64
G	1183.5730	2	15	-
H	1349.7360	2	28	13
I	1620.8151	2	37	-

Results: LC-ESI-ion trap data
(Full scan mode (m/z 500-1500))

Freie Universität Berlin

Data set I (Peptide mix of nine known peptides)



Relative error of centroids:

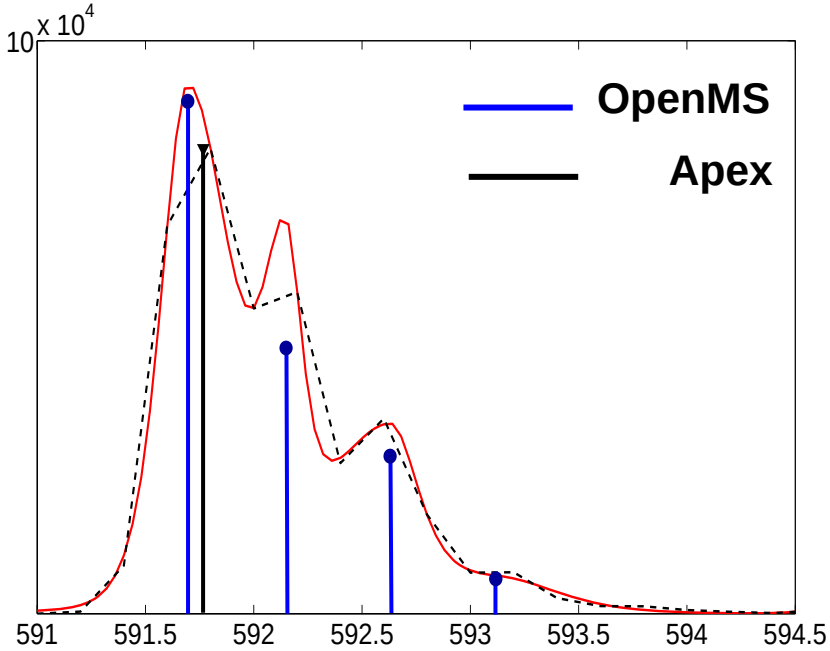
21ppm, 1.2ppm, 35ppm, 16ppm

	$mass_{theo}$	z	rel. err. [ppm] OpenMS	Apex
A	556.2693	1	31	39
B	573.3071	1	16	16
C	574.2257	1	44	21
D	1007.4365	1	25	94
E	1084.4379	1	18	12
F	1061.5614	2	56	64
G	1183.5730	2	15	-
H	1349.7360	2	28	13
I	1620.8151	2	37	-

Results: LC-ESI-ion trap data
(Full scan mode (m/z 500-1500))

Freie Universität Berlin

Data set I (Peptide mix of nine known peptides)



Relative error of centroids:

21ppm, 1.2ppm, 35ppm, 16ppm

	$mass_{theo}$	z	rel. err. [ppm] OpenMS	Apex
A	556.2693	1	31	39
B	573.3071	1	16	16
C	574.2257	1	44	21
D	1007.4365	1	25	94
E	1084.4379	1	18	12
F	1061.5614	2	56	64
G	1183.5730	2	15	-
H	1349.7360	2	28	13
I	1620.8151	2	37	-

Sanity check of our peak picking method

Freie Universität Berlin

Data set II: tryptic digest of bovine serum albumin (BSA from Sigma Aldrich).

Measuring method:

MALDI-TOF using a Ultraflex II Lift mass spectrometer operated in the reflectron mode and using Panorama delayed ion extraction (m/z 500-5000)

Sanity check of our peak picking method

Freie Universität Berlin

Evaluation:

- compare with vendor supplied Bruker FlexAnalysis software on the same spectra
- no sophisticated calibration

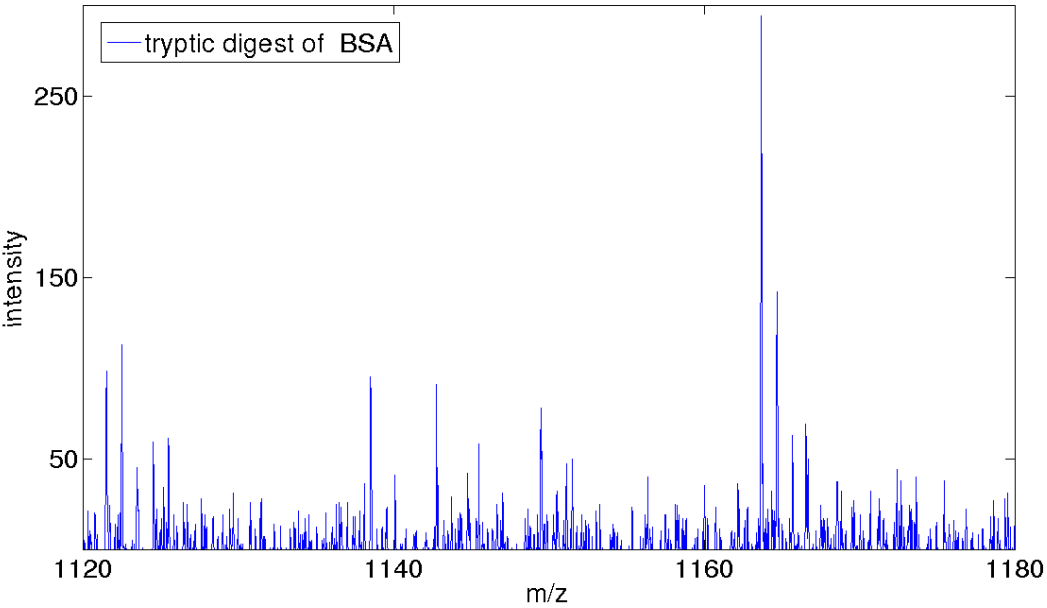
Performance criterion:

1. compute the sequence coverage using the determined peak list as a MASCOT peptide mass fingerprinting query

Results: MALDI-TOF data

Freie Universität Berlin

Data set II (tryptic digest of bovine serum)



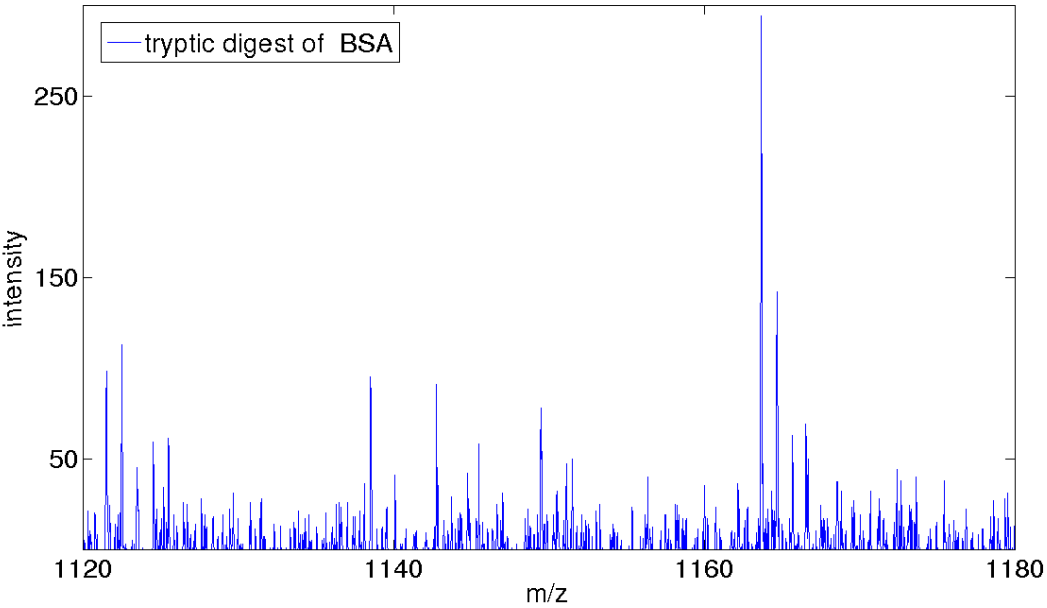
Sequence coverage	
OpenMS	Bruker
52% - 67%	44%

rel. err. [ppm]	
OpenMS	Bruker
80 - 93	95

Results: MALDI-TOF data

Freie Universität Berlin

Data set II (tryptic digest of bovine serum)



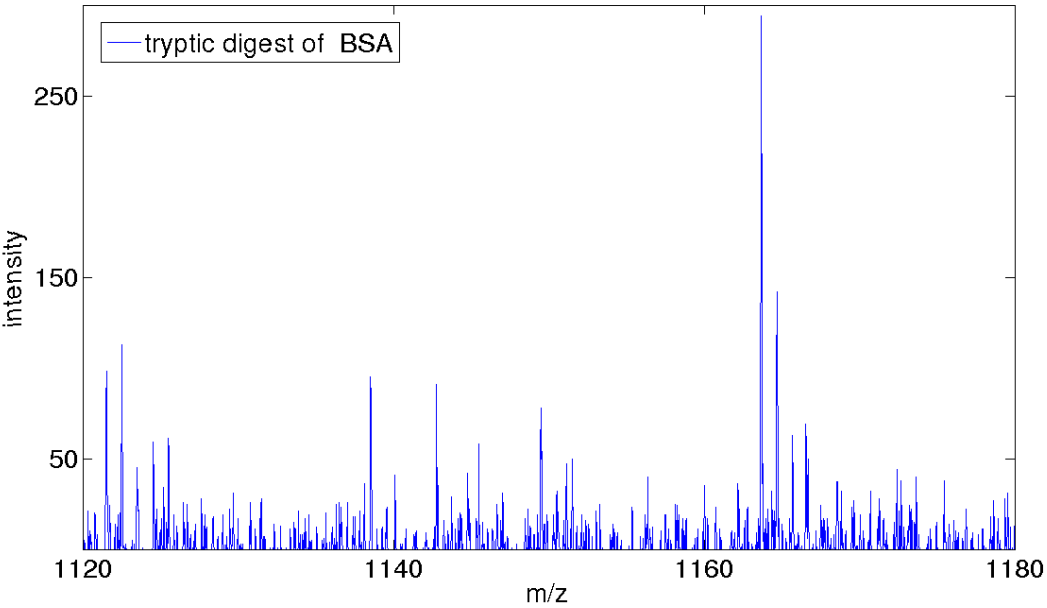
Sequence coverage	
OpenMS	Bruker
52% - 67%	44%

rel. err. [ppm]	
OpenMS	Bruker
80 - 93	95

Results: MALDI-TOF data

Freie Universität Berlin

Data set II (tryptic digest of bovine serum)



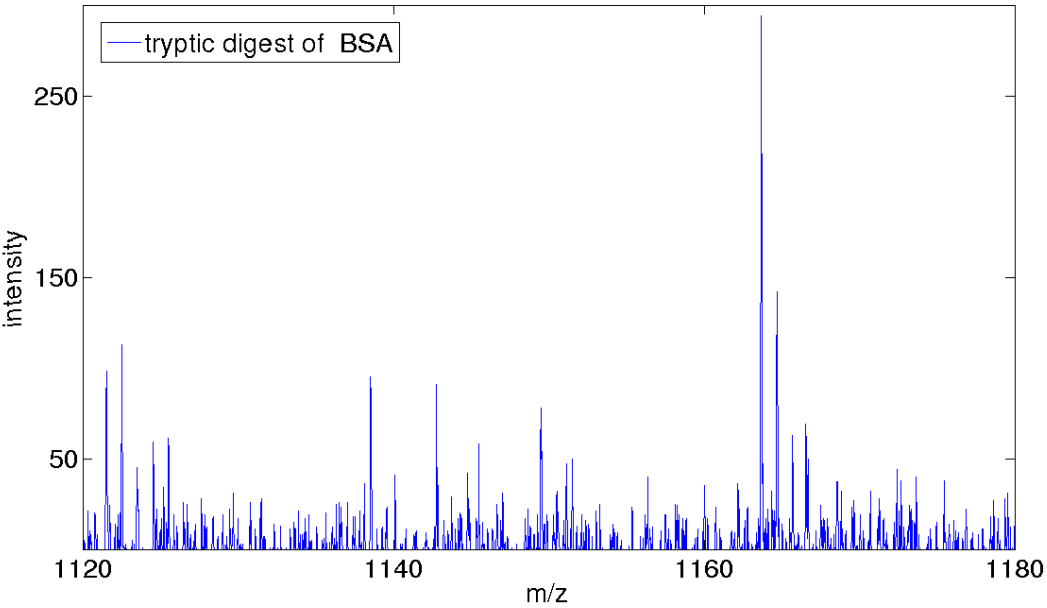
Sequence coverage	
OpenMS	Bruker
52% - 67%	44%

rel. err. [ppm]	
OpenMS	Bruker
80 - 93	95

Results: MALDI-TOF data

Freie Universität Berlin

Data set II (tryptic digest of bovine serum)



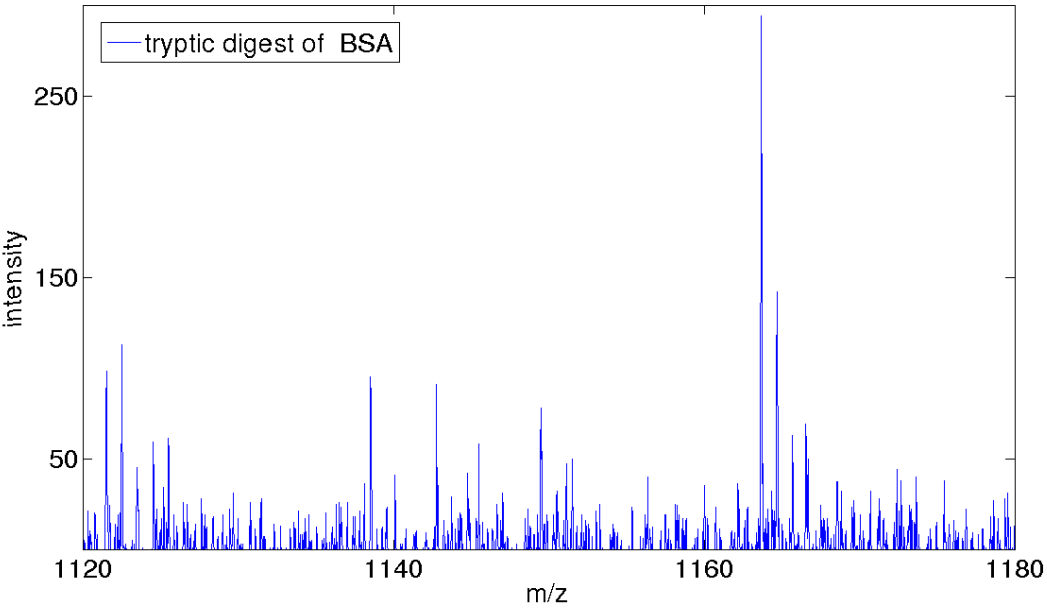
Sequence coverage	
OpenMS	Bruker
52% - 67%	44%

rel. err. [ppm]	
OpenMS	Bruker
80 - 93	95

Results: MALDI-TOF data

Freie Universität Berlin

Data set II (tryptic digest of bovine serum)



Sequence coverage	
OpenMS	Bruker
52% - 67%	44%

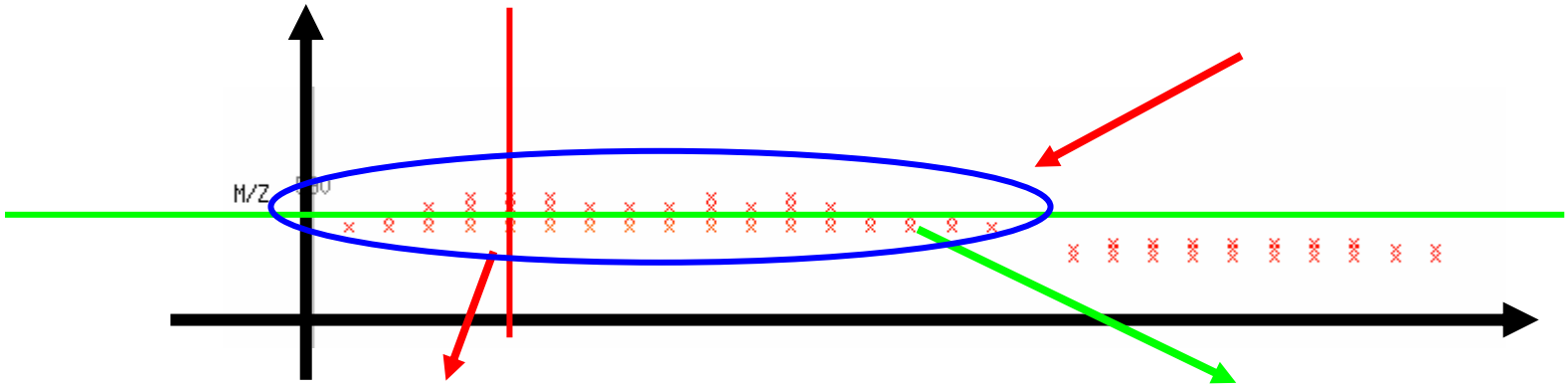
rel. err. [ppm]	
OpenMS	Bruker
80 - 93	95

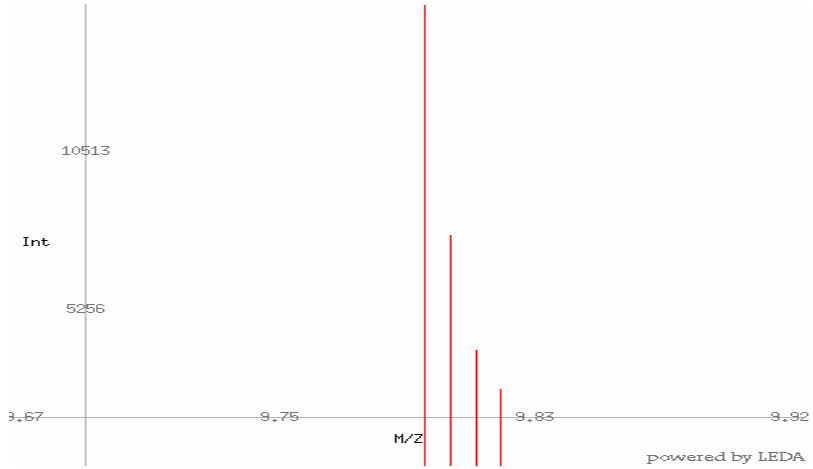
Overview

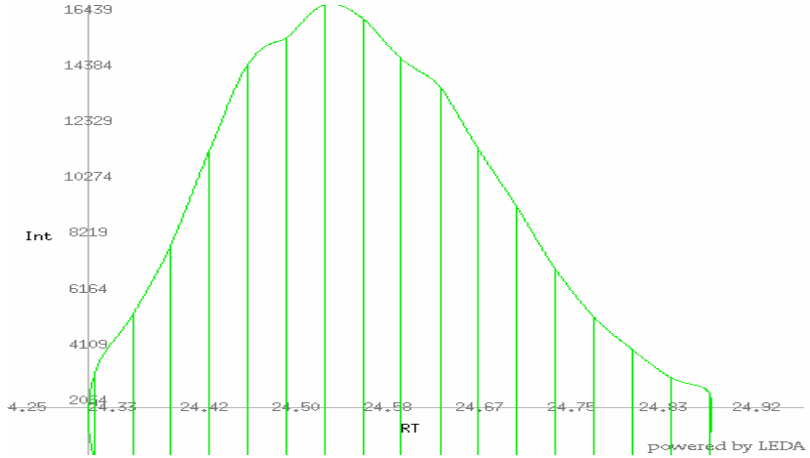
Freie Universität Berlin

- Motivation for OpenMS
- Bioinformatics Issues in quantitative Proteomics
  - Signal Processing
  - Feature Finding
  - Map Mapping
  - Differential Quantitation
  - Identification, Clustering
  - Software Engineering, Databases

Feature Finding



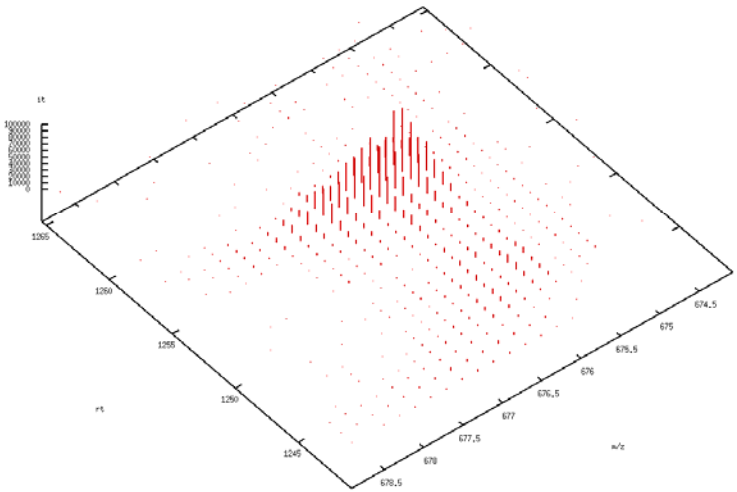


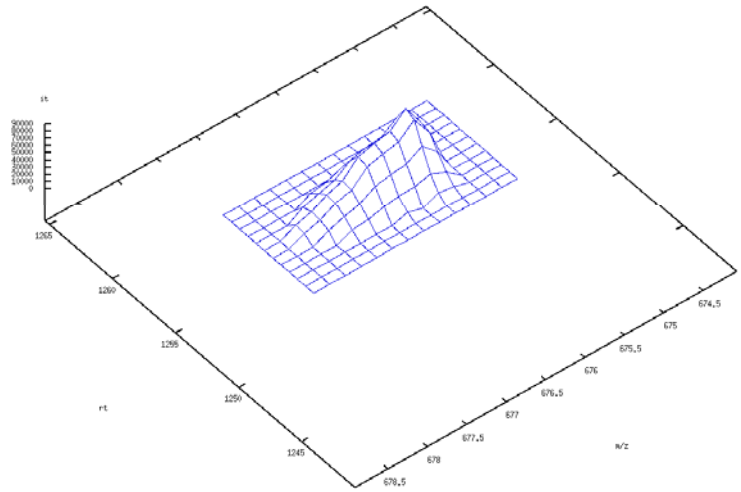


Capture **ALL** peaks belonging to a peptide for **quantitation**!

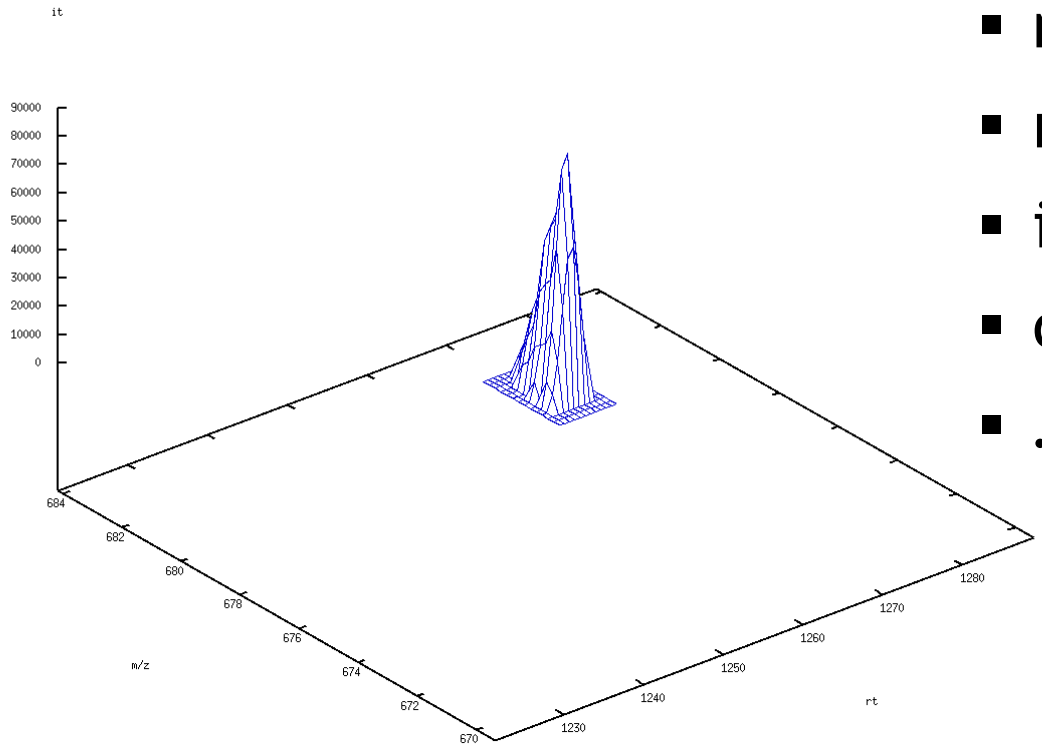
Feature Finding

- A two-dimensional *model* has to be adjusted to the raw data





Feature Finding ...



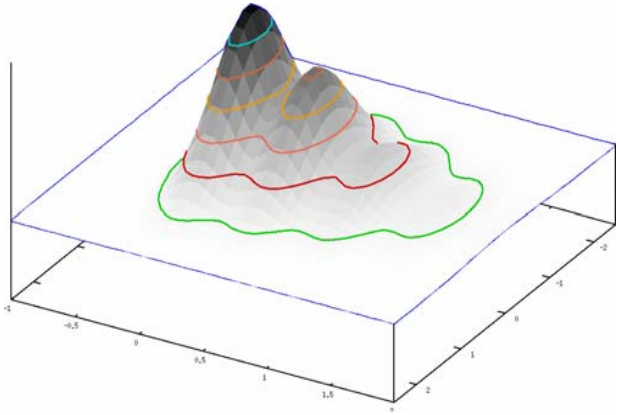
Attributes of a feature:

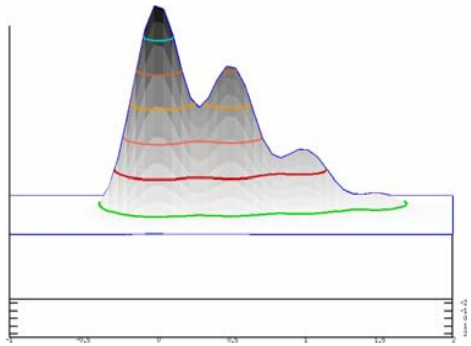
- mass-to-charge ratio
- retention time
- intensity, volume
- quality
- … … …

Feature Finding

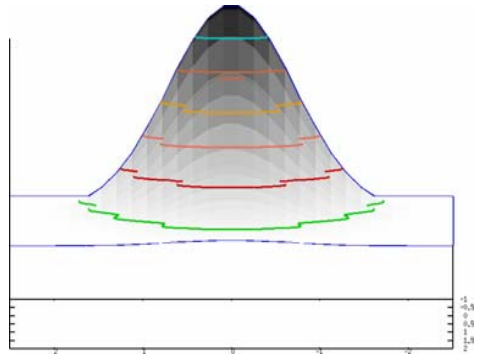
Freie Universität Berlin

feature model = isotope pattern + elution profile





m/z



rt

Isotope Patterns

- Natural isotopes occur with well-known abundances
- Peak intensities determined by binomial convolution
- Depend on molecular formula of peptide

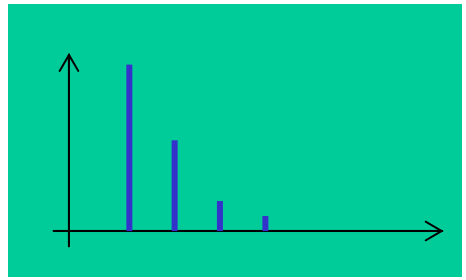
^{12}C	98.90%	^{13}C	1.10%		
^{14}N	99.63%	^{15}N	0.37%		
^{16}O	99.76%	^{17}O	0.04%	^{18}O	0.20%
^{1}H	99.98%	^{2}H	0.02%		

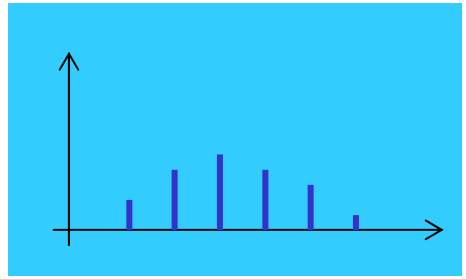
Estimating the Isotope Pattern

Randomly sample peptides from protein database and compute average composition.

Table with average isotope pattern for specified peptide **mass**

m [Da]	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

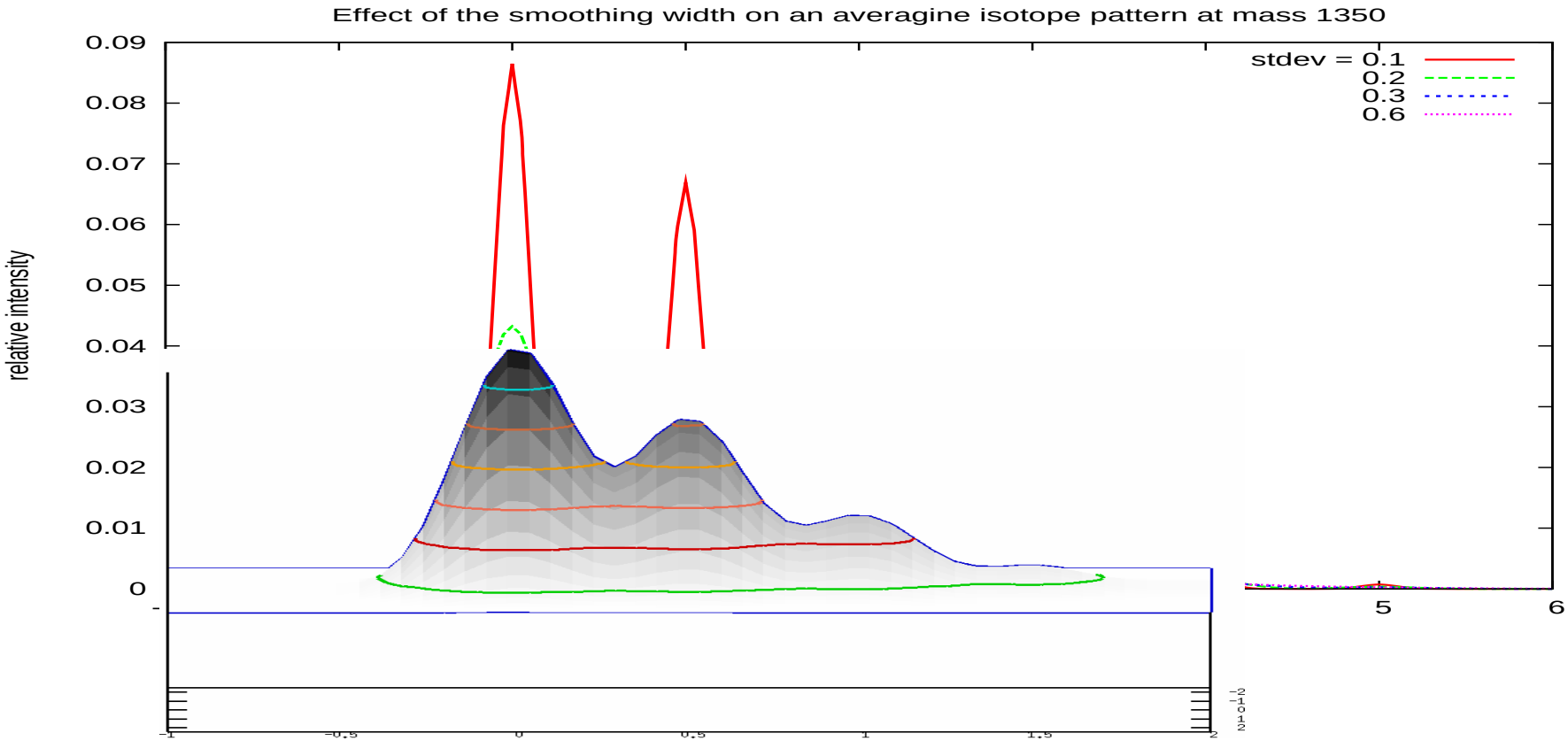




Feature Models: m/z

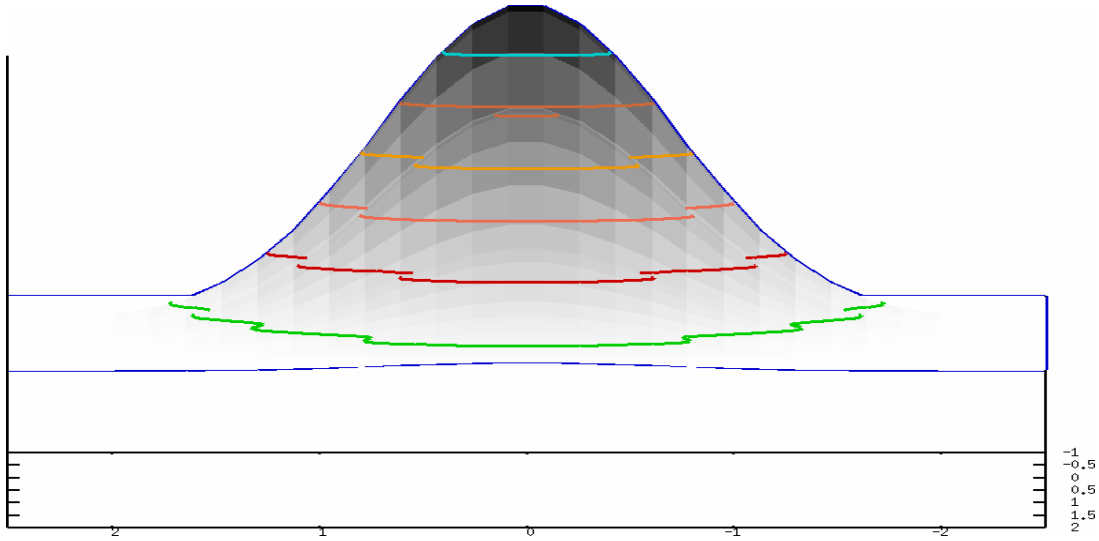
Freie Universität Berlin

- Isotope patterns

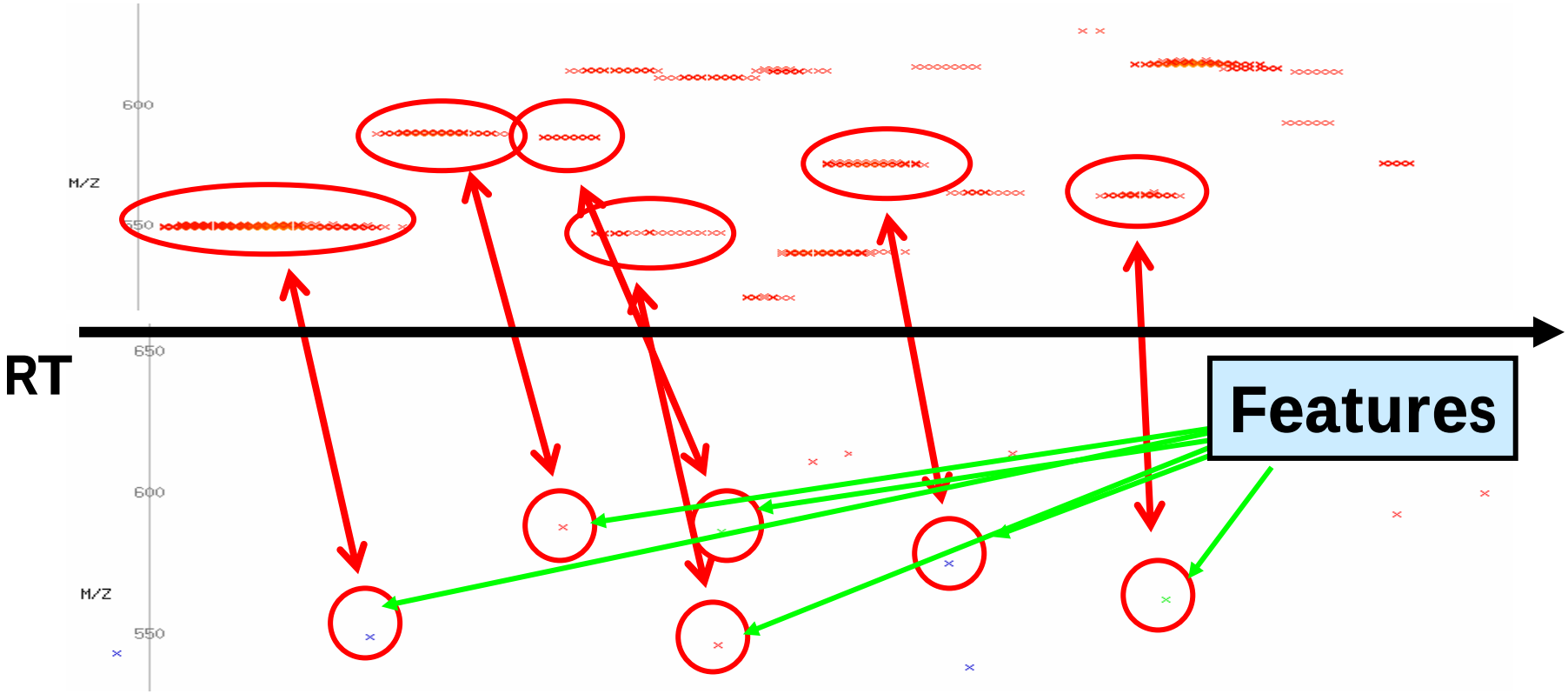


Feature Models: rt

- Elution profiles
  - Currently modeled by a normal distribution, other shapes possible



Peak Aggregation

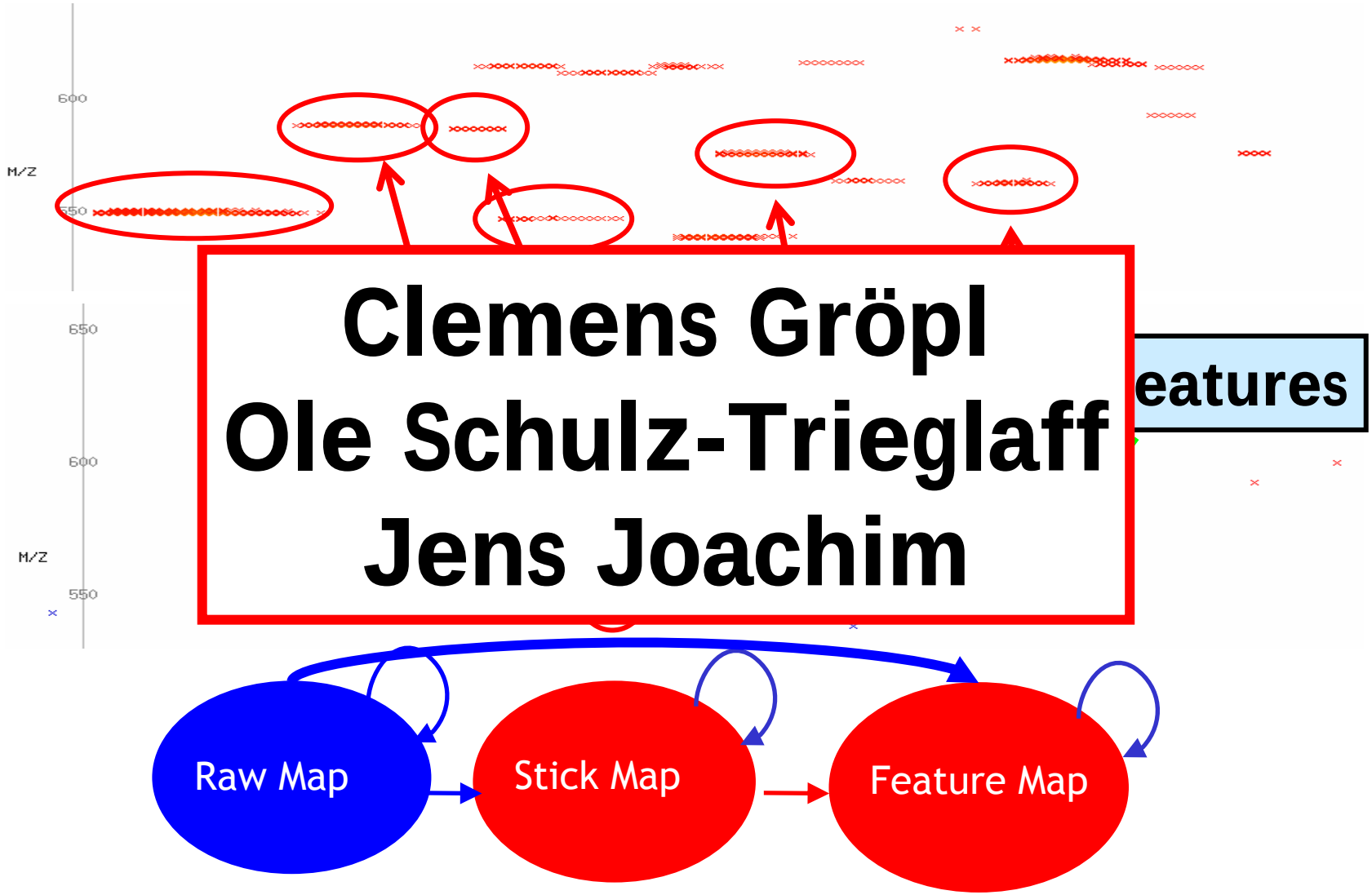


Peak aggregation to **features** yields

- 1,000-fold compression in data volume
- reduction to meaningful entity: a peptide
- accurate quantitation: integration of all related peaks

Peak Aggregation

Freie Universität Berlin



Overview Data Reduction

Freie Universität Berlin

- Noise/baseline filtering
- Peak picking, integration
- Peak aggregation
  - **De-isotoping**: joining peaks from same isotope pattern
  - **De-charging**: joining features from same peptide, different charge states

$\Rightarrow$ **Aggregated Map**

Reduction in Data Volume: up to 10^4

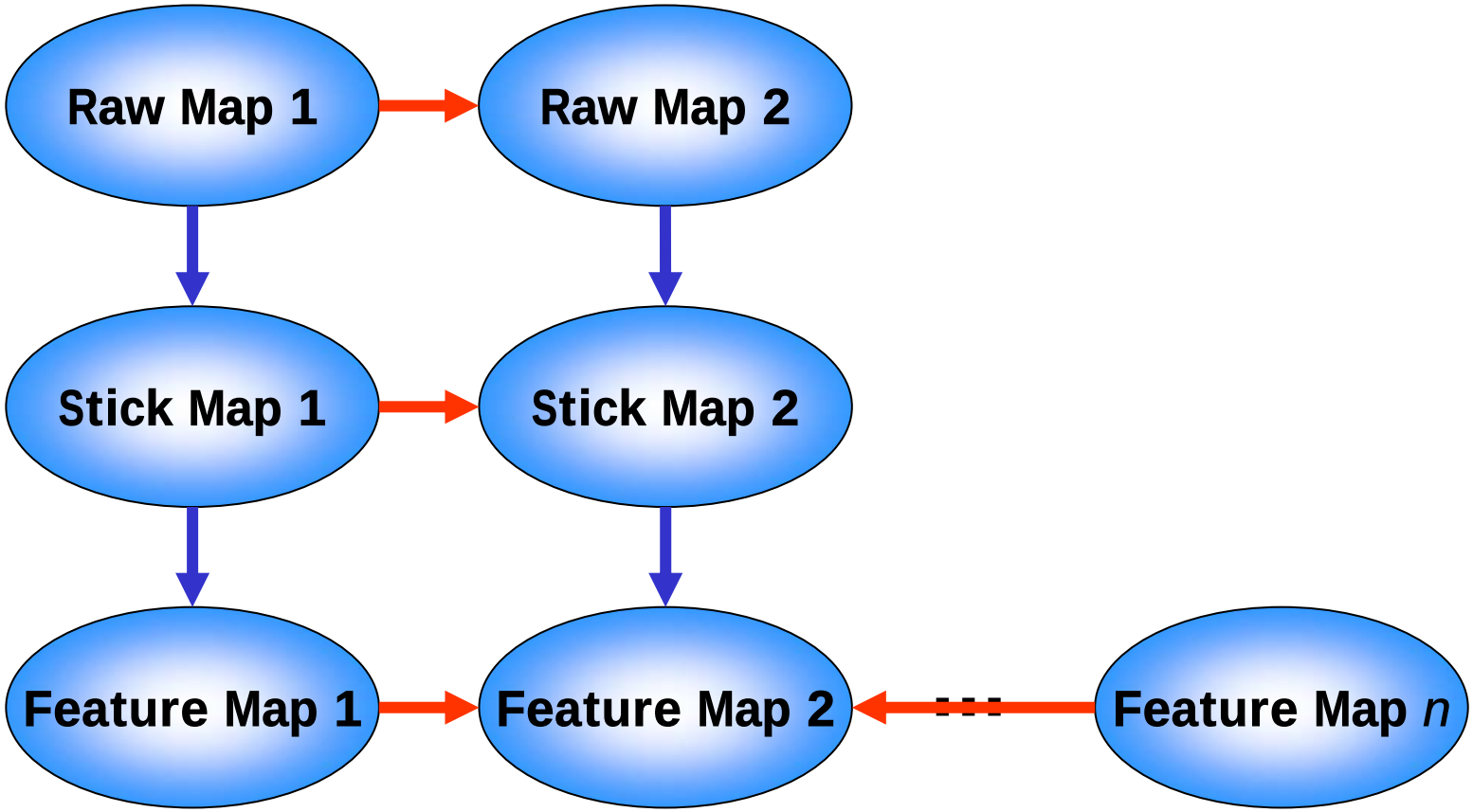
Overview

Freie Universität Berlin

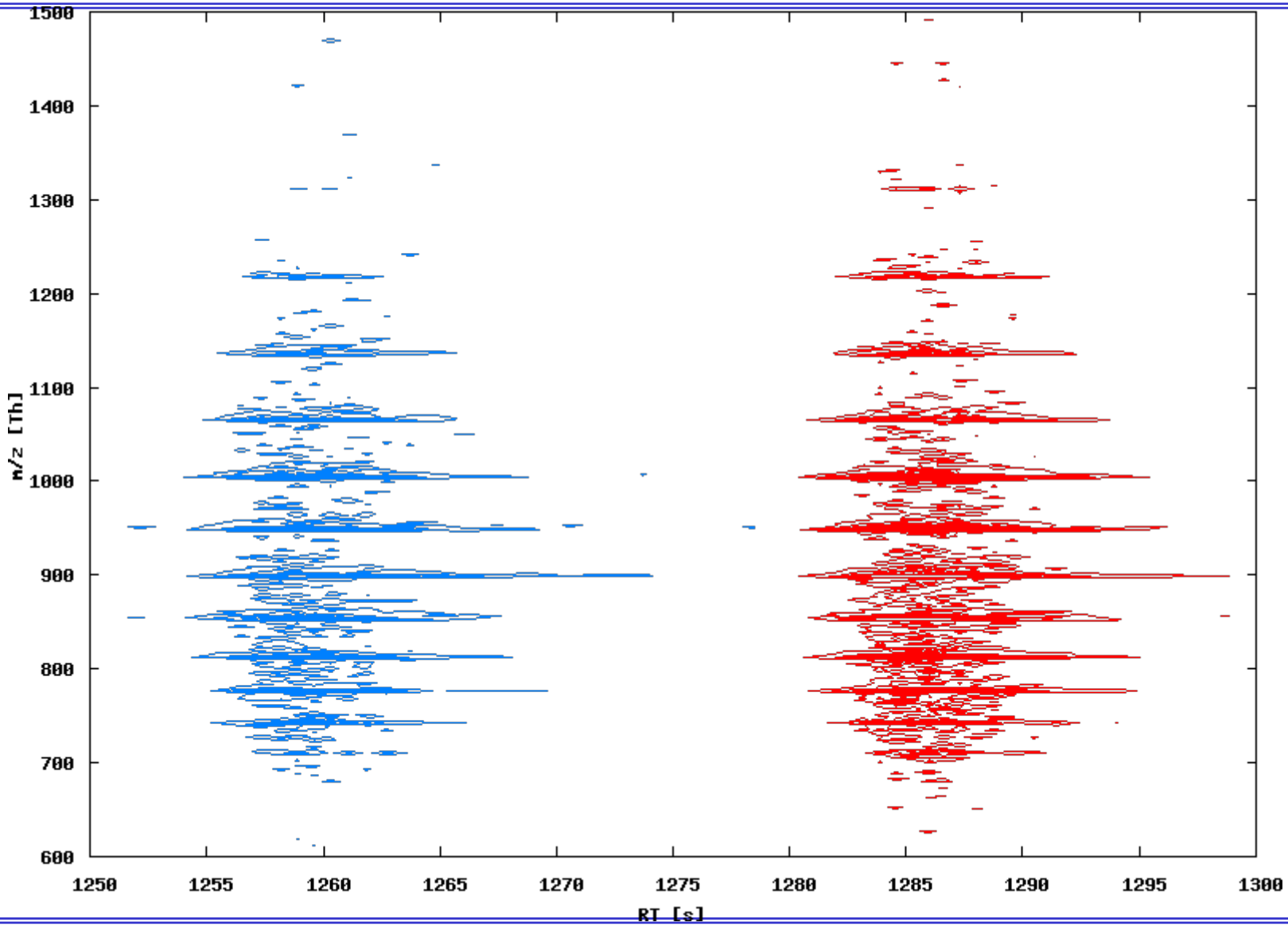
- Motivation for OpenMS
- Bioinformatics Issues in quantitative Proteomics
  - Signal Processing
  - Feature Finding
  - Map Mapping
  - Differential Quantitation
  - Identification, Clustering
  - Software Engineering, Databases
- Subsequent talks give more details

Map Mapping – At All Levels

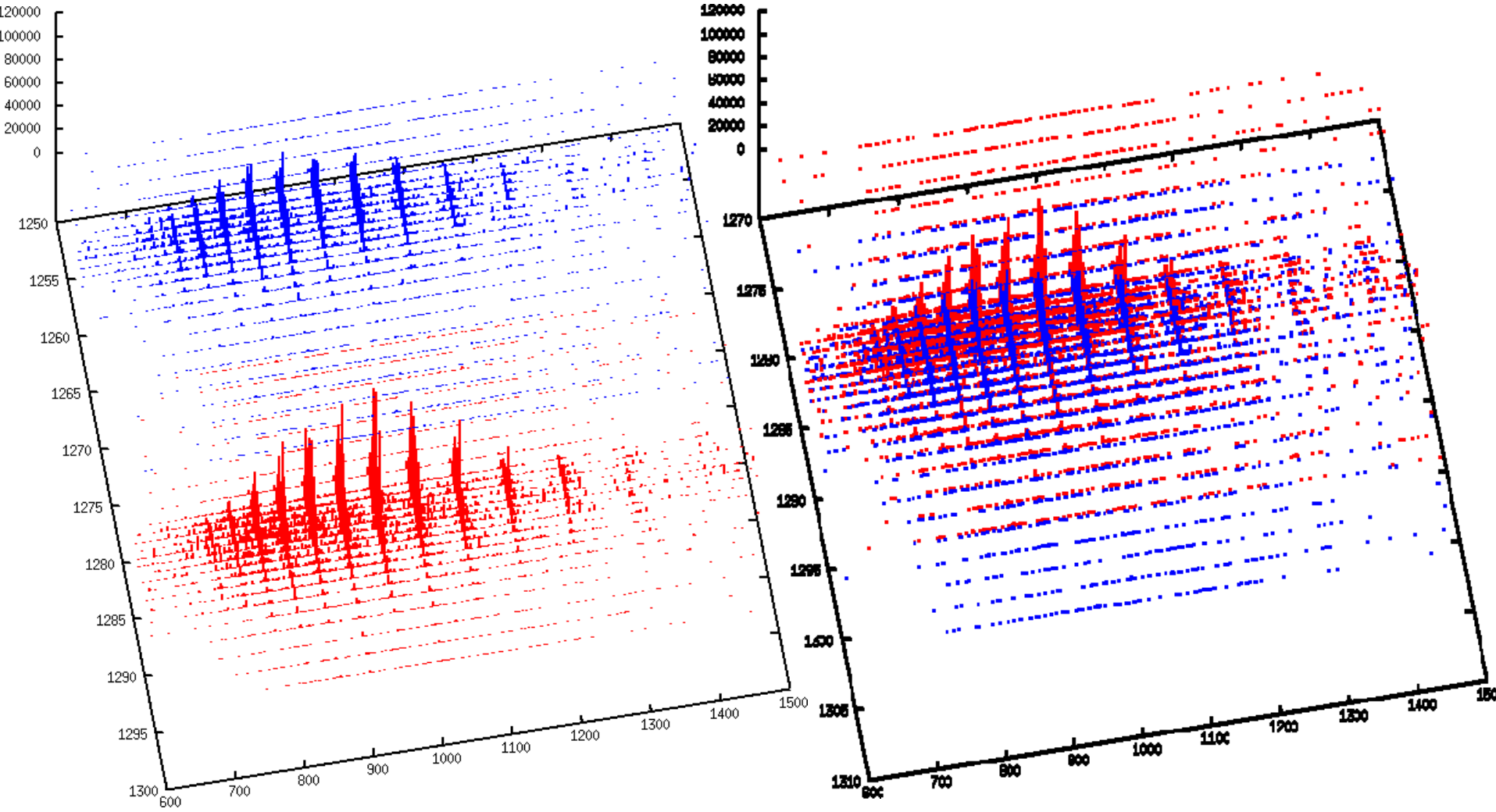
→ Data reduction → Mapping



Mapping of Raw Maps

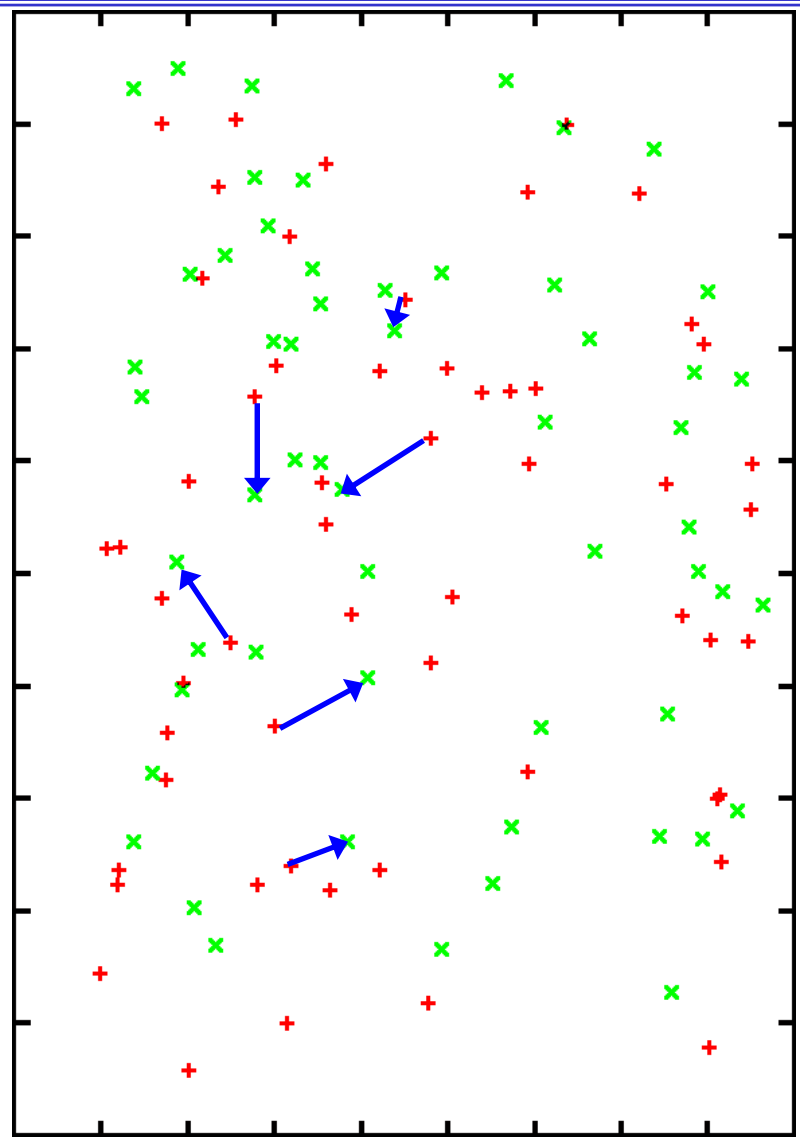


Mapping Discrete Peaks



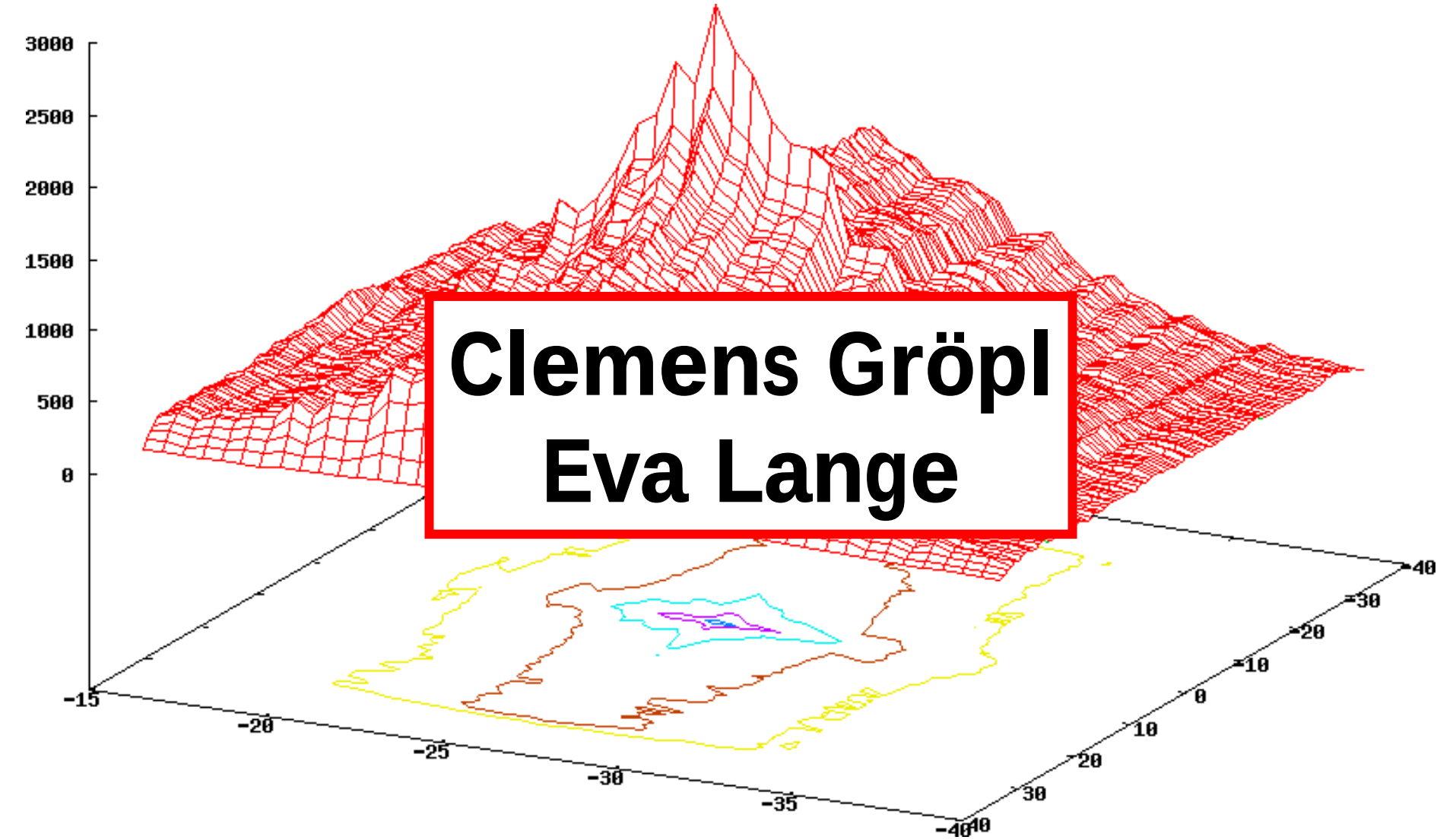
Geometric Hashing

- Consider difference vectors → from features in map 1 ("+") to features in map 2 ("×")
- Restrict difference vector to reasonable differences along RT and m/z dimensions
- Hash these vectors into bins ("2D histogram")
- In sufficiently similar maps, the optimal translation will be the most frequent one



Finding the Offset Vector

Freie Universität Berlin



Clemens Gröpl
Eva Lange

Overview

Freie Universität Berlin

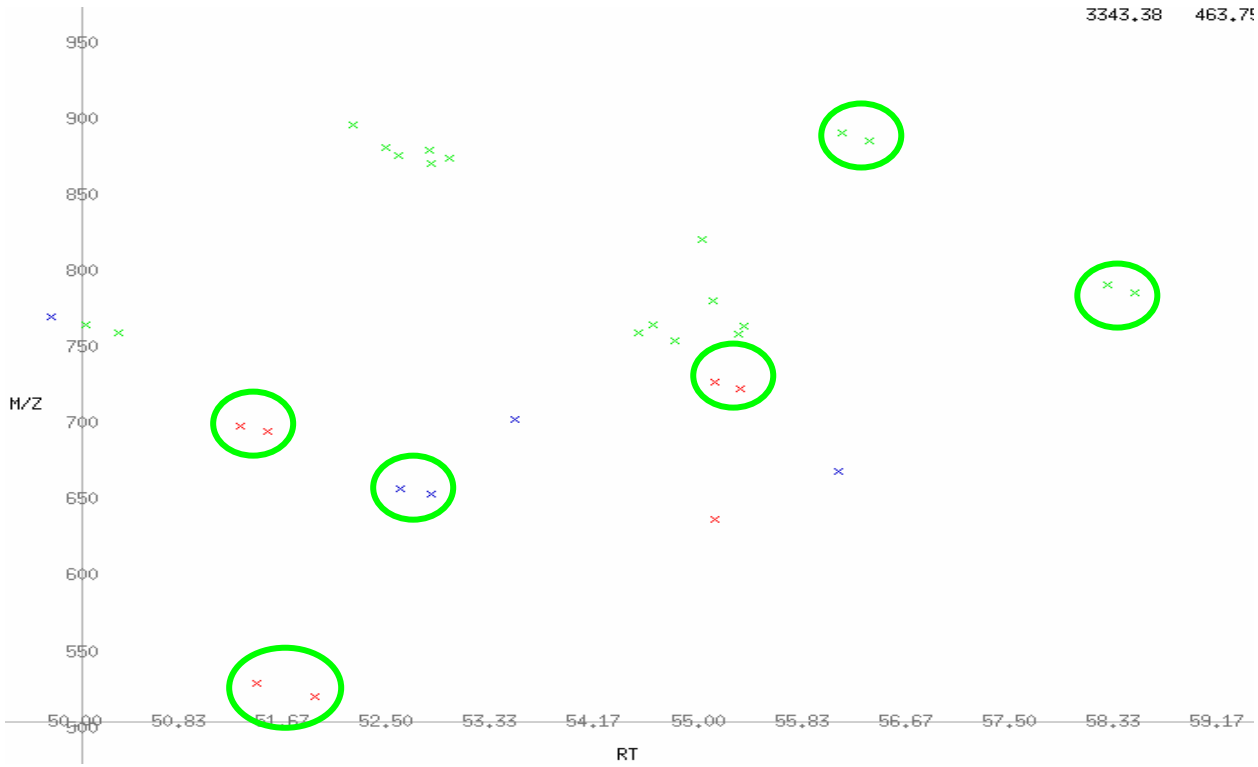
- Motivation for OpenMS
- Bioinformatics Issues in quantitative Proteomics
  - Signal Processing
  - Feature Finding
  - Map Mapping
  - Differential Quantitation
  - Identification, Clustering
  - Software Engineering, Databases
- Subsequent talks give more details

Differential Analysis

Freie Universität Berlin

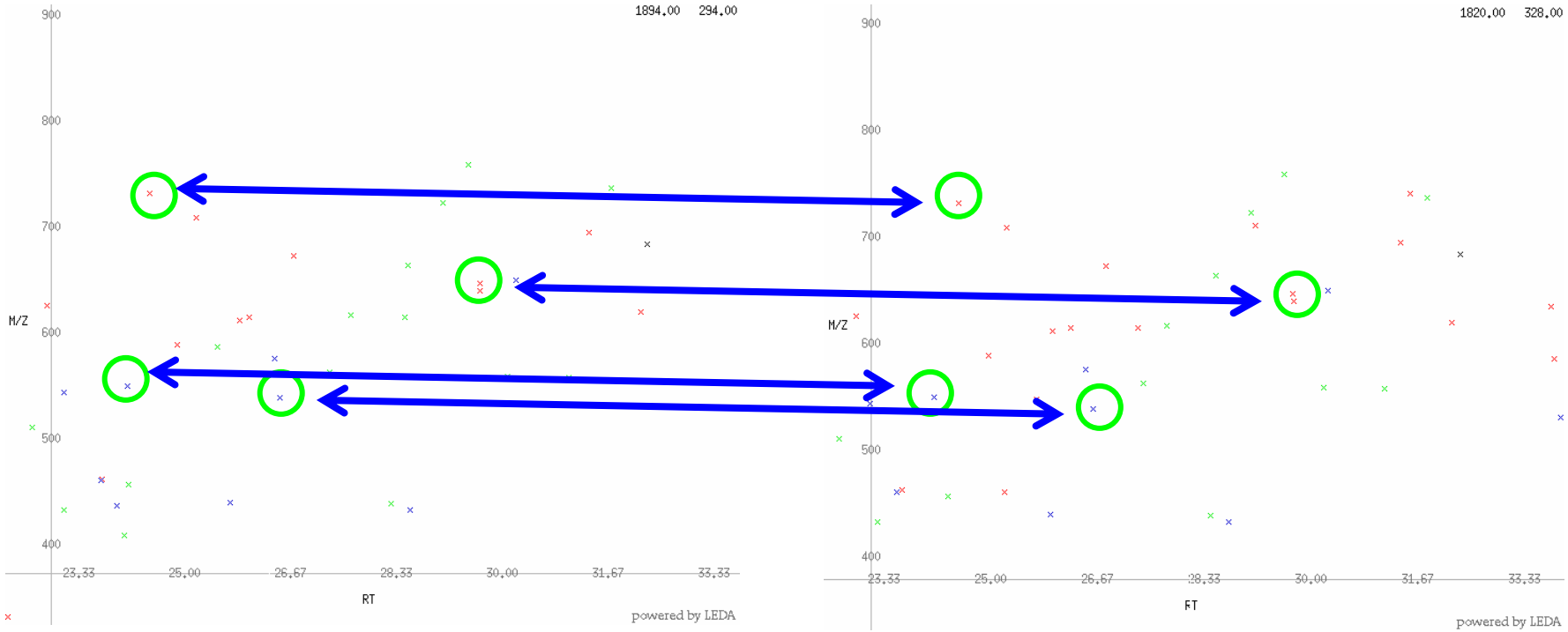
- Identify „differential peptides“
- Depends on quantitation technique
  - Labeling techniques:
    Find pairs **within same map**
  - Unlabeled (DDQ):
    Assign matching pairs **across maps**
- Usually done manually
- In HT settings this *has* to be done automatically!
- Reduction to features helps enormously

ICAT Pair Identification



Assuming charge *z* and *n* Cys in a peptide,
check for pairs $(8n)/z$ Thomson apart in **ONE MAP**

DDQ – Peptide Assignment



Assign pairs across two or more maps

RT of peptide may vary between maps

⇒ compute suitable mapping

(linear function usually suffices)

Multidimensional Search

Freie Universität Berlin

- Use efficient data structures for range searches (e.g. Kd-trees or range trees)
- Employ memory management for cached range queries to allow for large data sets (e.g. large stick maps)
- Making use of robust and efficient implementations of these data structures in CGAL (Computational Geometry Algorithms Library)
- CGAL provides standard data structures for *d*-dimensional computational geometry

Map Normalization

- Intensities will vary between maps, even for the same sample (e.g. volume error during injection)
- Usually this results in a constant ratio of feature intensities
- Map normalization is required to determine this ratio for accurate relative quantitation
- Assuming a sufficient similarity between two maps, the majority of features will be identical, only few will be differ in concentration between the two samples

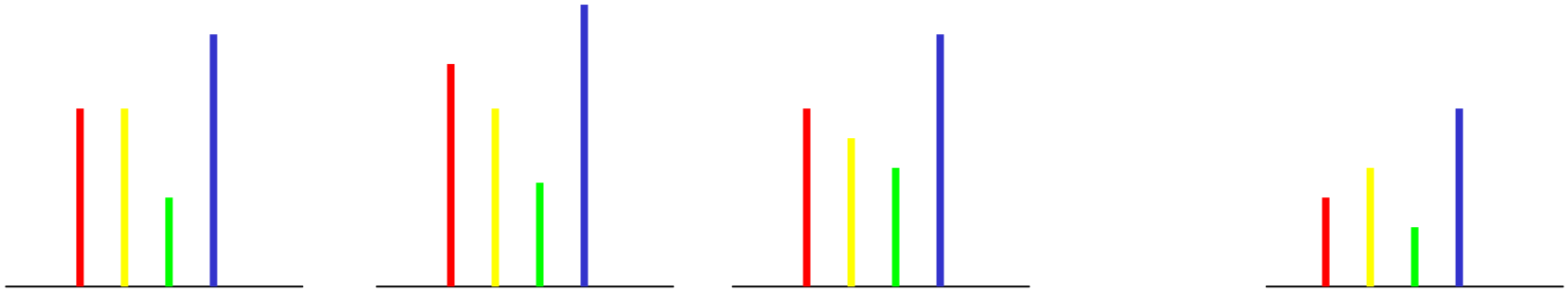
Improve Sensitivity and Specificity

Freie Universität Berlin

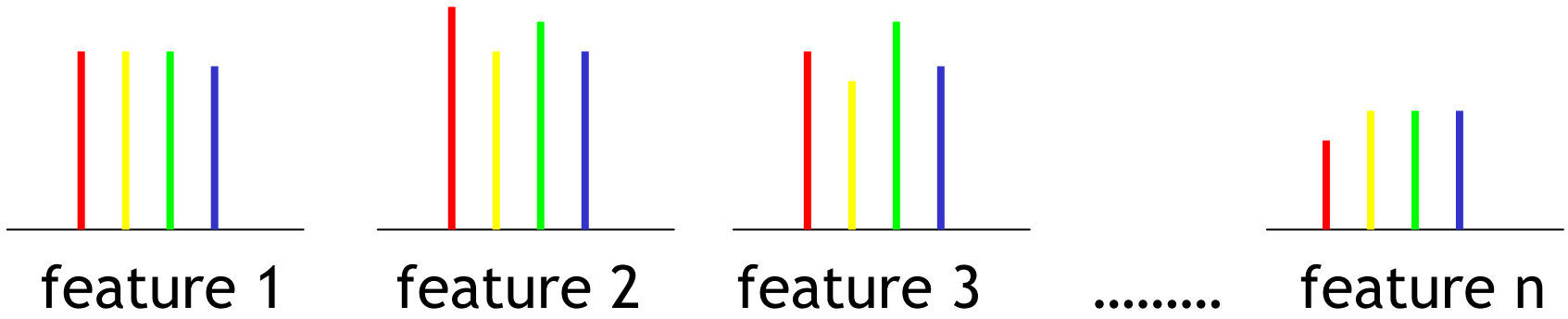
Run several experiments on the same sample

- Too time-consuming for manual assignment
- Automatic assignment increases reliability
  1. Features that show up most of the time are real
  2. Features that do not are probably noise
  3. Allows statements about statistical significance

Normalize and group measurements

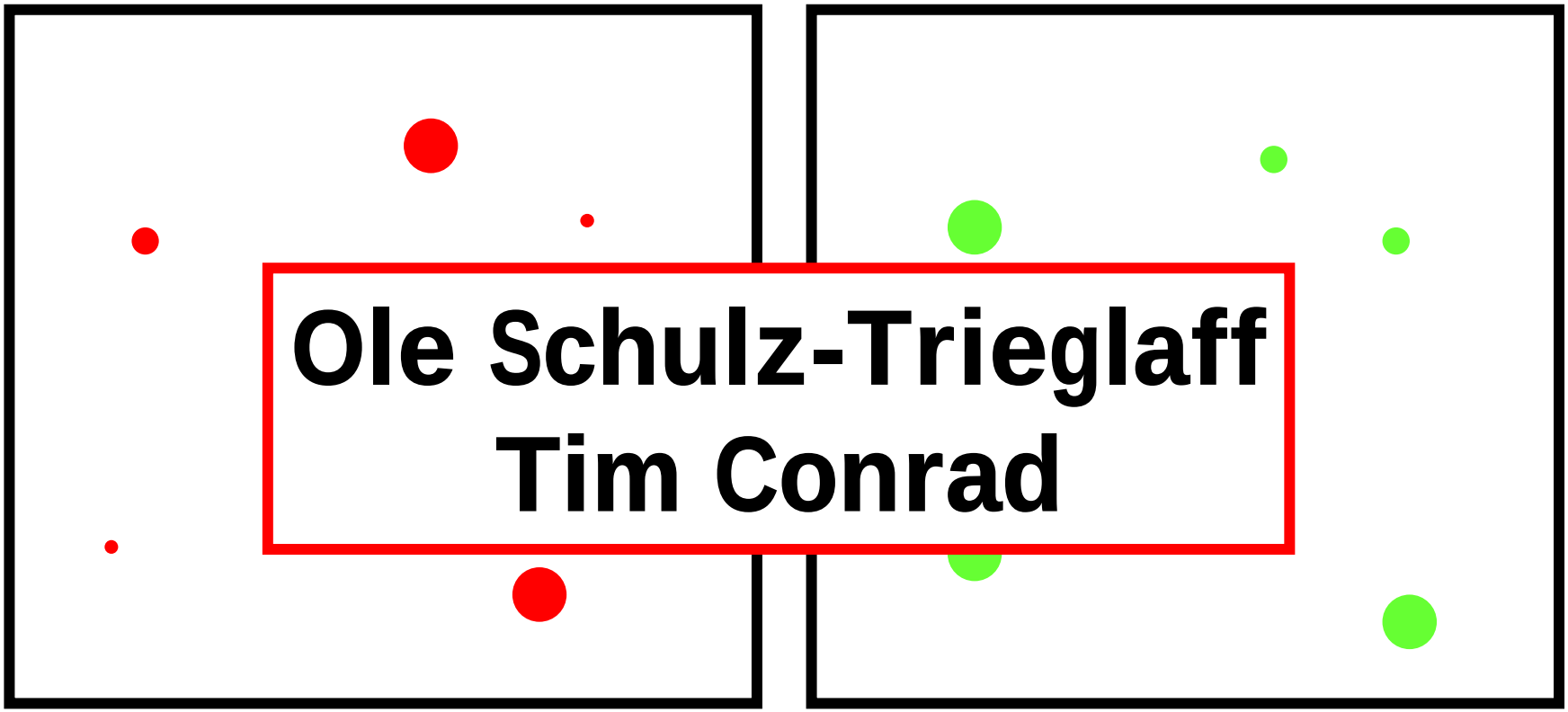


Compute scaling between pairs of maps, group them



Statistical analysis of differential patterns

Freie Universität Berlin

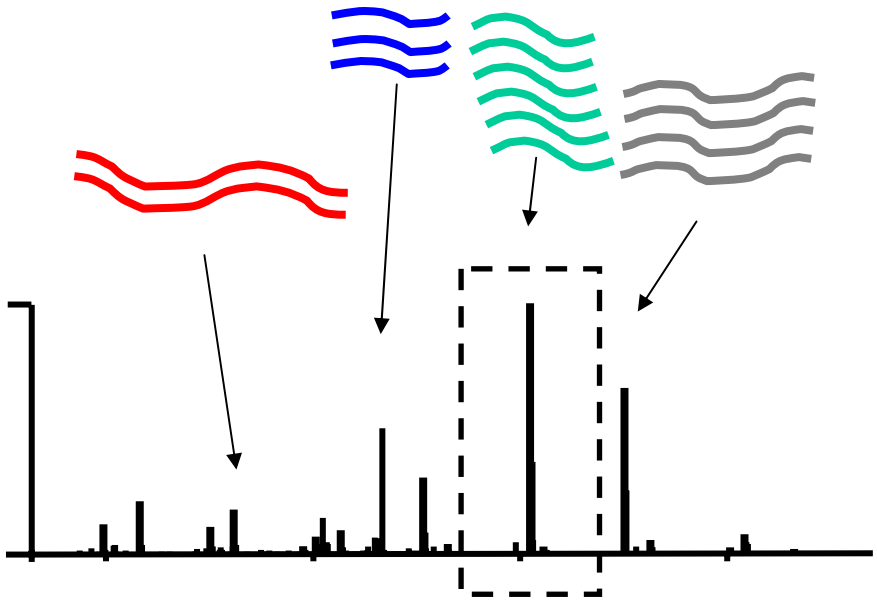


Overview

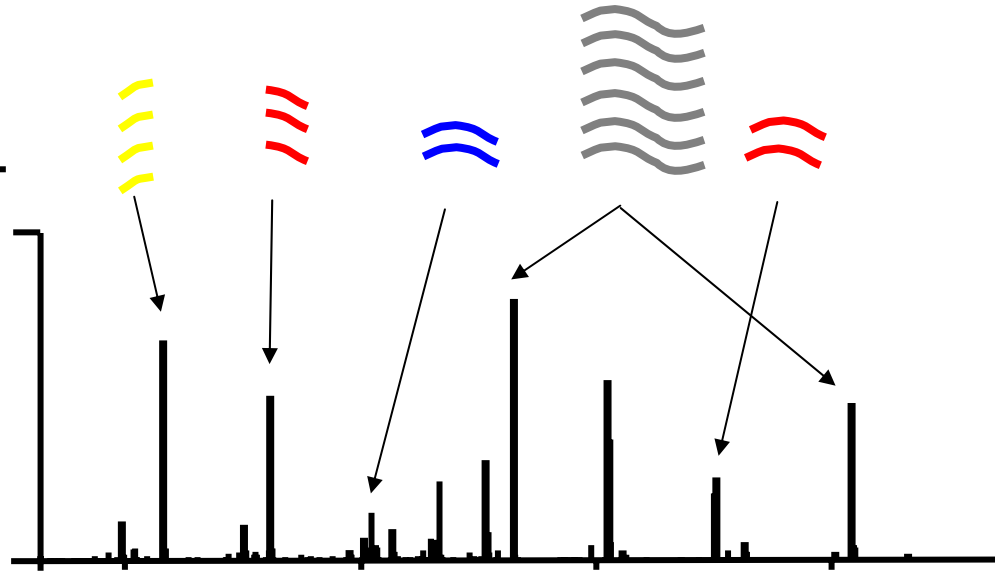
Freie Universität Berlin

- Motivation for OpenMS
- Bioinformatics Issues in quantitative Proteomics
  - Signal Processing
  - Feature Finding
  - Map Mapping
  - Differential Quantitation
  - Identification, Clustering
  - Software Engineering, Databases

Peptide Identification (MS-MS)



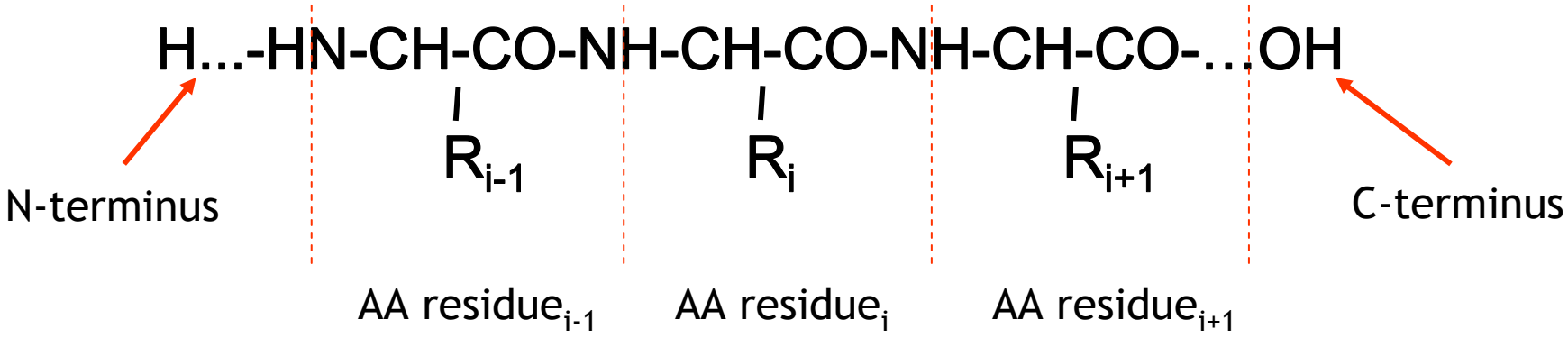
Secondary fragmentation



Fragment Spectrum

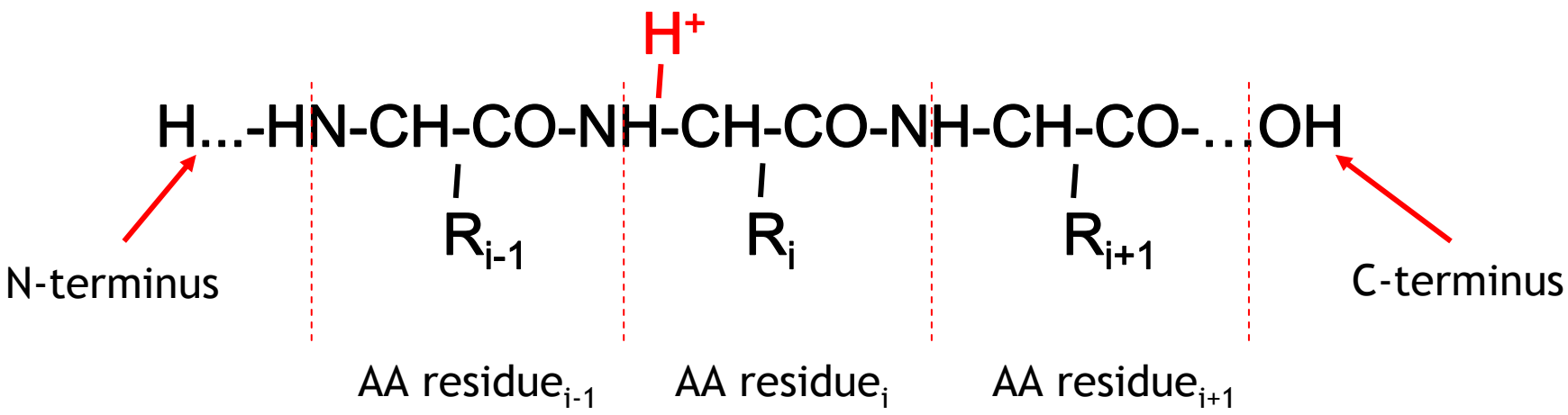
Peptide Fragmentation

The peptide backbone breaks to form fragments with characteristic masses.



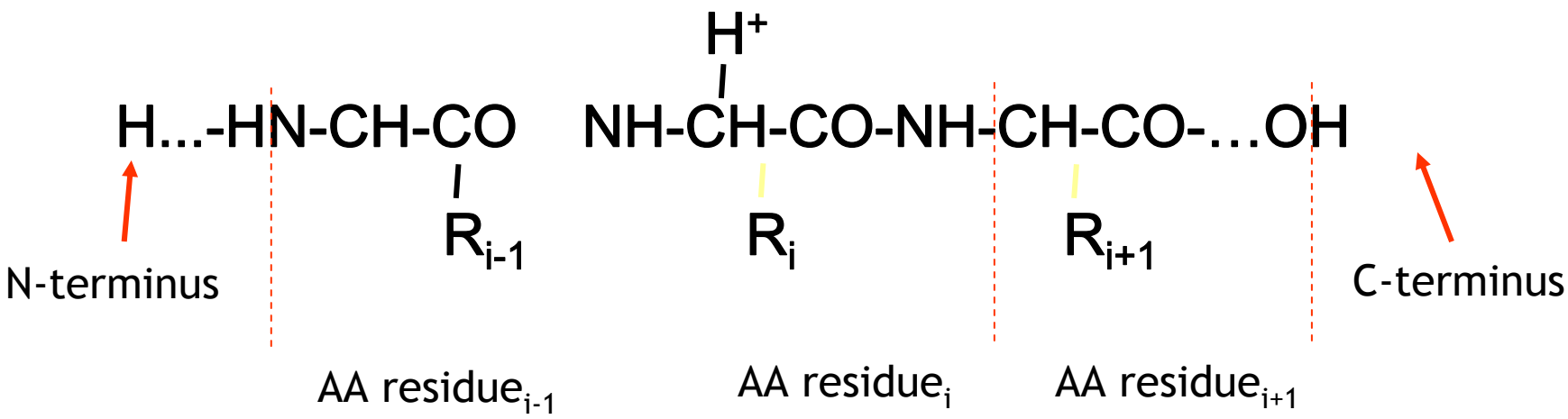
Peptide Fragmentation

The peptide backbone breaks to form fragments with characteristic masses.



Peptide Fragmentation

The peptide backbone breaks to form fragments with characteristic masses.

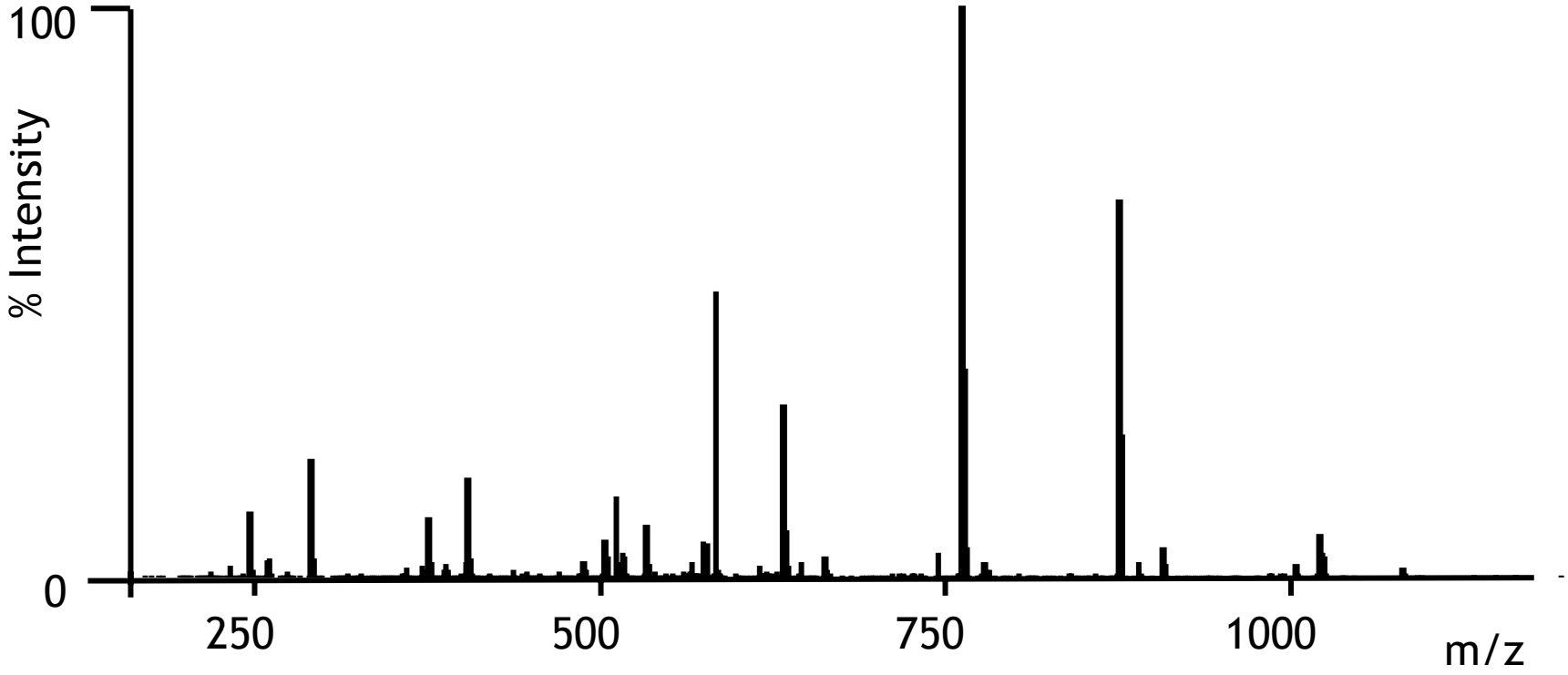


Ionized peptide fragment

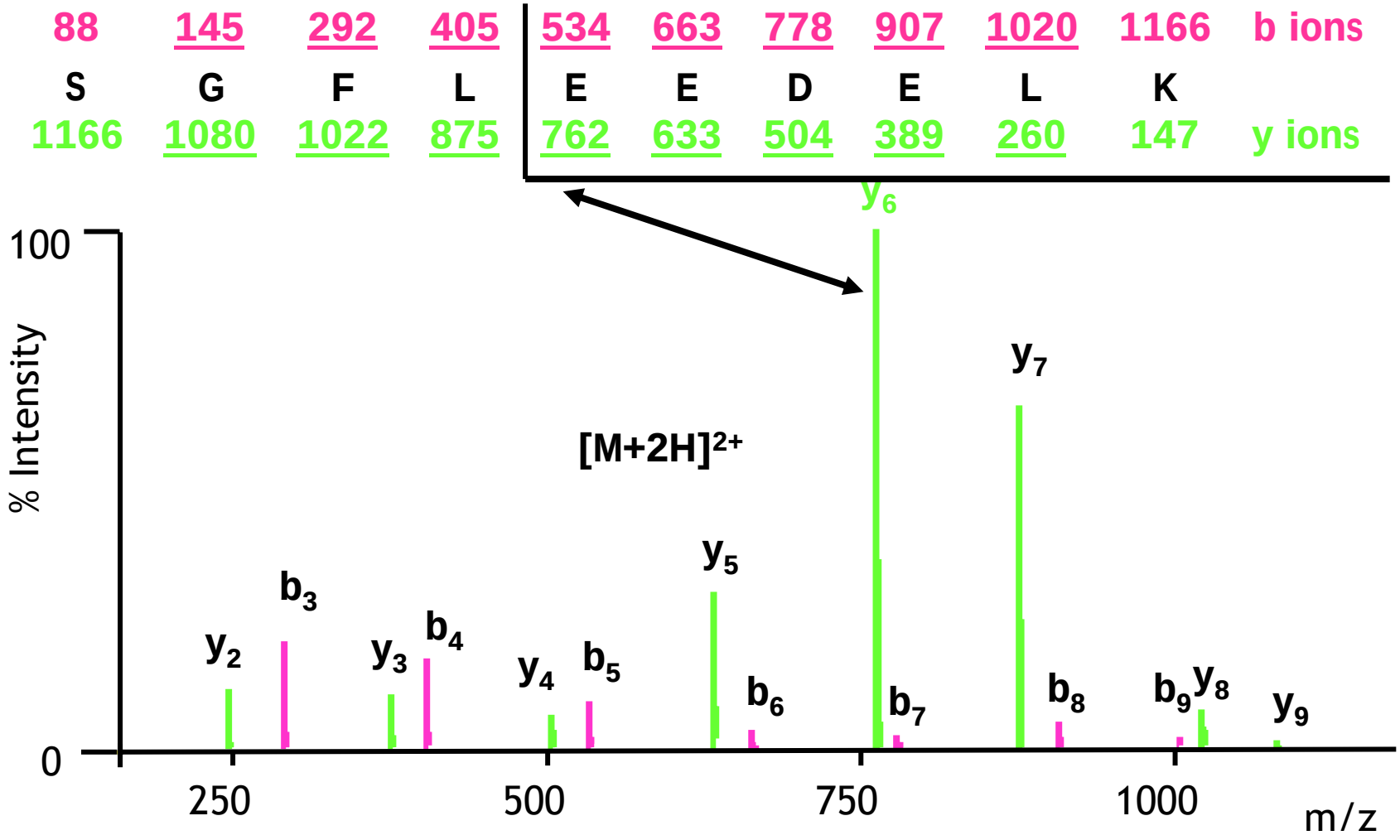
Peptide Ions in Spectrum

Freie Universität Berlin

88	145	292	405	534	663	778	907	1020	1166	b ions
S	G	F	L	E	E	D	E	L	K	
1166	1080	1022	875	762	633	504	389	260	147	y ions



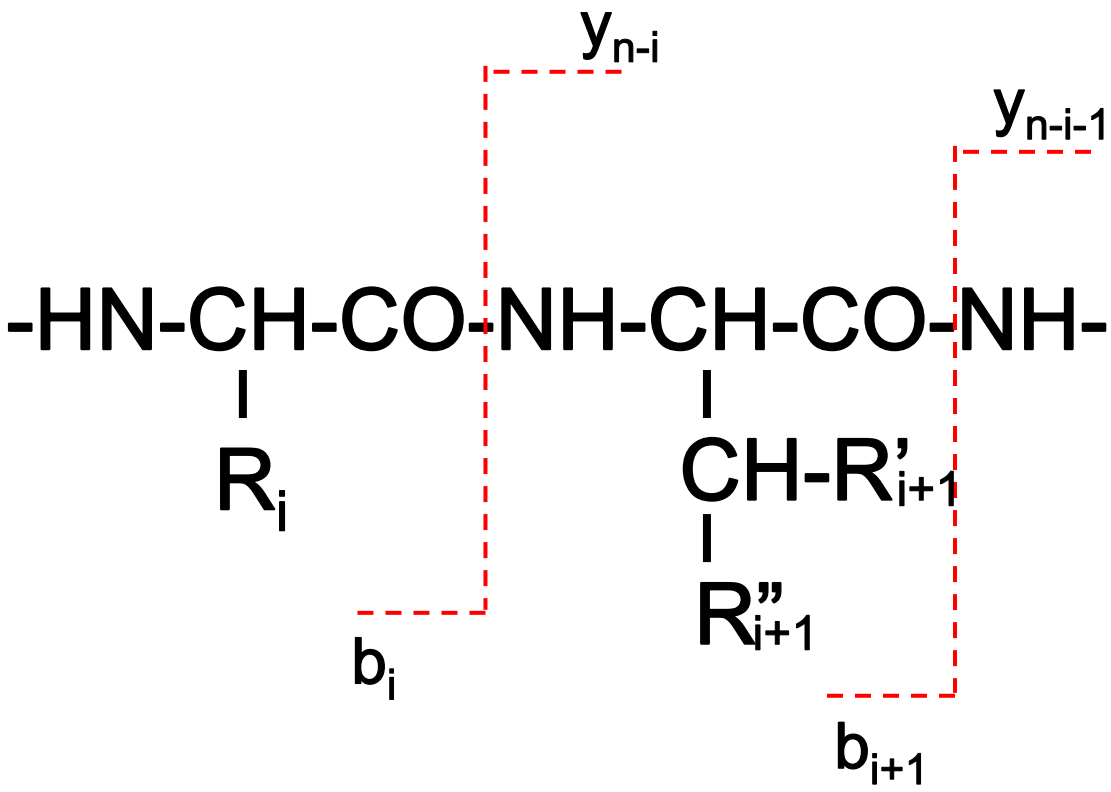
Peptide Ions in Spectrum



What's the problem?

Freie Universität Berlin

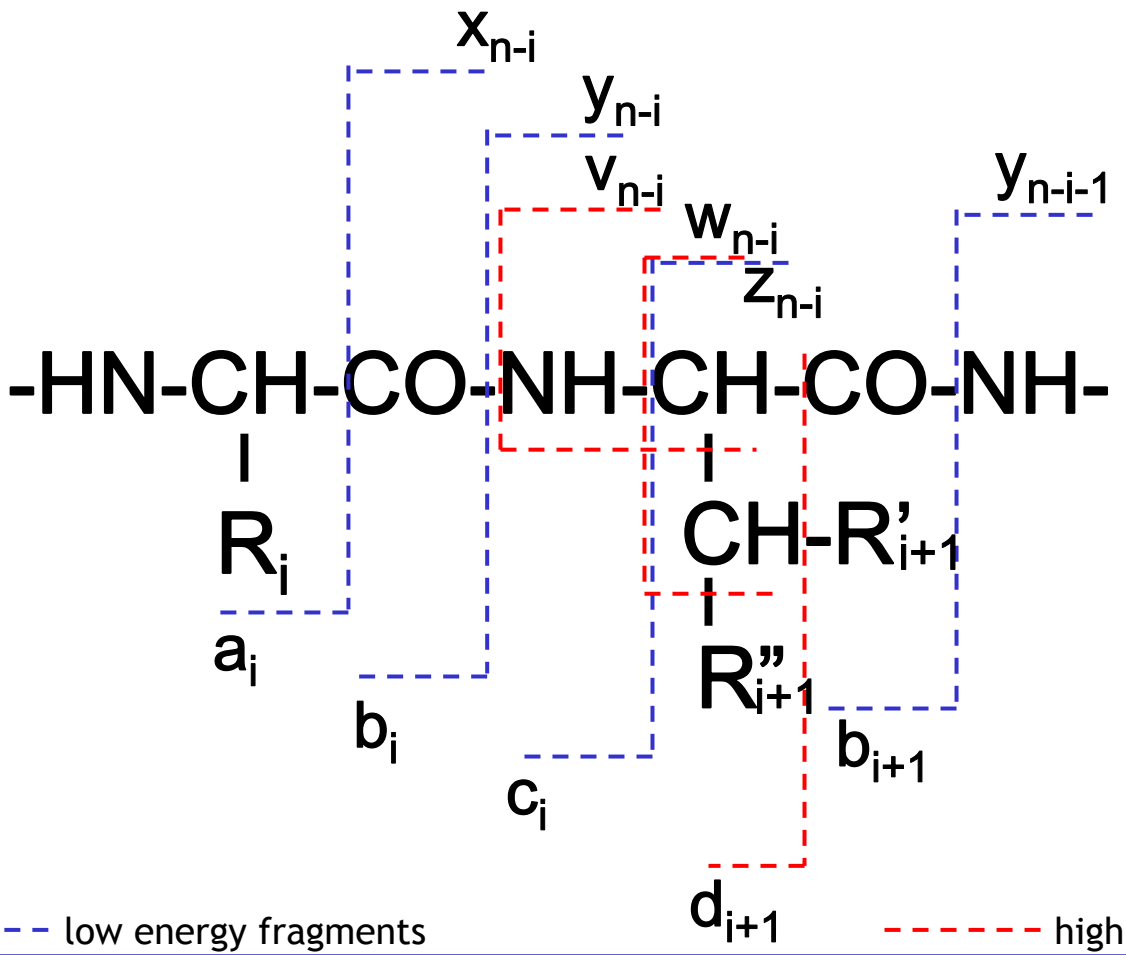
Peptide fragmentation possibilities (ion types)



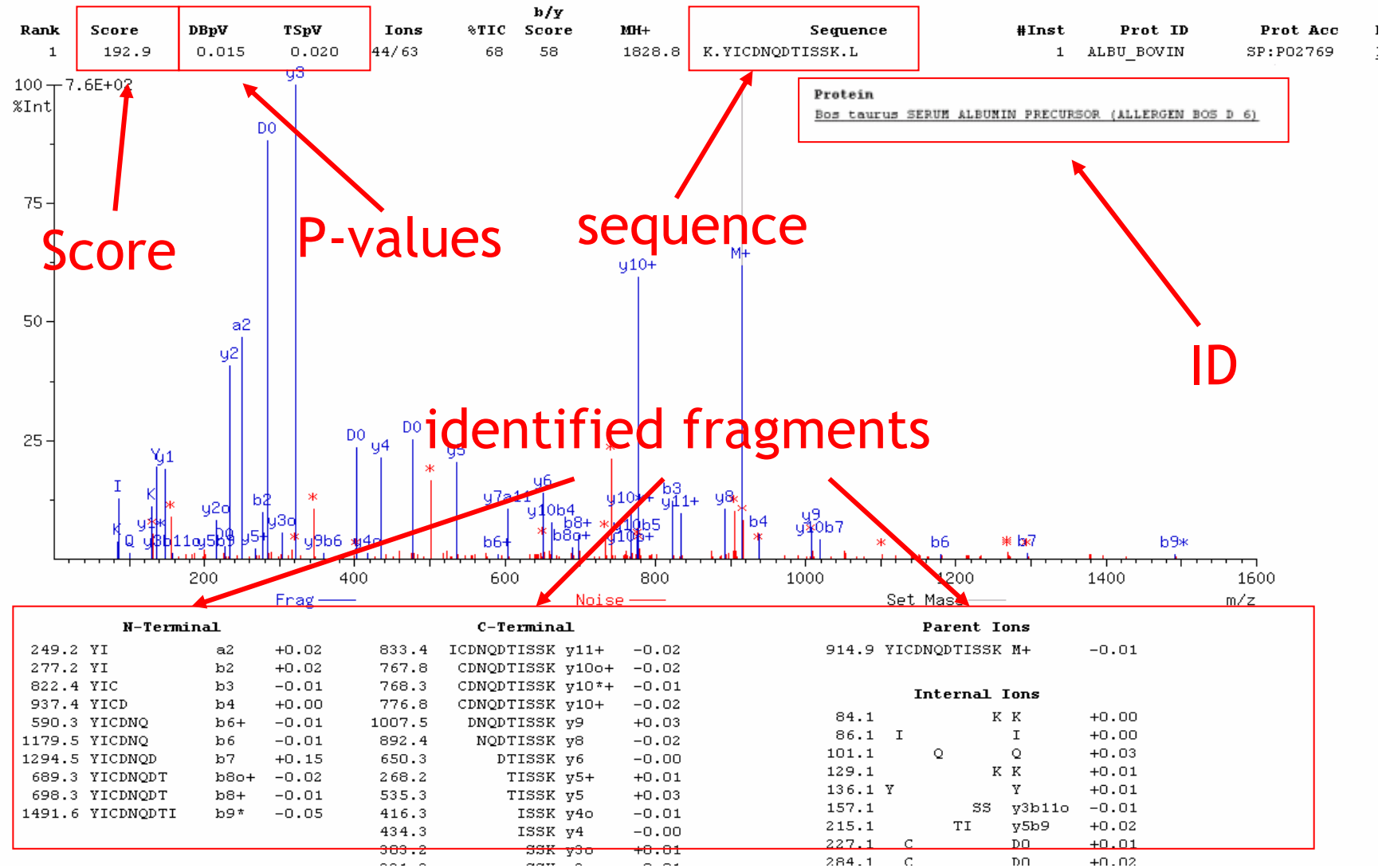
What's the problem?

Freie Universität Berlin

Peptide fragmentation possibilities (ion types)



Identification – SCOPE



Vineet Bafna, Nathan Edwards, Proc. ISMB 2001

Overview

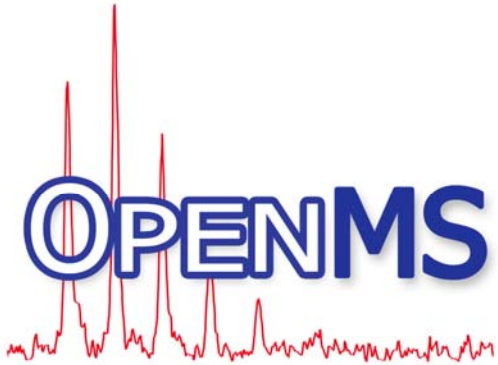
Freie Universität Berlin

- Motivation for OpenMS
- Bioinformatics Issues in quantitative Proteomics
  - Signal Processing
  - Feature Finding
  - Map Mapping
  - Differential Quantitation
  - Identification, Clustering
  - Software Engineering, Databases

OpenMS

Freie Universität Berlin

- Software **framework** for shotgun proteomics
- **ISO/ANSI C++ compliant**
- Features
  - **Open Source** (LGPL)
  - **Data structures & algorithms**
    - *d*-dimensional core
    - Signal processing for raw data conversion
  - Data **import/export** in standard formats
  - **Database** storage of all data
    (using an extended PEDRO/MIAPE model)
  - DB-driven **workflow** management
  - **Visualization** for spectra, maps

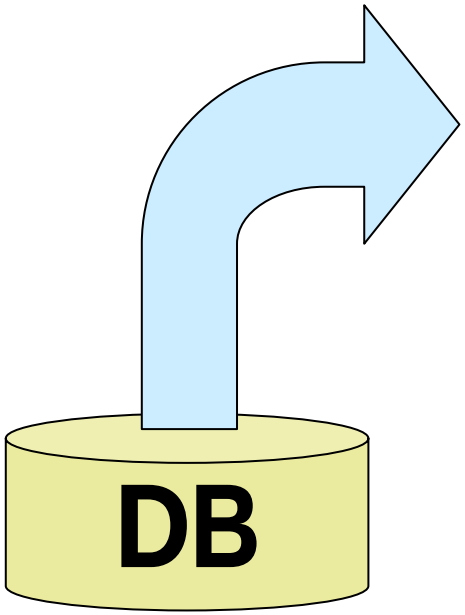


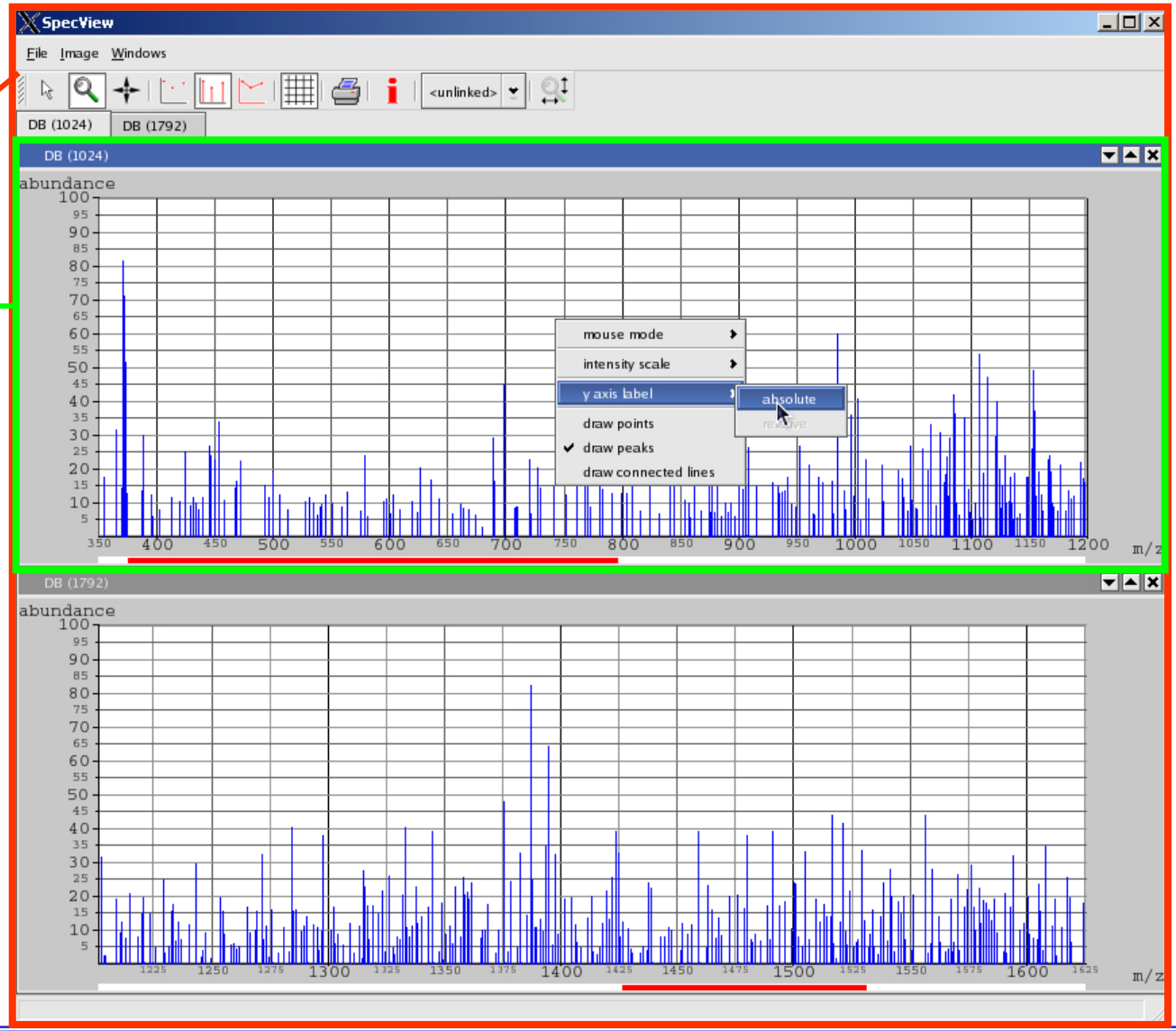
SpecView – 1D Window

Freie Universität Berlin

MDI Window

1D Window

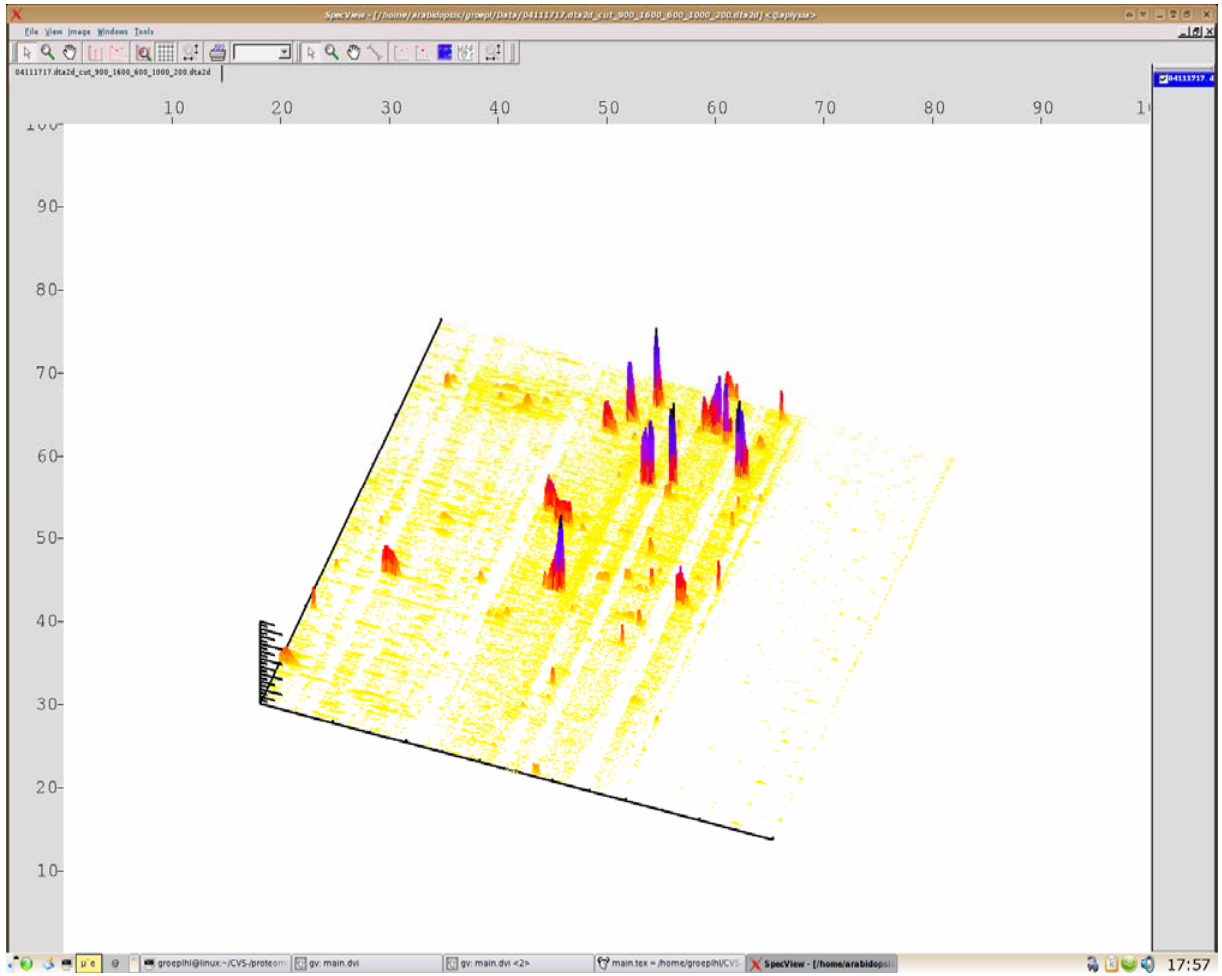




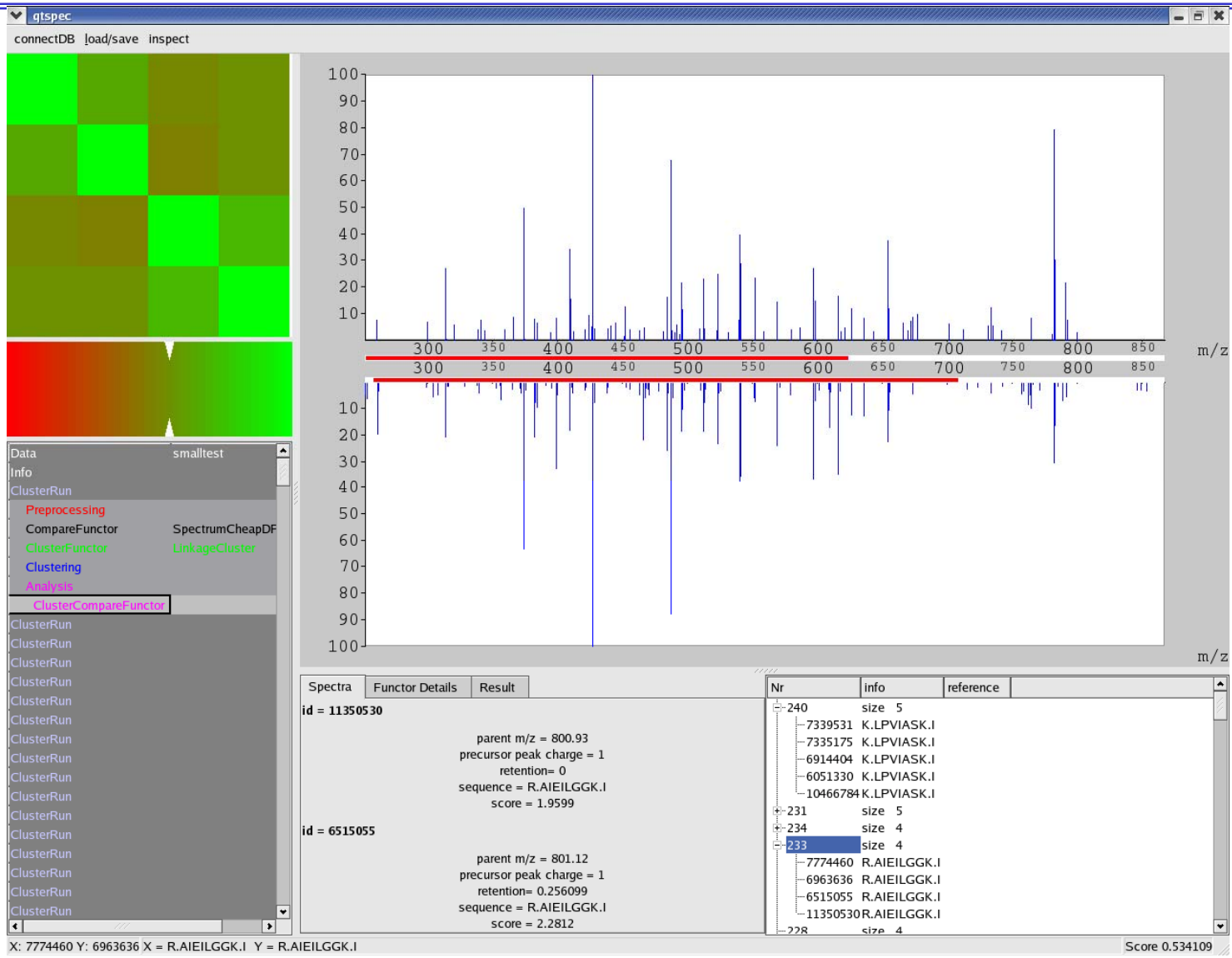
SpecView – 3D Window

Freie Universität Berlin

- Visualization of raw, stick, and feature maps
- Rapid interactive zoom in/out through efficient quad tree implementation
- Data import from files or direct visualization from the DB



Clustering/ID View



Visualization Code Example


```
// create application with a single widget
QApplication app(argc, argv);
Spectrum1DWidget* wid = new Spectrum1DWidget;
app.setMainWidget(wid);

// load data from DB
DBAdapter dba;
dba.connectDB("OpenMS", "<user>", "<password>");
dba.executeQuery("SELECT id FROM PeakList … ", true);
wid->setDataSet(dba.objects()[0]);

// execute application
wid->show();
return app.exec();
```

Acknowledgements

Dr. Clemens Gröpl
Eva Lange, Tim Conrad,
Ole Schulz-Trieglaff
(Algorithmische Bioinformatik, FU Berlin)

Prof. Hartmut Schlüter
(Universitätsmedizin Berlin, Charité)

Prof. Dr. Oliver Kohlbacher
Marc Sturm, Andreas Bertsch
Jens Joachim
(SBS/WSI, Tübingen)

Andreas Hildebrandt
(Uni Saarbrücken)

Prof. Dr. Christian Huber
Bettina Mayr et al.
(Instr. Analytik & Bioanalytik, Univ. des Saarlandes, Saarbrücken)

Dr. Albert Sickmann
(Virchow-Zentrum, Würzburg)

Herbert Thiele
Jens Decker
(Bruker Daltonics, Bremen)

Dr. Christoph Klein
(IRMM, Geel; now IHCP Ispra)

Überblick Assemblierung

Freie Universität Berlin

1. Überblick : DNA Sequenzierung
2. Überblick : BAC-by-BAC Assembler
3. Überblick : WGS Assembler
4. Überblick : Compartmentalized Assembler
5. Details der CSA Pipeline

Danke für die Aufmerksamkeit !!!

DNA Sequenzierung

Freie Universität Berlin

Shotgun DNA Sequenzierung (Technologie)