



BIOINFORMATICS FOR BIOANALYTICAL ANALYSES

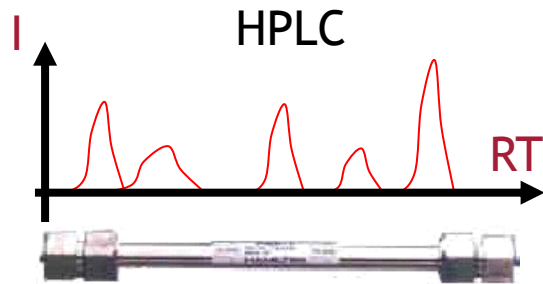
Professor Dr. Sven Nahnsen

Quantitative Biology Center (QBiC)

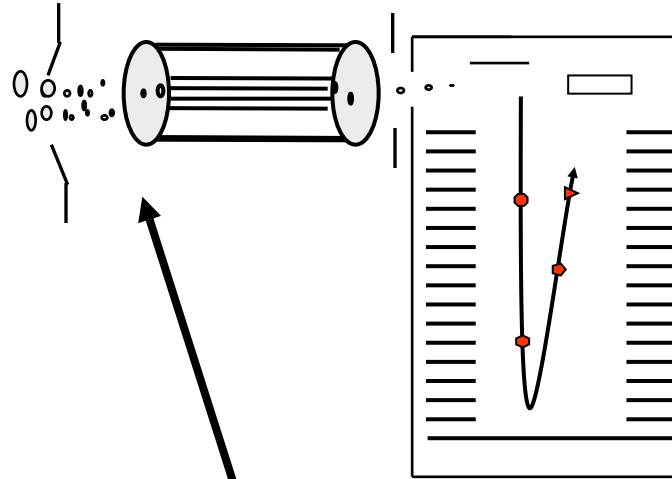
Summer term 2022



- Quantification using mass spectrometry
 - Recall of basic principles
 - Basic terms from analytical chemistry
- Identification using mass spectrometry
 - Database search in proteomics
 - Algorithms for peptide and protein identification
 - Identification in metabolomics
 - Accurate mass ID
- Downstream analysis
 - Principle component analysis

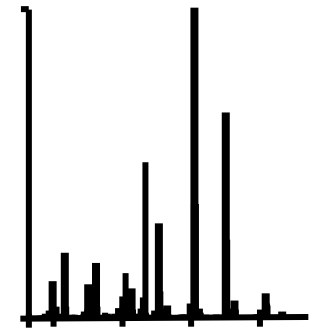


ESI



TOF

Spectrum (scan)



Separation 1
separate peptides
by their retention
time on column

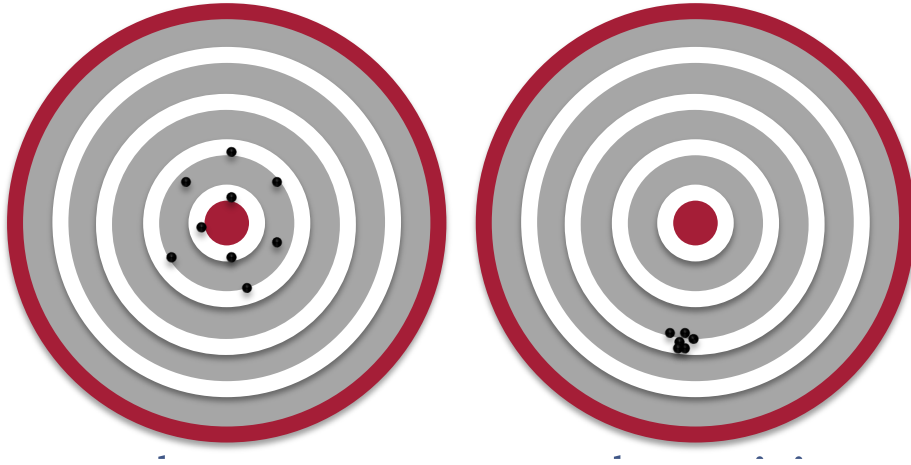
Ionization
electrospray,
transfers charge
to the peptides

Separation 2
MS separates by
mass-to-charge
ratio (m/z)

- **Analyte** – the stuff we want to analyze
- **Matrix** – the components of the sample that are not analytes
- The matrix can significantly impact the way the whole analysis is performed
- Example:
 - Proteomics analysis from urine
 - Urine contains
 - Proteins and peptides – the analyte
 - Water
 - Metabolites
 - Urea
 - ...

} **matrix**

Precision and Accuracy



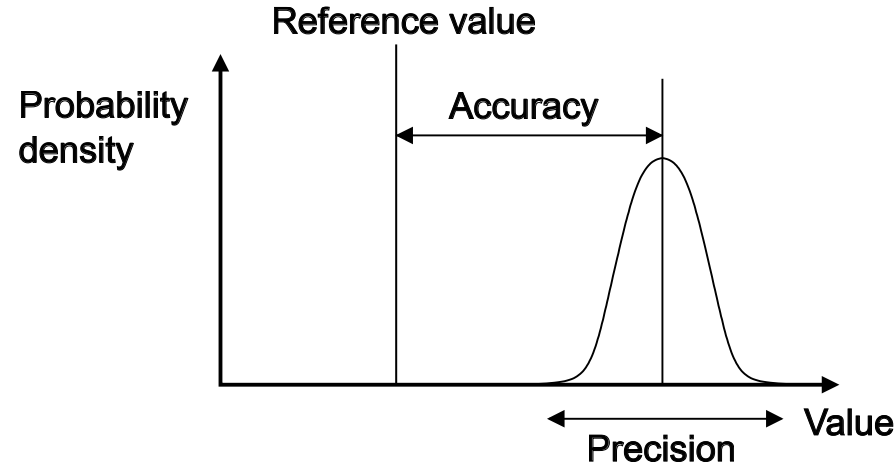
Precision and Accuracy



good accuracy,
poor precision

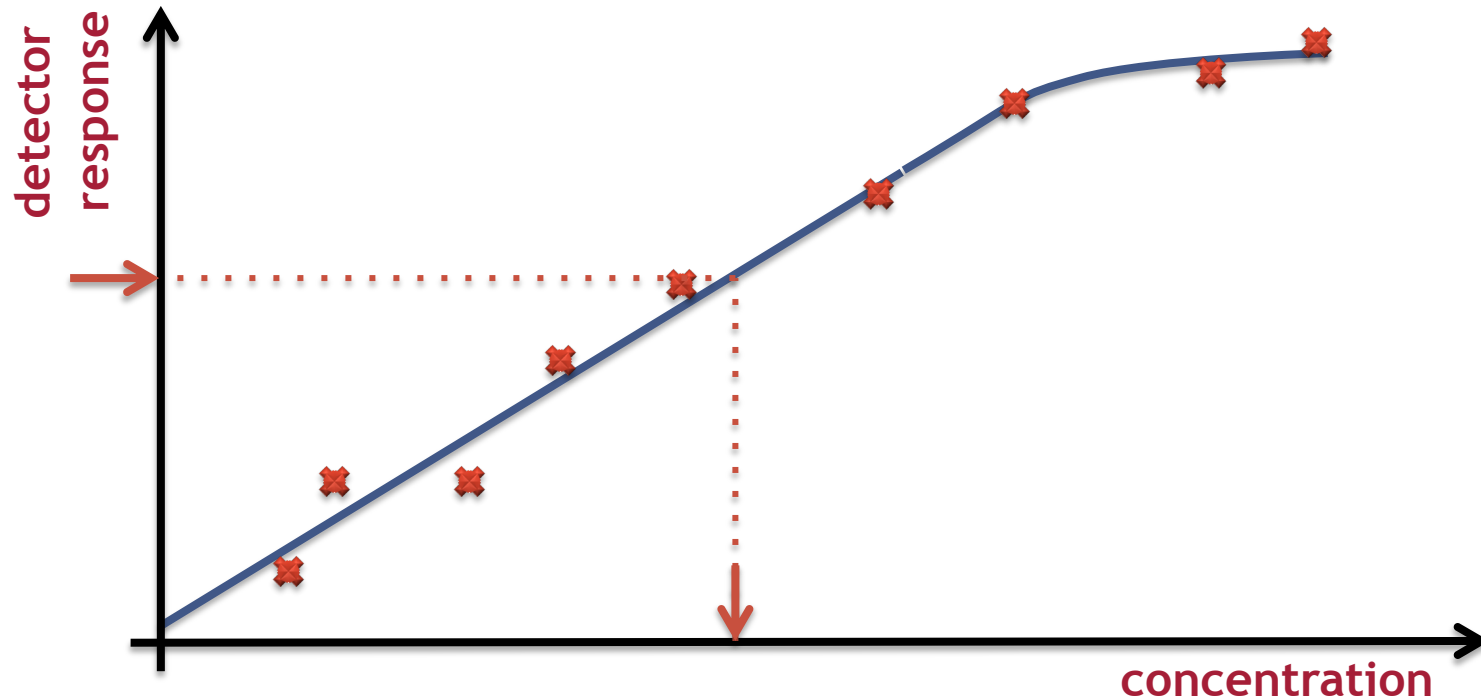


good precision,
poor accuracy

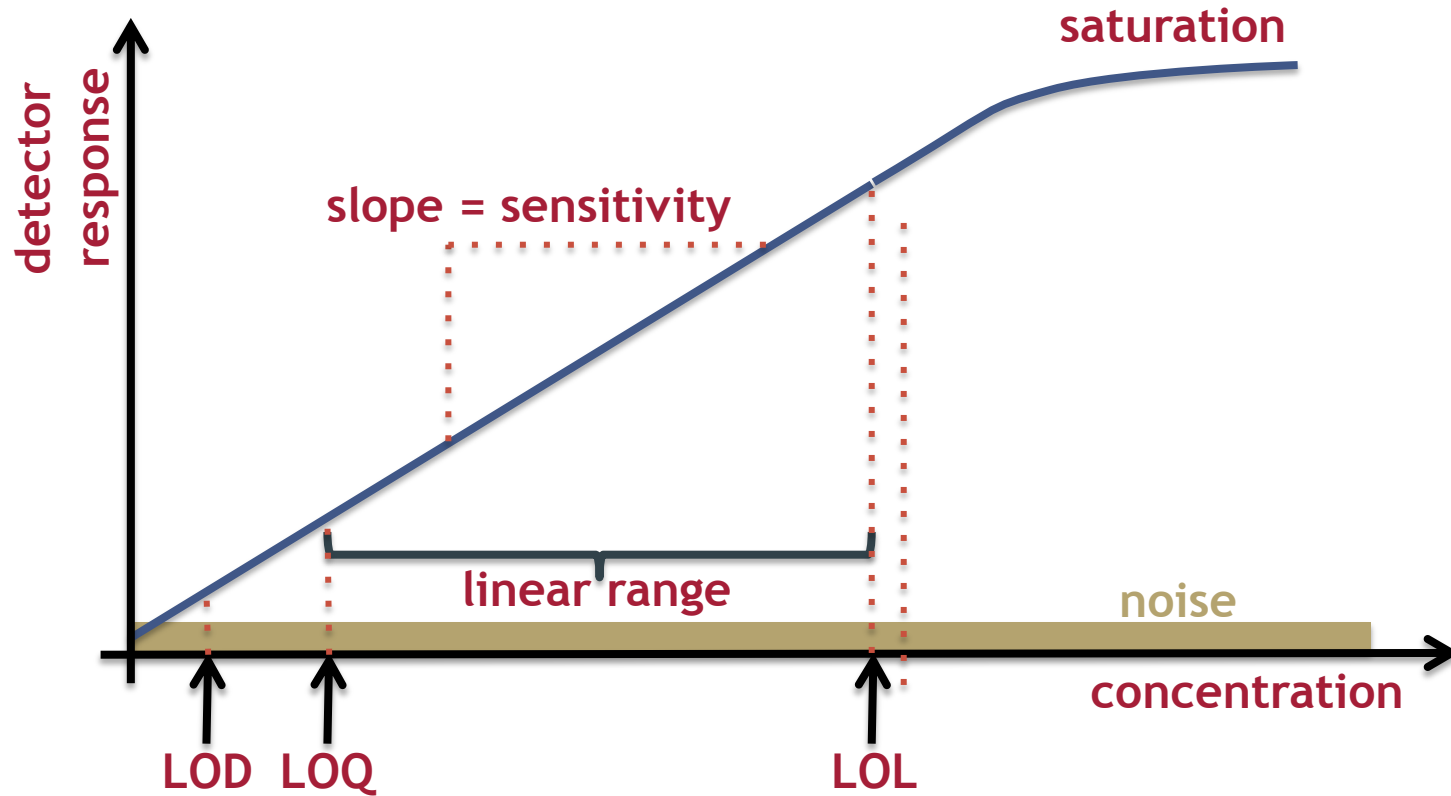


- **Accuracy**: closeness to the true value (mostly influenced by systematic error) – repetition of the experiment will not improve the result
- **Precision**: repeatability of the measurement (mostly influenced by random error) – repetition of the experiment will yield more a value closer to the true value
- An ideal experiment combines high accuracy with high precision

Calibration Curve



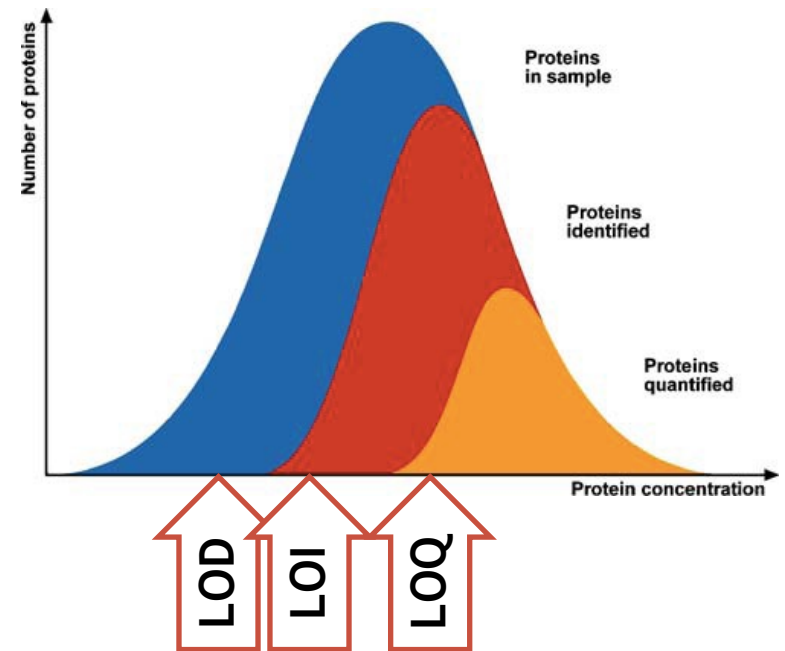
- Measurement of the detector response for various (known) concentrations allows the construction of a calibration curve
- Most detector responses are chosen in a way that the response changes linearly with the concentration
- Once the calibration curve has been measured, it allows the determination of the concentration of an unknown sample



- **LOD**: level of detection – at what concentration can we decide that the analyte is present
- **LOQ**: level of quantification – at what concentration can we accurately quantify it
- **LOL**: limit of linearity – saturation effects start here
- **Linear range (dynamic range)**: the concentration range where we get a response that is linear in the concentration



- Proteomics
 - More peptides/proteins are usually identified than quantified
 - Identification: MS/MS, quantification usually by MS
-> independent processes
 - Many things can be seen (detected) but cannot be identified or quantified

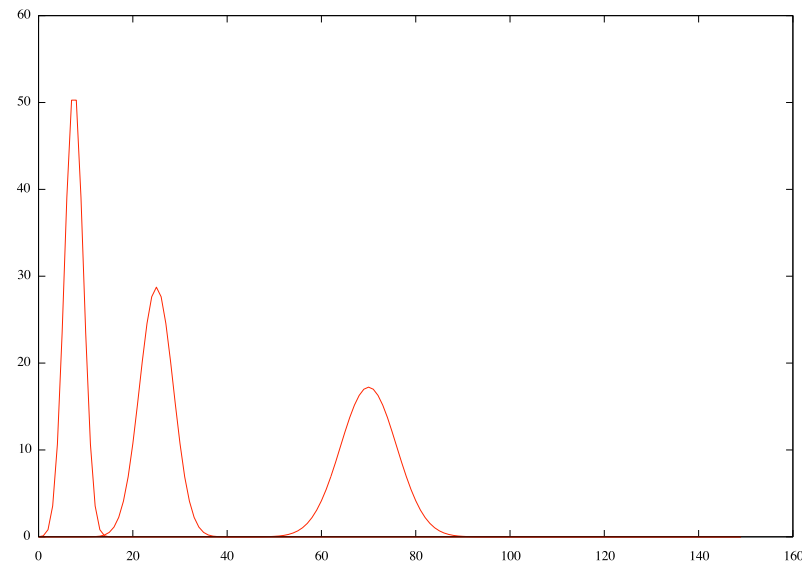


LOI: “Level of identification”

Peak Width in Chromatography

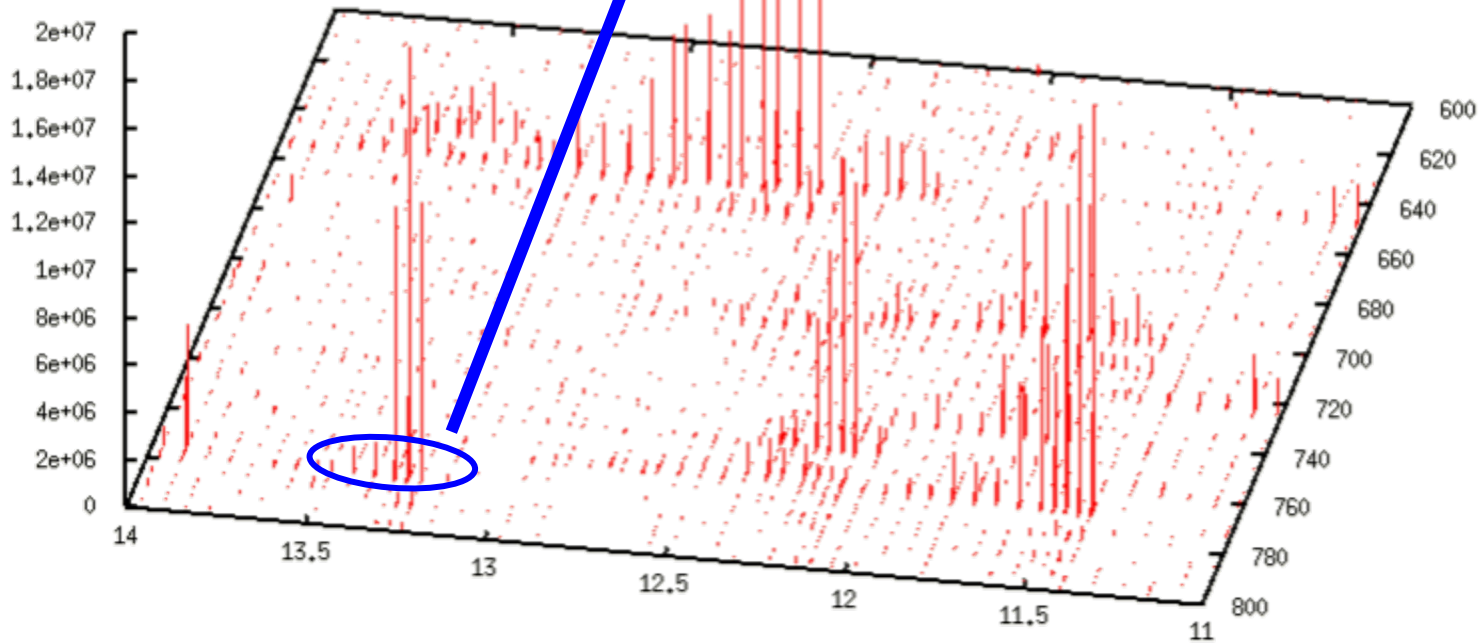


- Peaks broaden with retention time
- Early peaks are sharp and narrow
- Later peaks tend to be broader
 - Diffusion along the column during the separation
 - Analytes eluting later had more time to diffuse
- Peak area remains constant, though
 - No analyte is getting lost
 - Increased width is compensated by reduced height



Quantification

(15 nmol/ μ l, 3x over-expressed, ...)

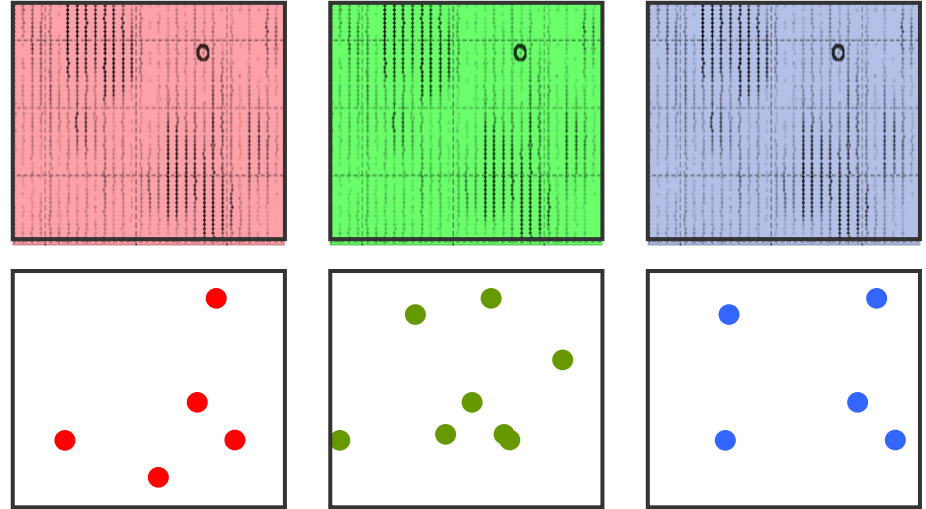




- Label-free quantification is probably the most natural way of quantifying
 - **No labeling required**, removing further sources of error, no restriction on sample generation, cheap
 - Data on different samples acquired in different measurements – **higher reproducibility needed**
 - **Manual analysis difficult**
 - **Scales very well** with the number of samples, basically no limit, no difference in the analysis between 2 or 100 samples

LFQ – Analysis Strategy

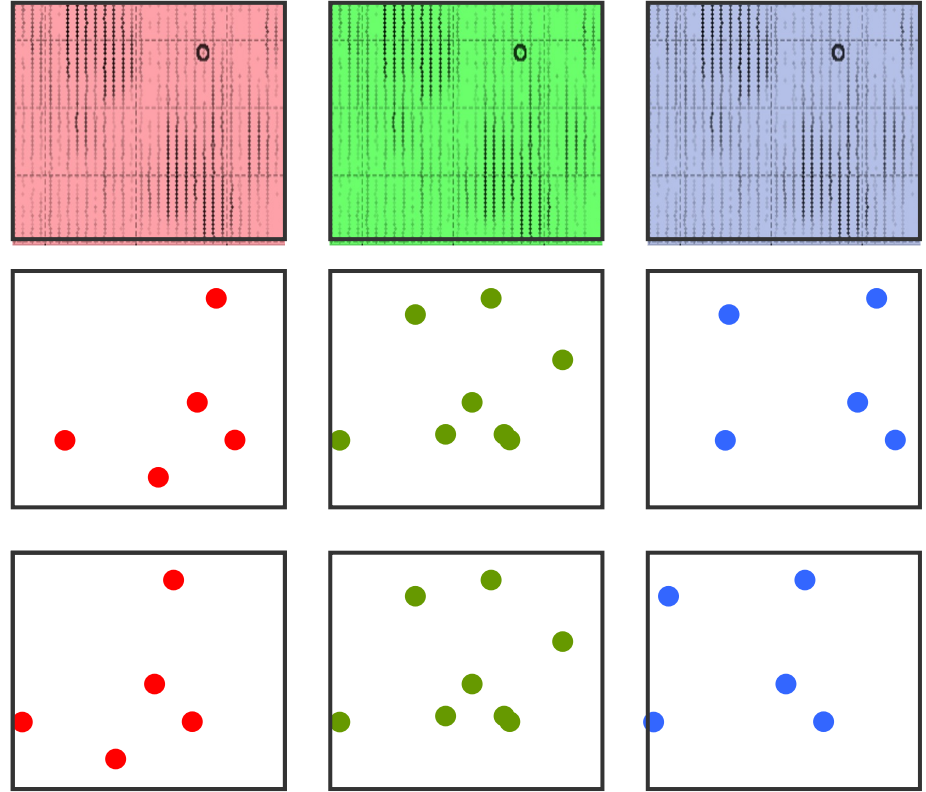
1. Find features in all maps



LFQ – Analysis Strategy



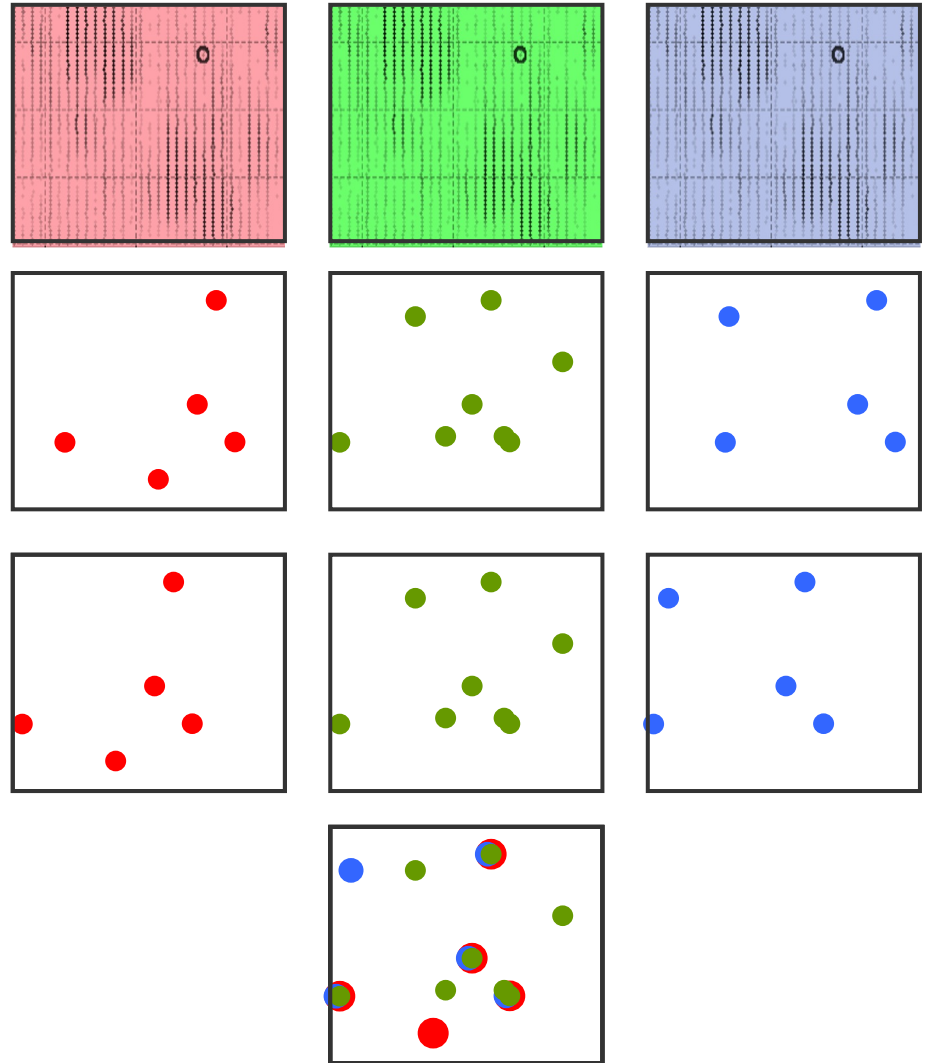
1. **Find** features in all maps
2. **Align** maps



LFQ – Analysis Strategy



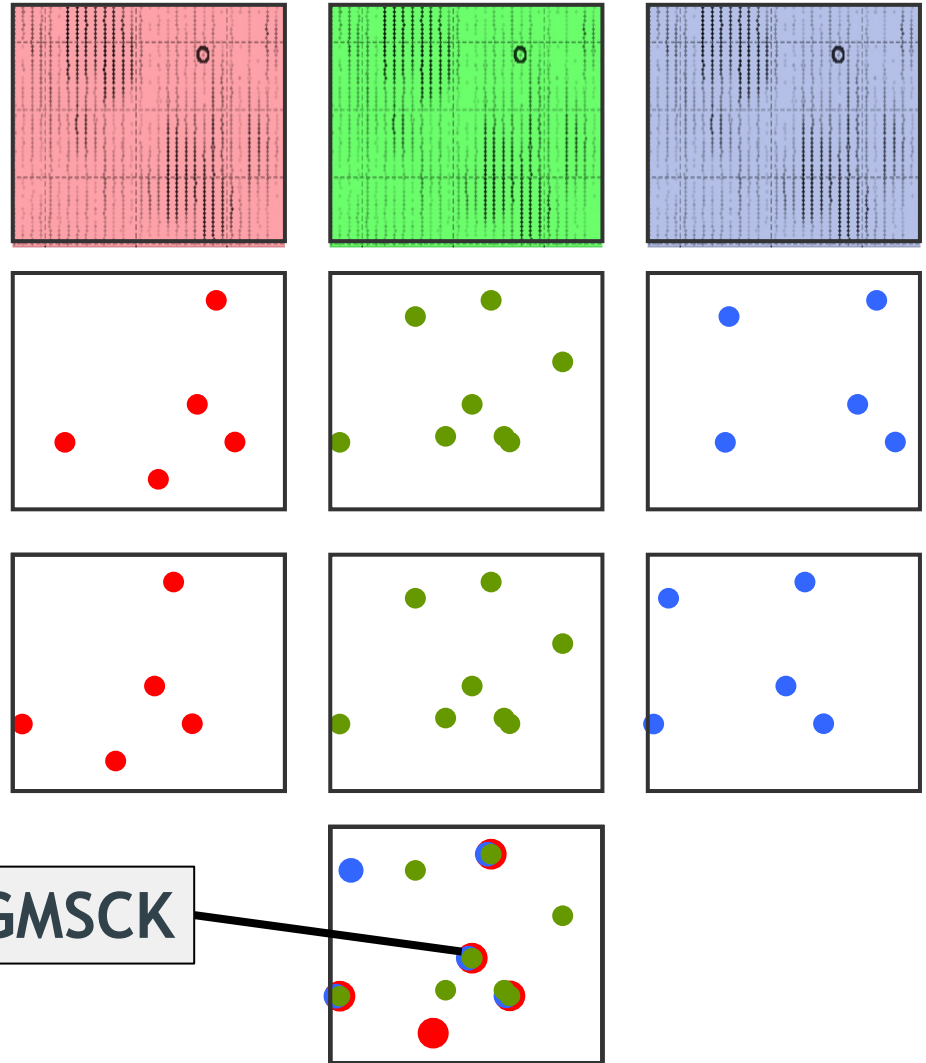
1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features



LFQ – Analysis Strategy



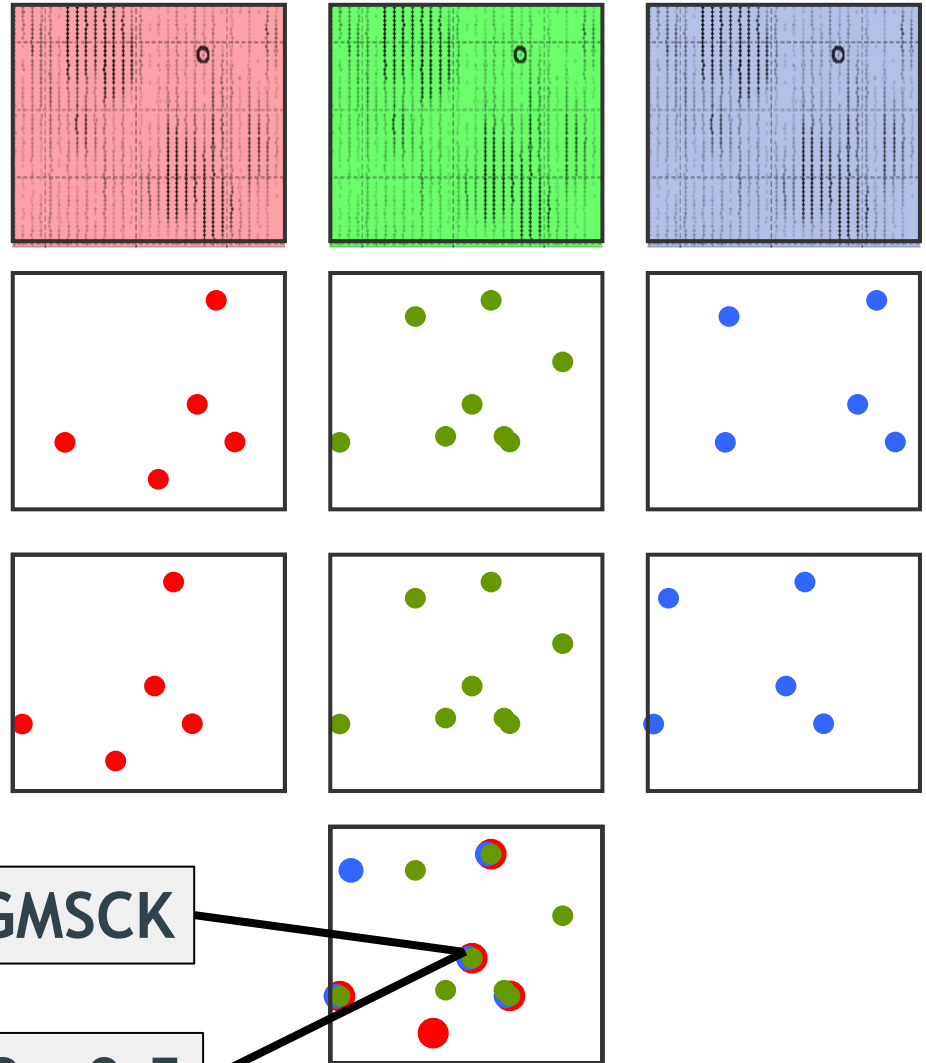
1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features



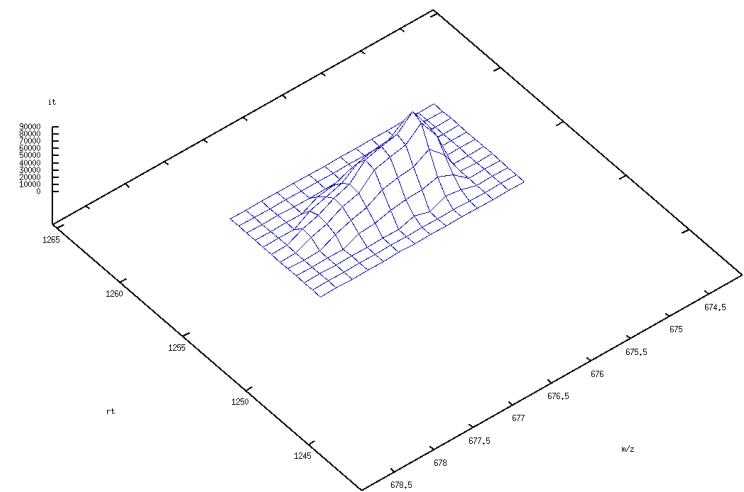
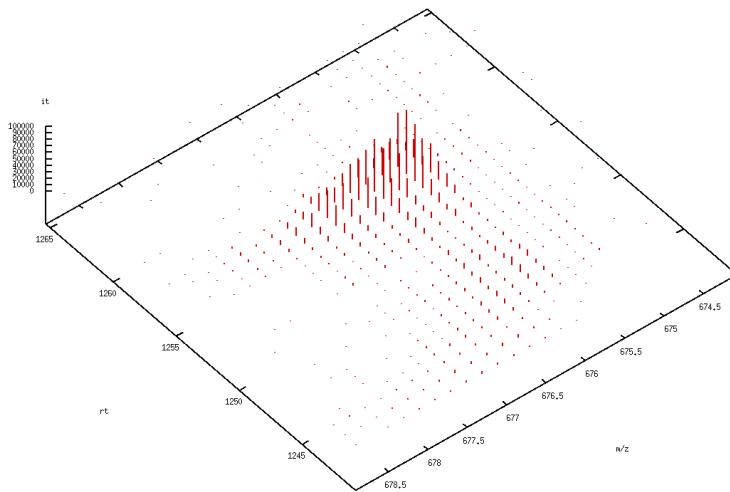
LFQ – Analysis Strategy



1. **Find** features in all maps
2. **Align** maps
3. **Link** corresponding features
4. **Identify** features
5. **Quantify**



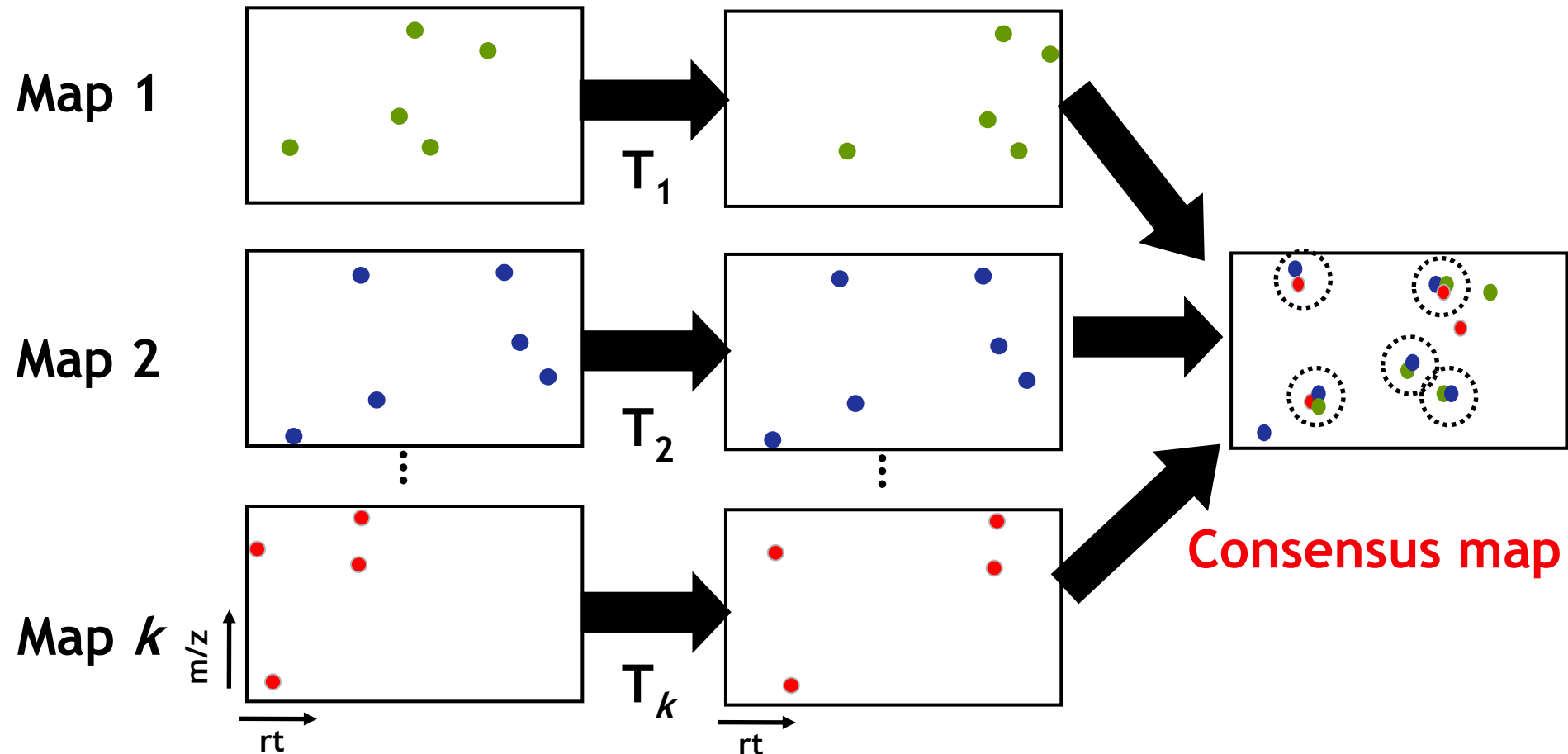
- Identify all peaks belonging to one peptide
- Key idea:
 - Identify suspicious regions
 - Fit a **model** to that region and identify peaks explained by it



Multiple Alignment



- Dewarp k maps onto a comparable coordinate system
- Choose one map (usually the one with the largest number of features) as reference map (here: map 2 $\rightarrow T_2 = 1$)



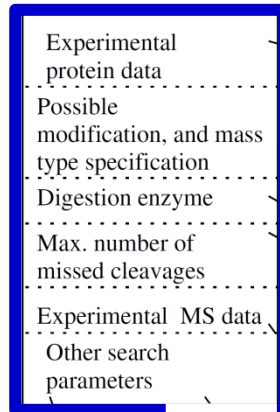


- Identification in proteomics

Identification workflow



Experimental
parameters



Preprocess DB data
and identify relevant
protein sequences

Protein database

Database entry e_j

Filtered sequence
database

Database
sequence d_i

DB settings

Remove contaminants

Experimental mass list

In silico
digestion

Theoretical mass list

Compare
theoretical masses to
experimental masses

List of matched
masses

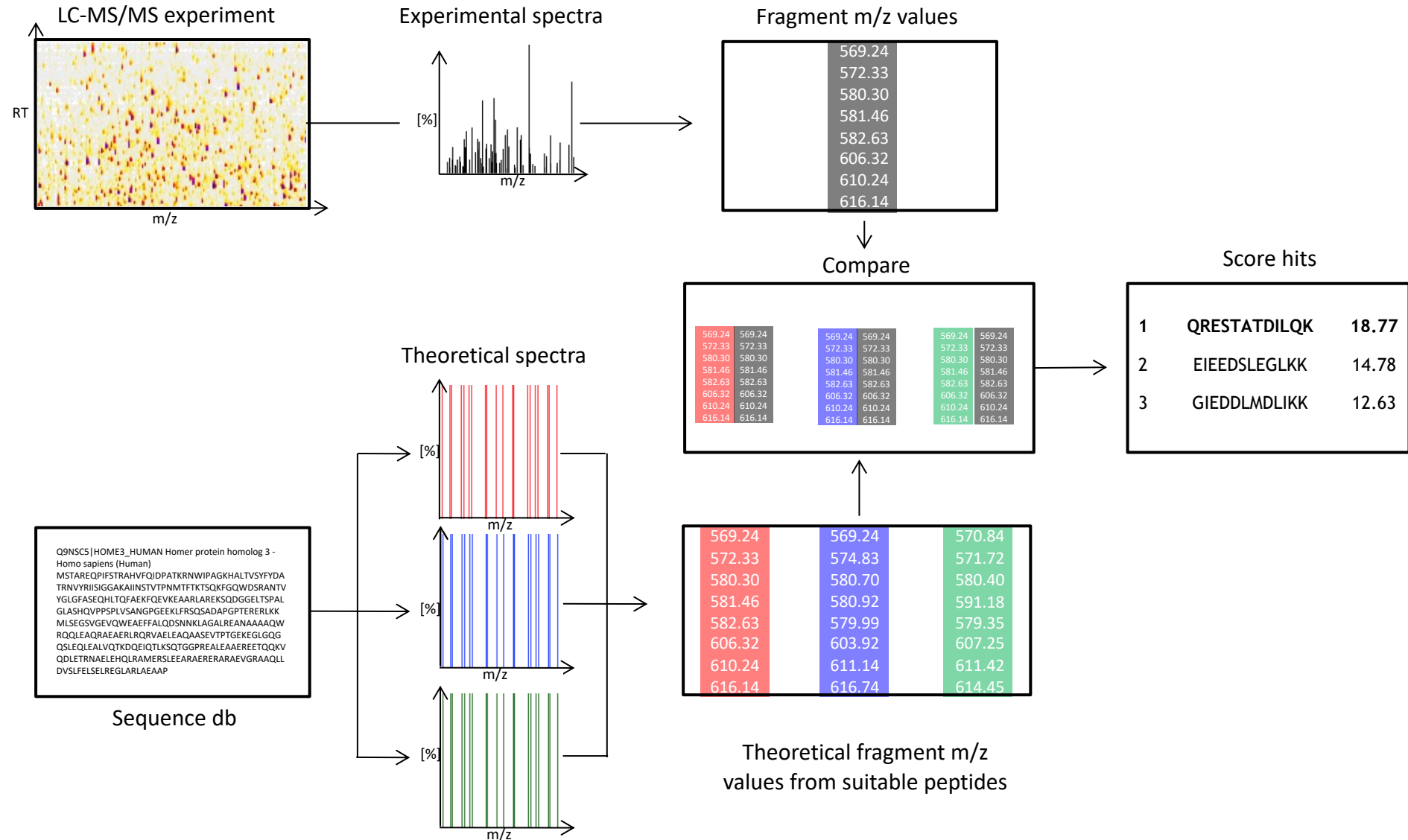
Calculate score
and statistical
significance

Presentation

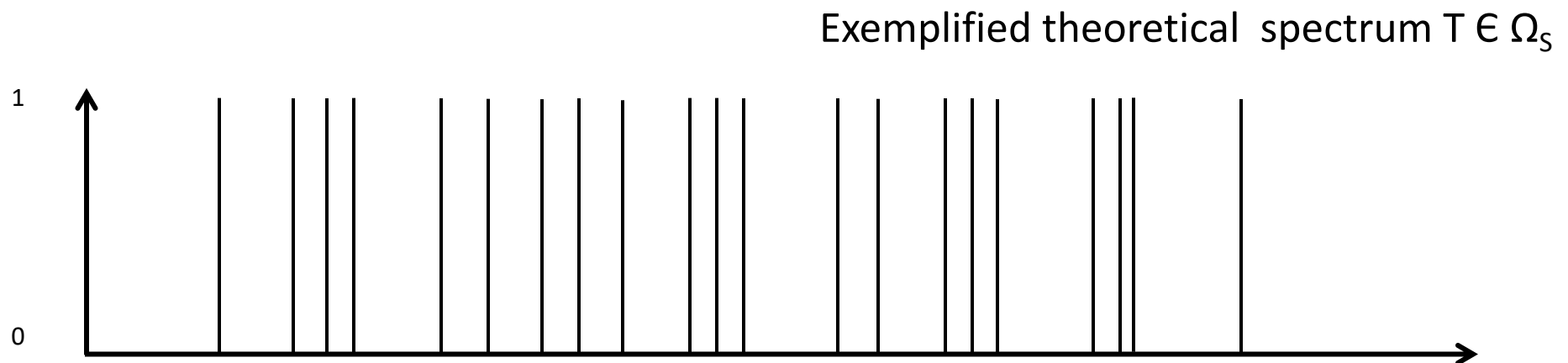
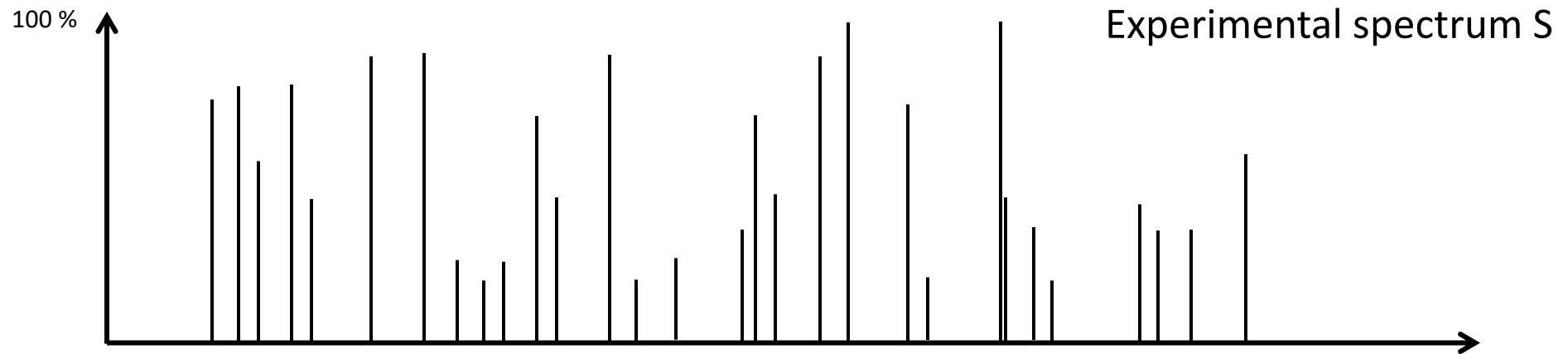
Search engine



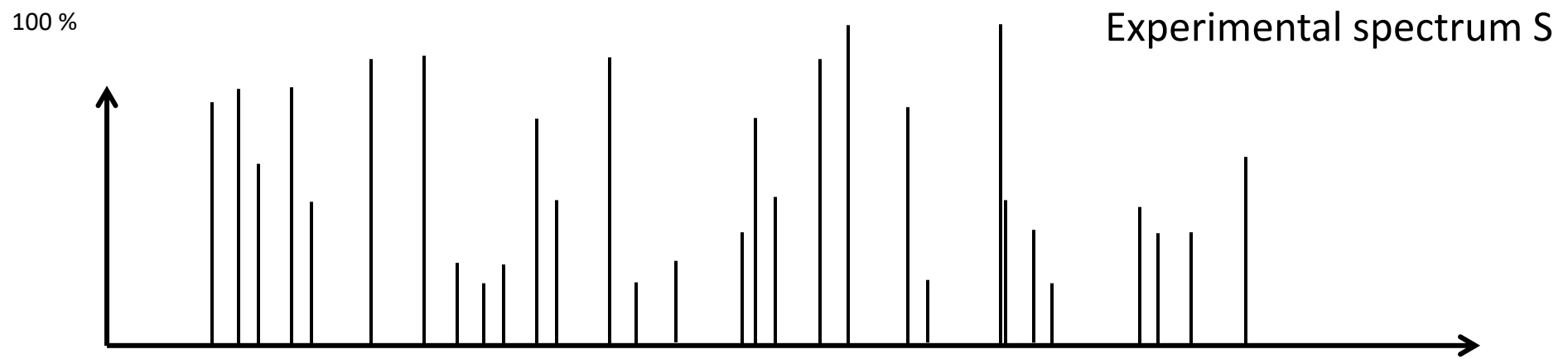
Peptide identification



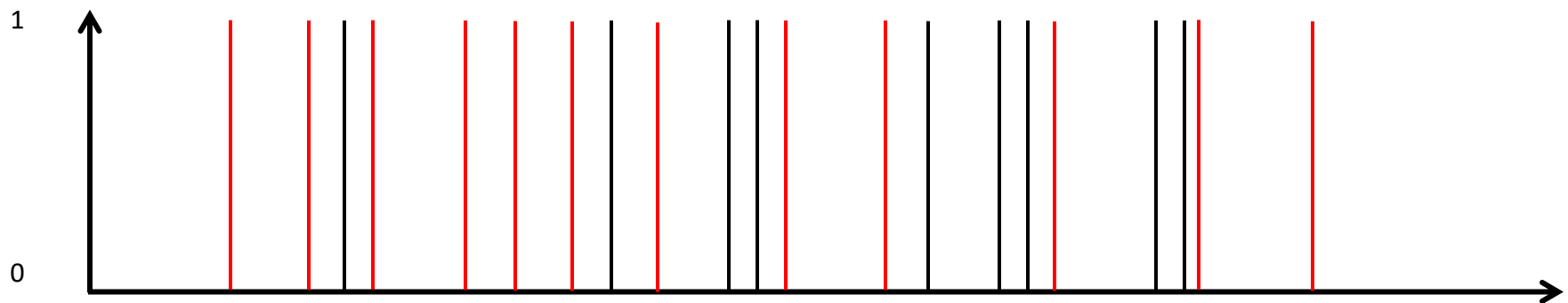
1. From the database, extract all sequences that fit the precursor mass of the MS2 spectrum with a given error tolerance
2. For each of these candidates a theoretical spectrum is generated
3. All theoretical spectra are aligned / compared to the experimental spectrum
4. The alignments are scored and the candidates are ranked according to the score
5. The top ranked candidate is assumed to be the correct PSM (Peptide Spectrum Matching)



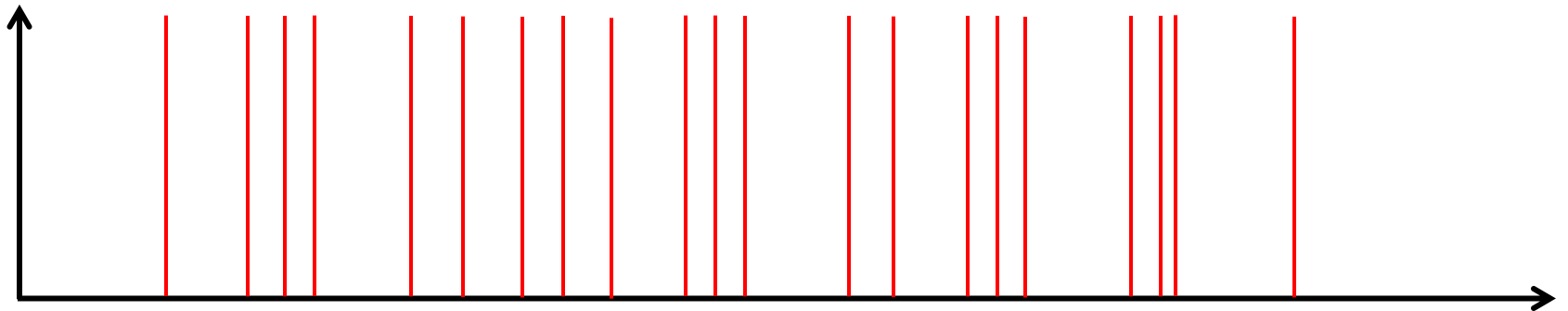
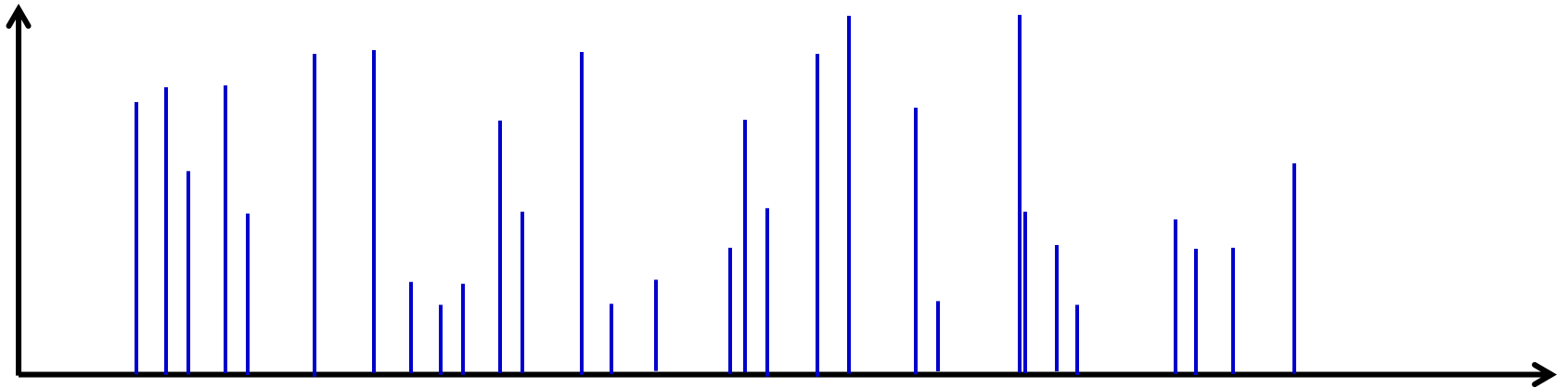
Sequest – Cross correlation



- Sum all the peaks that overlap between theoretical and experimental spectrum
- This score is called **Cross-correlation**



Sequest – Autocorrelation



Sequest – X_{corr} score

- By shifting the spectra, the assumption is that the peaks should not overlap. The spectra are displaced by x Da
- The peaks that overlap upon spectra shifting are used to calculate the autocorrelation

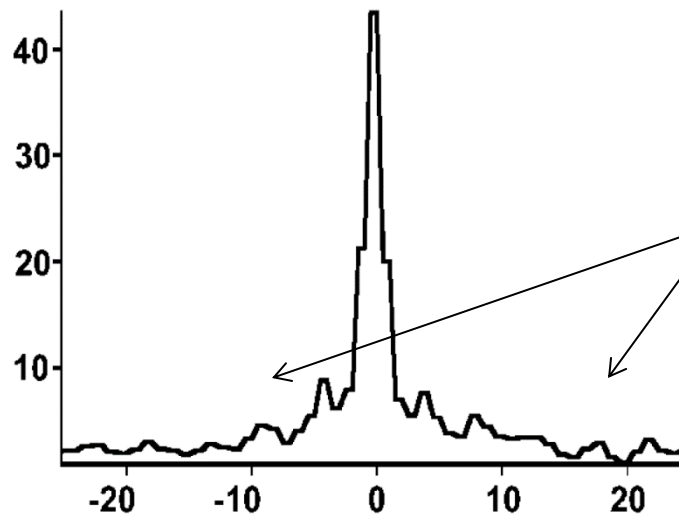
- Sequest reports X_{corr} scores

$$X_{corr} = \frac{Cross_{corr}}{Average(Auto_{corr})_{-75 \leq x \leq 75}}$$

for displacement
 $x [\text{Da}] \in \{-75, 75\}$

correlation count

$\times 10^3$



Displacement $x = 0$,
denotes the cross
correlation

Displacement $x \neq 0$,
denotes the auto-
correlation

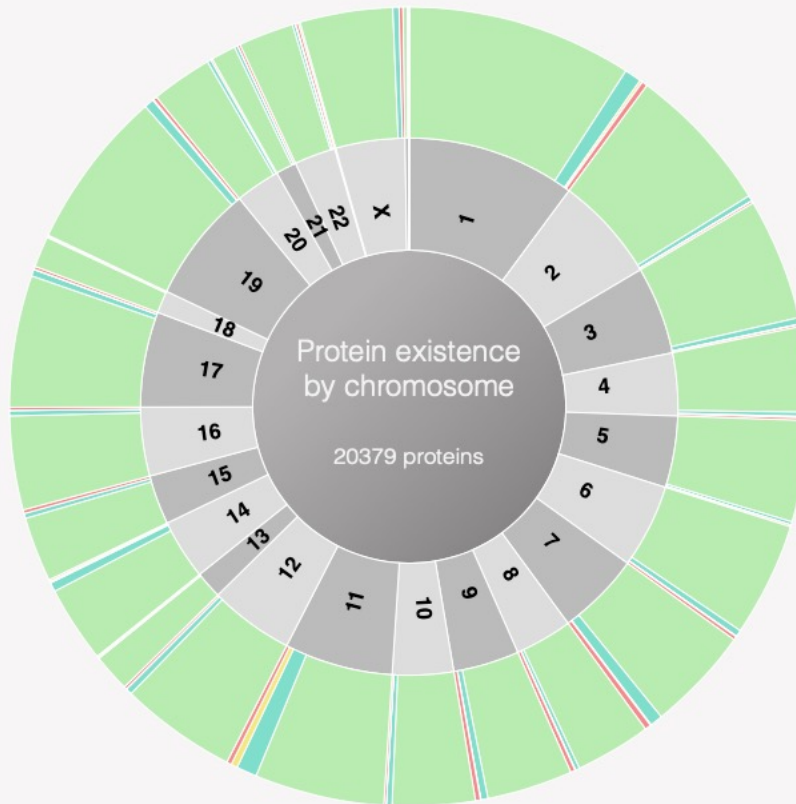
Displacement x in Da

- X_{corr} scores can be calculated for every theoretical spectrum in the search space Ω_S for an experimental spectrum S
- Additionally to the X_{corr} score, Sequest also calculates the ΔC_n score for the top scoring PSM (best X_{corr})
- This score measures how good the best score is in relation to the second best

$$\Delta C_n = \frac{X_{cross1} - X_{cross2}}{X_{cross1}}$$



- If we are interested in proteomics (in contrast to peptide identification in metabolomics, MHC ligandomics etc.), we want to **quantify proteins**
- Non-unique peptide sequences can stem from different proteins
- **Uniqueness**
 - depends on the chosen database
 - becomes more likely for longer peptide sequences
- Reasons for non-uniqueness
 - Chance hits
 - Different isoforms
 - Conserved regions shared within a protein family



PE1 : Evidence at protein level PE2 : Evidence at transcript level PE3 : Inferred from homology PE4 : Predicted PE5 : Uncertain						
Chr	PE1	PE2	PE3	PE4	PE5	Total
1	1859	139	21	2	46	2067
2	1228	45	3	1	18	1295
3	1005	52	5	0	17	1079
4	701	35	11	0	22	769
5	835	26	2	0	14	877
6	940	54	4	1	30	1029
7	863	105	6	1	50	1025
Chr	PE1	PE2	PE3	PE4	PE5	Total
All	18357	1265	147	9	601	20379
	90.08%	6.21%	0.72%	0.04%	2.95%	100%

<https://www.nextprot.org>

Protein

Function
Medical
Expression
Interactions
Localisation
Sequence
Proteomics

Structures

Identifiers

Gene

Exons
Identifiers

References

Curated publications (13)
Additional publications (6)
Patents (0)
Submissions (3)
Web resources (0)

PDE9A » High affinity cGMP-specific 3',5'-cyclic phosphodiesterase 9A [EC 3.1.4.35]

favorite label

Gene name: PDE9A

extend overview

1 22 16

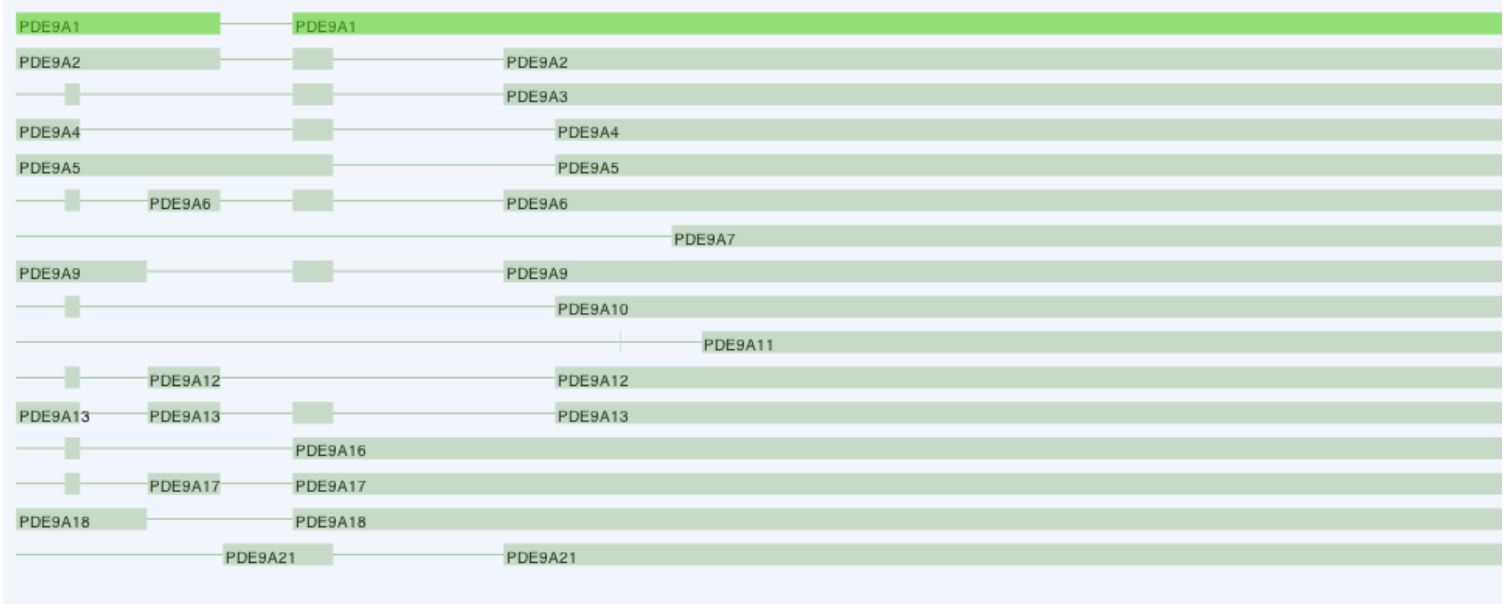
GENE REF ISO

Family name: Cyclic nucleotide phosphodiesterase » PDE9

One or more isoforms of this protein have been shown to exist at protein level

Displayed isoform: PDE9A1

change isoform

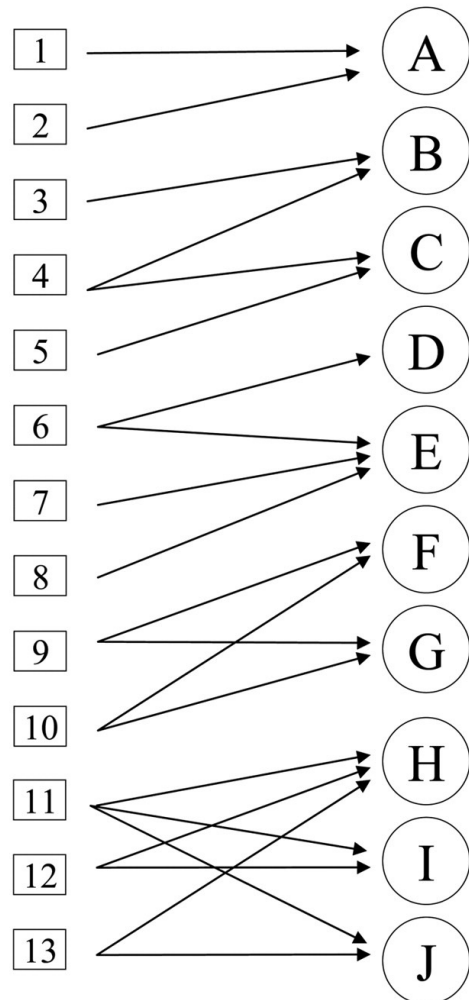


- phosphodiesterase 9A has 16 documented isoforms
- Peptides stemming from the second half of the sequence are entirely indistinguishable between isoforms



peptides

proteins



protein summary list

minimal list of proteins:

1. Protein A
peptides 1, 2
2. Protein B
peptides 3, 4*
3. Protein C
peptides 4*, 5
4. Protein E
peptides 6*, 7, 8
5. Protein F, Protein G
peptides 9*, 10*
6. Protein group:
 - (1) Protein H
peptides 11*, 12*, 13*
 - (2) Protein I
peptides 11*, 12*
 - (3) Protein J
peptides 11*, 13*

“protein” count: 6

no conclusive evidence:

7. Protein D
peptides 6*



Terabytes of *raw data* as starting data



Quantification and identification workflows reduce the bulk raw data to Megabytes of *processed data*



These processed data can then be subject to biological interpretation

Downstream bioinformatics is used to find biomarkers or kinetic models of biological mechanisms

- Multivariate analysis: identifying groups of samples from a dataset with $\# \text{ features} > \# \text{ samples}$
- Cluster analysis: identify similar expression profiles (e.g. in time course data)
- Gene set enrichment analyses: identify functional groups (e.g. oxidative stress) that are differentially regulated
- Mathematical modeling: (partial) differential equations are statistical models can be used to describe the quantitative behavior of the data

- Frequently used methods for interpretation of processed proteomics or metabolomics include multivariate statistical methods or clustering methods

Definition: **Multivariate statistics** is a form of statistics encompassing the simultaneous observation and analysis of more than one outcome variable. The application of multivariate statistics is multivariate analysis.

[en.wikipedia.org/wiki/Multivariate_statistics, access: 224/11/2014, 1 PM]

- Many statistical techniques focus on just one or two variables
- Multivariate analysis (MVA) techniques allow more than two variables to be analysed at once
 - Multiple regression is not typically included under this heading, but can be thought of as a multivariate analysis

Simpson's paradox

- Gender bias among graduate school admissions to the University of Berkley
- Example: 44% of male applicants are admitted by a university, but only 33% of female applicants
- Does this mean there is unfair discrimination?
- To conclude anything the data need to be broken down

	Applicants	Admitted
Men	8442	44%
Women	4321	35%

https://en.wikipedia.org/wiki/Simpson's_paradox#Low_birth_weight_paradox

Source: <https://www.stat.auckland.ac.nz/~balemi/>

Simpson's paradox

- Six out of 85 departments were significantly biased against men, whereas only four were significantly biased against women
- the pooled and corrected data showed a "small but statistically significant bias in favor of women

Department	Men		Women	
	Applicants	Admitted	Applicants	Admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%

Six largest
departments

https://en.wikipedia.org/wiki/Simpson's_paradox#Low_birth_weight_paradox

- **Conclusion:** Detailed and deeper analysis is needed



- Identify underlying dimensions or principal components of a distribution
- Helps understand the joint or common variation among a set of variables
- Probably the most commonly used method of deriving “factors” in factor-based experimental designs

- The first principal component is identified as the vector (or equivalently the linear combination of variables) on which the most data variation can be projected
- The 2nd principal component is a vector perpendicular to the first, chosen so that it contains as much of the remaining variation as possible
- And so on for the 3rd principal component, the 4th, the 5th etc.

Example – Crime Rates by State

Crime Rates per 100,000 Population by State

Obs	State	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
1	Alabama	14.2	25.2	96.8	278.3	1135.5	1881.9	280.7
2	Alaska	10.8	51.6	96.8	284.0	1331.7	3369.8	753.3
3	Arizona	9.5	34.2	138.2	312.3	2346.1	4467.4	439.5
4	Arkansas	8.8	27.6	83.2	203.4	972.6	1862.1	183.4
5	California	11.5	49.4	287.0	358.0	2139.4	3499.8	663.5
...	...	...	...	...	...	...	...	...

The PRINCOMP Procedure

Observations	50
Variables	7

Simple Statistics							
	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
Mean	7.444000000	25.73400000	124.0920000	211.3000000	1291.904000	2671.288000	377.5260000
Std	3.866768941	10.75962995	88.3485672	100.2530492	432.455711	725.908707	193.3944175

Correlation Matrix							
	Murder	Rape	Robbery	Assault	Burglary	Larceny	Auto_Theft
Murder	1.0000	0.6012	0.4837	0.6486	0.3858	0.1019	0.0688
Rape	0.6012	1.0000	0.5919	0.7403	0.7121	0.6140	0.3489
Robbery	0.4837	0.5919	1.0000	0.5571	0.6372	0.4467	0.5907
Assault	0.6486	0.7403	0.5571	1.0000	0.6229	0.4044	0.2758
Burglary	0.3858	0.7121	0.6372	0.6229	1.0000	0.7921	0.5580
Larceny	0.1019	0.6140	0.4467	0.4044	0.7921	1.0000	0.4442
Auto_Theft	0.0688	0.3489	0.5907	0.2758	0.5580	0.4442	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.11495951	2.87623768	0.5879	0.5879
2	1.23872183	0.51290521	0.1770	0.7648
3	0.72581663	0.40938458	0.1037	0.8685
4	0.31643205	0.05845759	0.0452	0.9137
5	0.25797446	0.03593499	0.0369	0.9506
6	0.22203947	0.09798342	0.0317	0.9823
7	0.12405606		0.0177	1.0000

Eigenvectors							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
Murder	0.300279	-.629174	0.178245	-.232114	0.538123	0.259117	0.267593
Rape	0.431759	-.169435	-.244198	0.062216	0.188471	-.773271	-.296485
Robbery	0.396875	0.042247	0.495861	-.557989	-.519977	-.114385	-.003903
Assault	0.396652	-.343528	-.069510	0.629804	-.506651	0.172363	0.191745
Burglary	0.440157	0.203341	-.209895	-.057555	0.101033	0.535987	-.648117
Larceny	0.357360	0.402319	-.539231	-.234890	0.030099	0.039406	0.601690
Auto_Theft	0.295177	0.502421	0.568384	0.419238	0.369753	-.057298	0.147046

- 2-3 components explain 76%-87% of the variance
- First principal component has uniform variable weights, so is a general crime level indicator
- Second principal component appears to contrast violent versus property crimes
- Third component is harder to interpret



- Principle component analysis

What is principal component analysis?

Markus Ringnér

Principal component analysis is often incorporated into genome-wide expression studies, but what is it and how can it be used to explore high-dimensional data?

Nature Computational Biology, 2008



- Requires some basic linear algebra, e.g. eigenvalues, eigenvectors and some statistical terms, e.g. variance, co-variance.
- A mathematical algorithm that reduces the dimensionality of a dataset while remaining most of the variance.

Application:

You compare metabolic profiles of 75 plants that have been growing under treatment A vs 75 control plants that have been grown without any treatment. Find metabolic markers that can be associated with treatment A within a data set of 100.000 metabolites profiled for each of the samples



- Resulting matrix

$$\begin{array}{cccc} k_{1,1} & k_{1,2} & \dots & k_{1,m} \\ k_{2,1} & k_{2,2} & \dots & k_{2,m} \\ \dots & \dots & \dots & \dots \\ k_{n,1} & k_{n,2} & \dots & k_{n,m} \end{array}$$

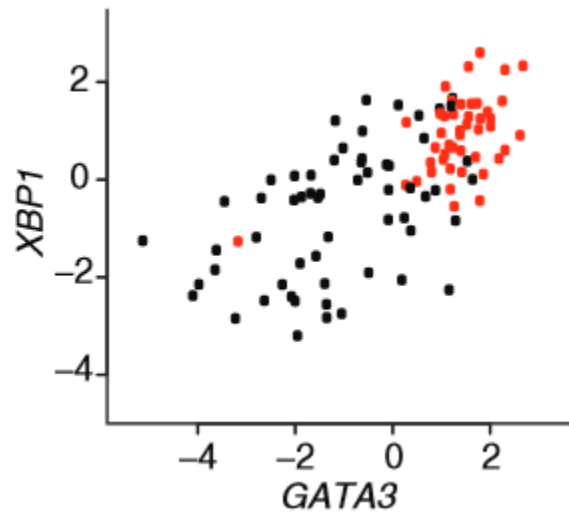
In this case: $m = 150$ samples und $n=100.000$ metabolites

- Manually, this is impossible, but PCA can automatically detect (if present) treatment A-associated markers

Example for multivariate analysis

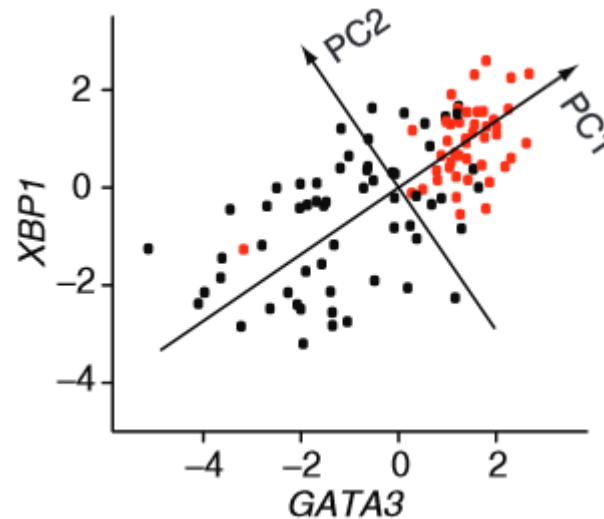


- For simplicity $n=2$ (GATA 3 und XBP1)





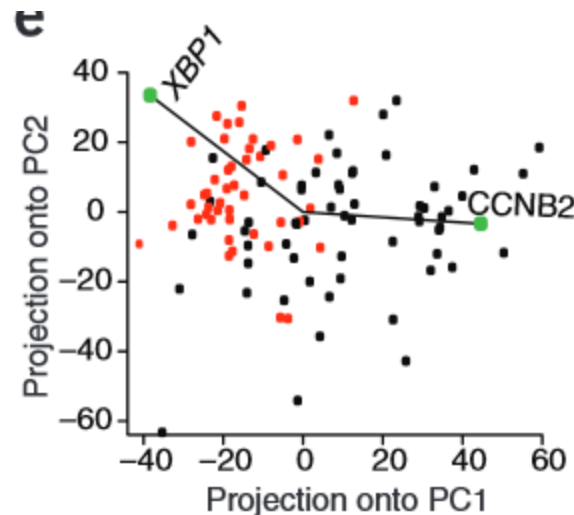
- Find areas of highest variance



- Principle components are linear combinations of the original axis, here: $PC1 = 0.83 \times GATA3 + 0.56 \times XBP1$



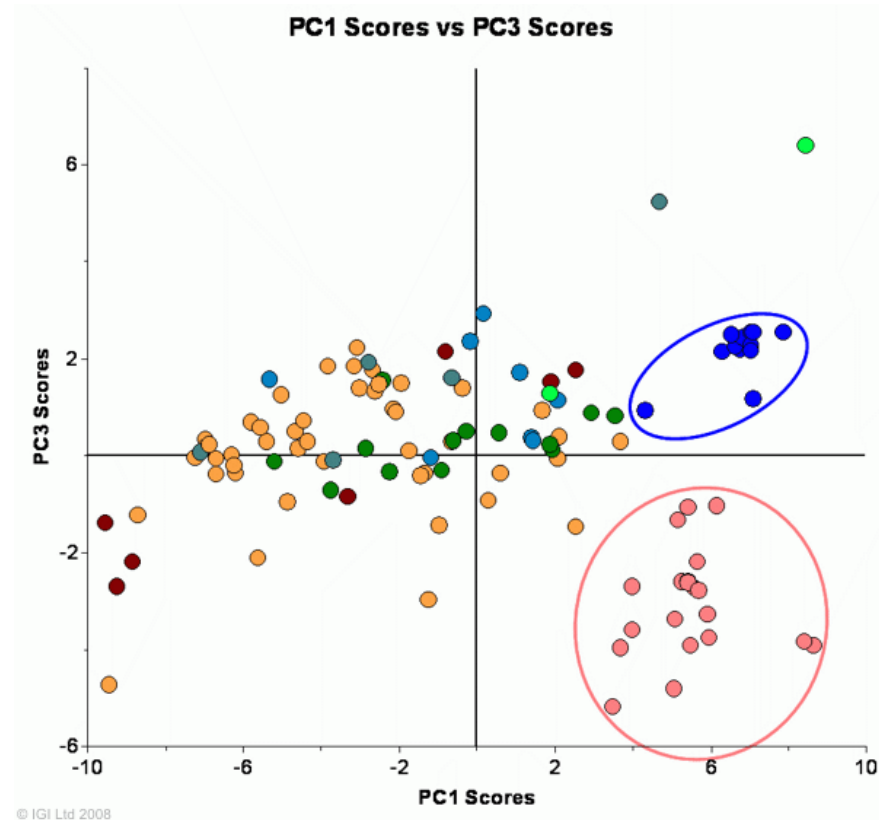
- Project to PC1 and PC2



- Now, do the same for 100.000 dimensions and read out your markers



- PCA can be used to identify groups of samples with similar profiles using the PCA loadings plot
- PCA can be used to identify the molecular markers using the PCA loadings plot (projected original variables)



© IGI Ltd 2008

http://www.igiltd.com/ig.NET%20Sample%20Pages/images/fig_pca_scores_plot.gif

- PCA can be used to identify groups of samples with similar profiles
- Other multivariate methods include:
 - PLS-DA [doi:10.1016/j.trac.2009.08.006](https://doi.org/10.1016/j.trac.2009.08.006)
 - Support vector machines
http://www.cs.columbia.edu/~kathy/cs4701/documents/jason_svm_tutorial.pdf

- Papers cited on the individual slides
- Quantification in general:
 - Bantscheff *et al.*, Quantitative mass spectrometry in proteomics: a critical review, *Anal Bioanal Chem* (2005), 389, 1017-1031 [PMID: 17668192]
- Experimental methods
 - SILAC: Ong, Mann, *Nat Prot* 1 (2007), 2650-2660.
 - iTRAQ: Ross *et al.*, *Mol Cell Prot* (2004), 3, 1154-1169.
- Pose clustering algorithm
 - Lange *et al.*, A geometric approach for the alignment of liquid-chromatography—mass spectrometry data, *Bioinformatics* (2007), 23:i273-i281 [PMID: 17646306]