SCIENCE MEETS LIFE

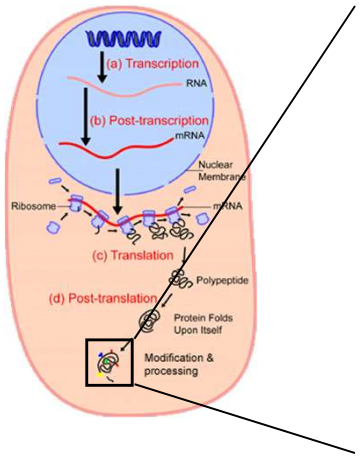
MS-BASED PROTEOMICS DATA ANALYSIS

lennart martens

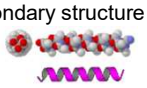
lennart.martens@vib-ugent.be
@compomics
computational omics and systems biology group
Ghent University and VIB, Ghent, Belgium

VIB GHENT UNIVERSITY Comp omics

Proteomics in the central paradigm of biology



- Primary structure (*sequence*)
  …YSFVATAER…
- Secondary structure (*structural elements*)



- Tertiary structure (*3D shape*)

- Modifications (*dynamic, function*)
  phosphorylation
- Processing (*targetting*, *activation*)
  trypsin
  platelet activity

Adapted from the NCBI Science Primer
http://www.ncbi.nih.gov/About/primer/genetics_cell.html

Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

Protein inference: bad, ugly, and not so good

Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

Database search algorithms in three phases

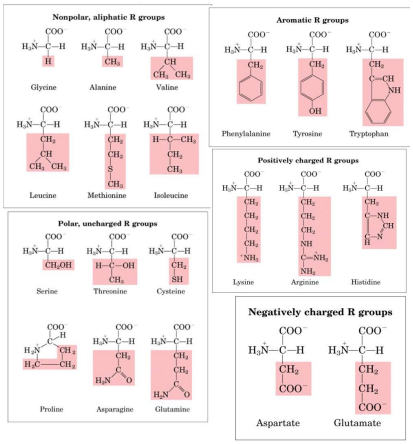
Sequencial search algorithms

Decoys and false discovery rate calculation

Protein inference: bad, ugly, and not so good

CC BY-SA 4.0

Amino acids vary considerably in their physico-chemical properties

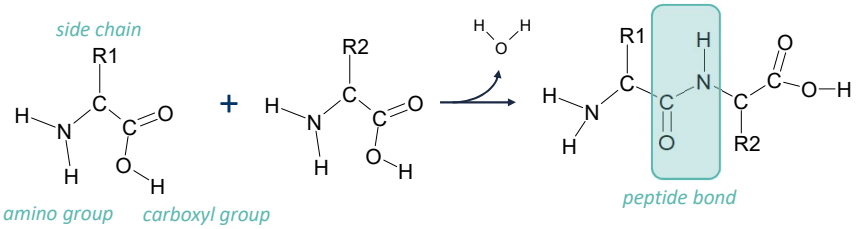


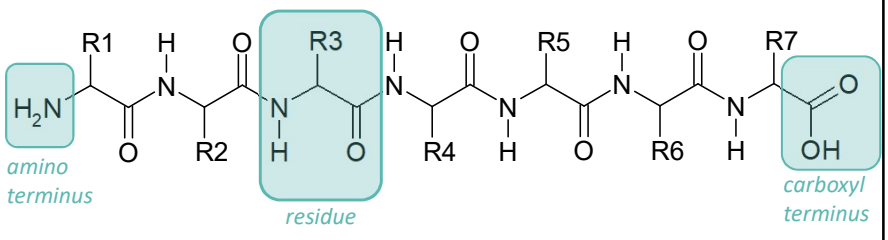
Name	3-Letter Symbol	1-Letter Symbol	Molecular weight	Molecular formula	Residue formula	Residue Weight (H_2O)	pK_a	pK_b	pK_x	pI
Alanine	Ala	A	89.10	$C_3H_7NO_2$	C_3H_5NO	71.08	2.34	9.69	—	6.00
Arginine	Arg	R	174.20	$C_6H_{14}N_4O_2$	$C_6H_{12}N_4O$	156.19	2.17	9.04	12.48	10.76
Asparagine	Asn	N	132.12	$C_4H_8N_2O_3$	$C_4H_6N_2O_2$	114.11	2.02	8.80	—	5.41
Aspartic acid	Asp	D	133.11	$C_4H_7NO_4$	$C_4H_5NO_3$	115.09	1.88	9.60	3.65	2.77
Cysteine	Cys	C	121.16	$C_3H_7NO_2S$	C_3H_5NOS	103.15	1.96	10.28	8.18	5.07
Glutamic acid	Glu	E	147.13	$C_5H_9NO_4$	$C_5H_7NO_3$	129.12	2.19	9.67	4.25	3.22
Glutamine	Gln	Q	146.15	$C_5H_{10}N_2O_3$	$C_5H_8N_2O_2$	128.13	2.17	9.13	—	5.65
Glycine	Gly	G	75.07	$C_2H_5O_2$	C_2H_3O	57.05	2.34	9.60	—	5.97
Histidine	His	H	155.16	$C_6H_9N_3O_2$	$C_6H_7N_3O$	137.14	1.82	9.17	6.00	7.59
Hydroxyproline	Hyp	O	131.13	$C_5H_9NO_3$	$C_5H_7NO_2$	113.11	1.82	9.65	—	—
Isoleucine	Ile	I	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	—	6.02
Leucine	Leu	L	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	—	5.98
Lysine	Lys	K	146.19	$C_6H_{14}N_2O_2$	$C_6H_{12}N_2O$	128.18	2.18	8.95	10.53	9.74
Methionine	Met	M	149.21	$C_5H_{11}NO_2S$	C_5H_9NOS	131.20	2.28	9.21	—	5.74
Phenylalanine	Phe	F	165.19	$C_9H_{11}NO_2$	C_9H_9NO	147.18	1.83	9.13	—	5.48
Proline	Pro	P	115.13	$C_5H_9NO_2$	C_5H_7NO	97.12	1.99	10.60	—	6.30
Pyroglutamatic	Glp	U	139.11	$C_5H_7NO_3$	$C_5H_5NO_2$	121.09	—	—	—	5.68
Serine	Ser	S	105.09	$C_3H_7NO_3$	$C_3H_5NO_2$	87.08	2.21	9.15	—	5.68
Threonine	Thr	T	119.12	$C_4H_9NO_3$	$C_4H_7NO_2$	101.11	2.09	9.10	—	5.60
Tryptophan	Trp	W	204.23	$C_{11}H_{12}N_2O_2$	$C_{11}H_{10}N_2O$	186.22	2.83	9.39	—	5.89
Tyrosine	Tyr	Y	181.19	$C_9H_{11}NO_3$	$C_9H_9NO_2$	163.18	2.20	9.11	10.07	5.66
Valine	Val	V	117.15	$C_5H_{11}NO_2$	C_5H_9NO	99.13	2.32	9.62	—	5.96

[1] pK_a is the negative of the logarithm of the dissociation constant for the -COOH group
[2] pK_b is the negative of the logarithm of the dissociation constant for the $-NH_3^+$ group
[3] pK_x is the negative of the logarithm of the dissociation constant for any other group in the molecule
[4] pI is the pH at the isoelectric point
References: D. R. Lide, *Handbook of Chemistry and Physics*, 72nd Edition, CRC Press, Boca Raton, FL, 1991.

http://courses.cm.utexas.edu/jrobertus/ch339k/overheads-1/ch5-amino-acids.jpg

Protein backbones are formed through amide (or peptide) bonds between residues





Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

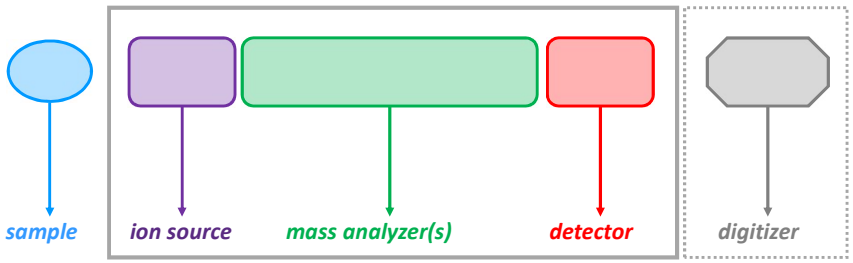
Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

Protein inference: bad, ugly, and not so good

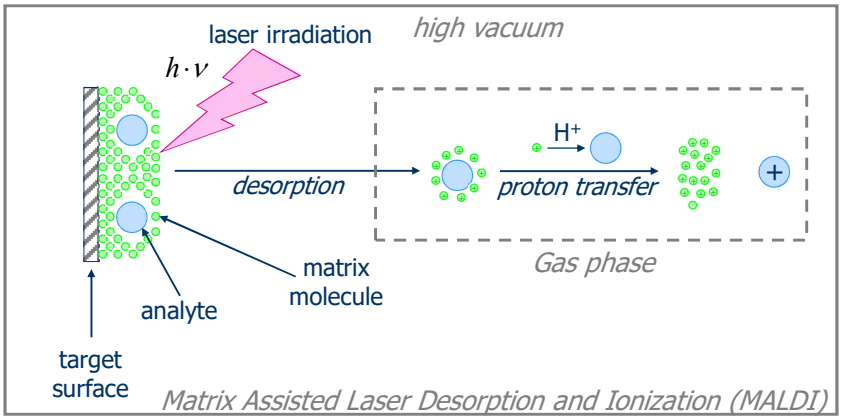
A generalized mass spectrometer consists of three main parts, along with a digitizer



Generalized mass spectrometer

All **mass analyzers** use electromagnetic fields to manipulate gas-phase ions. Results are plotted as a spectrum, with mass-over-charge (*m/z*) on the X-axis and ion intensity on the Y-axis. The latter can be absolute (counts) or relative. The **ion source** ensures that (a part of) the sample molecules are ionized and brought into the gas phase. The **detector** is responsible for actually recording the presence of ions. **Digitizers** (analog to digital converters; ADC) transform the continuous, analog detector signal into a digital, discretized spectrum.

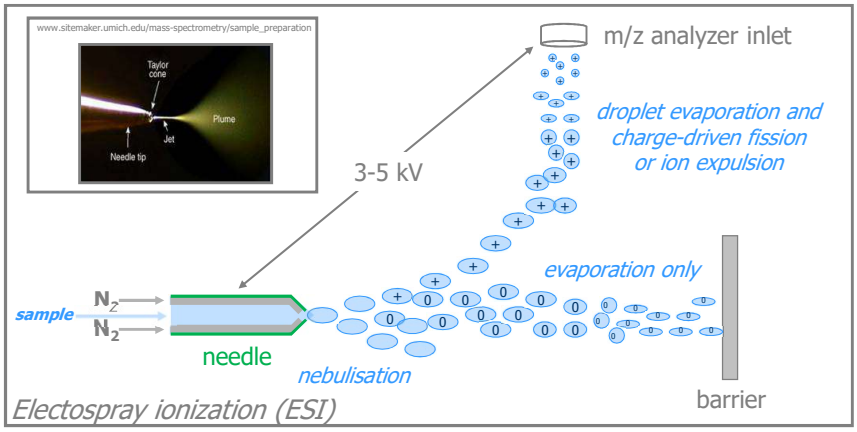
Ion sources: MALDI



MALDI sources for proteomics typically rely on a pulsed nitrogen UV laser (υ = 337 nm) and produce singly charged peptide ions. Competitive ionisation occurs.

The term 'MALDI' was coined by **Karas and Hillenkamp** (*Anal. Chem.*, 1985) and **Koichi Tanaka** received the 2002 Nobel Prize in Chemistry for demonstrating MALDI ionization of biological macromolecules (*Rapid Commun. Mass Spectrom.*, 1988)

Ion sources: ESI

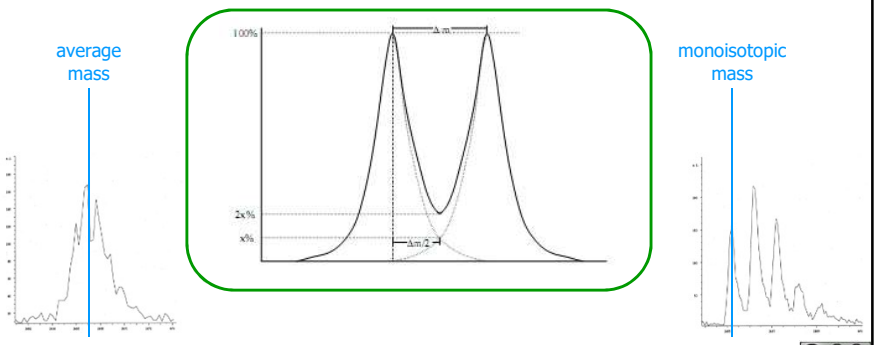


ESI sources typically heat the needle to 40° to 100° to facilitate nebulisation and evaporation, and typically produce multiply charged peptide ions (2^+, 3^+, 4^+)

John B. Fenn received the 2002 Nobel Prize in Chemistry for demonstrating ESI ionization of biological macromolecules (*Science*, 1989) – ESI is also used in fine control thrusters on satellites and interstellar probes…

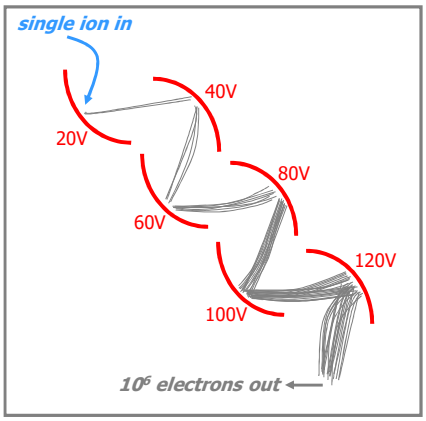
Mass resolution is an important characteristic for identification and quantification

Resolution in mass spectrometry is usually defined as the width of a peak at a given height (there is an alternative definition based on percent valley height). This width can be recorded at different heights, but is most often recorded at 50% peak height (FWHM).



From: Eidhammer, Flikka, Martens, Mikalsen – Wiley 2007

Detectors: electron multiplier amplification



Different variations of electron multiplier (EM) detectors are used, and these are the most common type of detector. An EM relies on several Faraday cup dynodes with increasing charges to produce an electron cascade from a few incident ions.

The primary principle in quantification is that detector signal relates to quantity

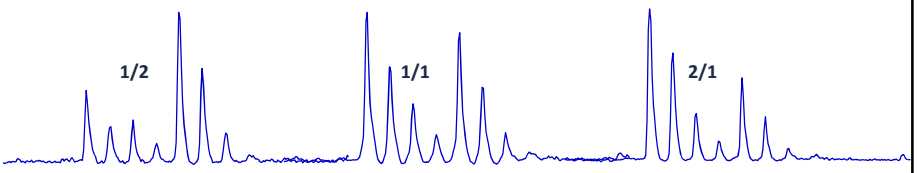
Make each sample distinguishable

introduce mass differences between the samples

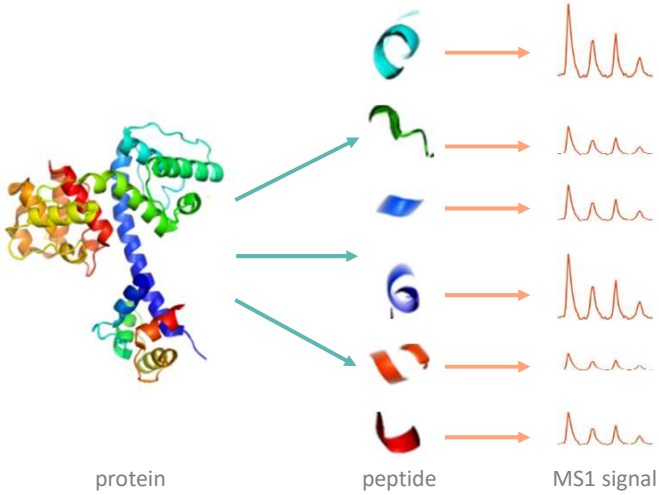
perform distinct experimental runs for each sample

Measure the intensity of the signal for each analyte in each sample

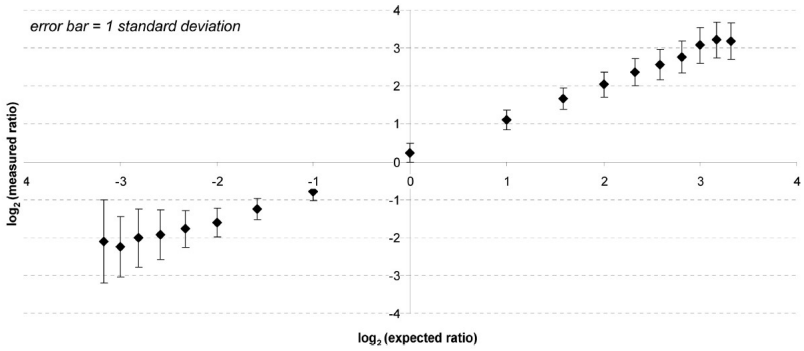
Statistically process the accumulated information



Not all peptides ionise equally, so we cannot compare signal strength across peptides

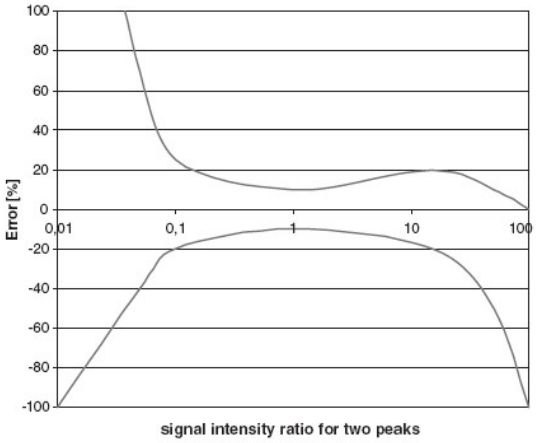


As intensities become more extreme, the detector response starts to level off



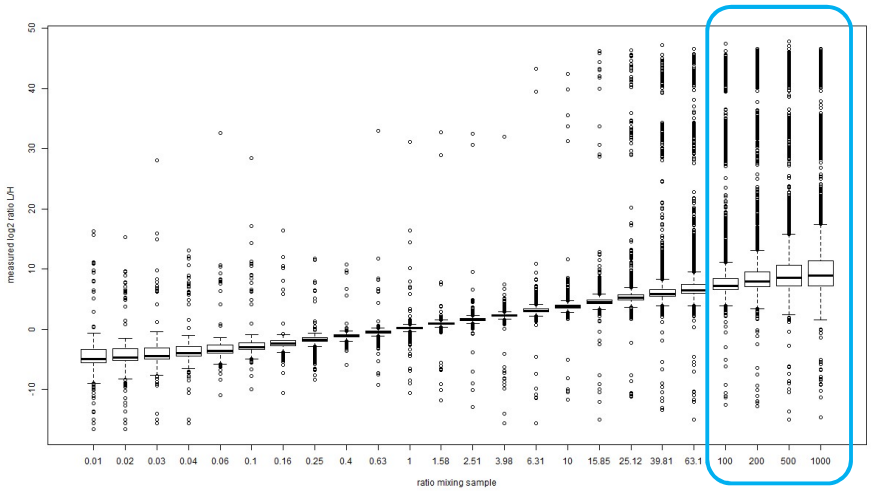
Gevaert, Proteomics, 2007

At the same time, the measurement error increases as the ratio deviates from 1/1



Vaudel, Proteomics, 2010

And these effects remain quite visible, even on modern instruments (Orbitrap)

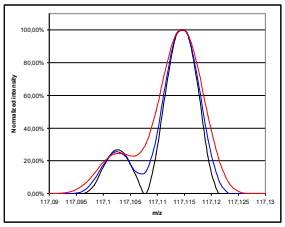


Raw data processing is somewhat imprecise, with expected errors on the order of 10%

Mass spectrometer specific processing required

Sets the dynamic range lower limit (S/N)

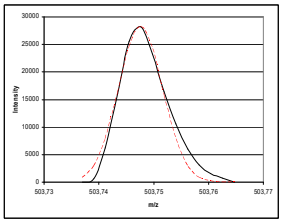
5-10% error in the final ratios due to peak-picker are often seen



Black: 0,02 Da

Blue: 0,04 Da

Red: 0,08 Da

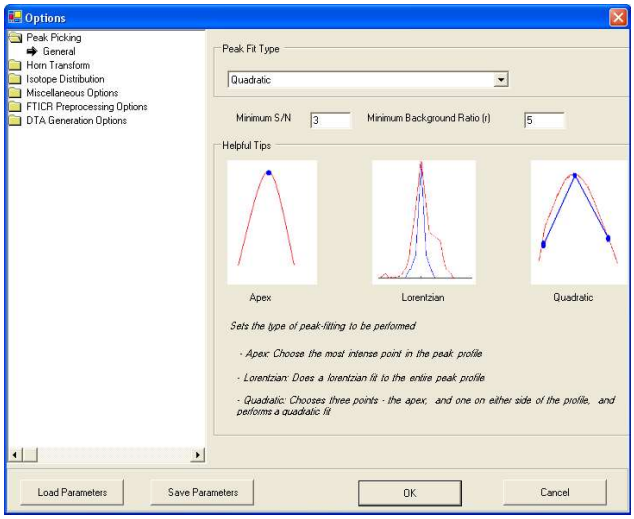


Non-adapted shape -> +10% error

Vaudel, Proteomics, 2010

Different model options are available in tools or libraries for MS peak detection

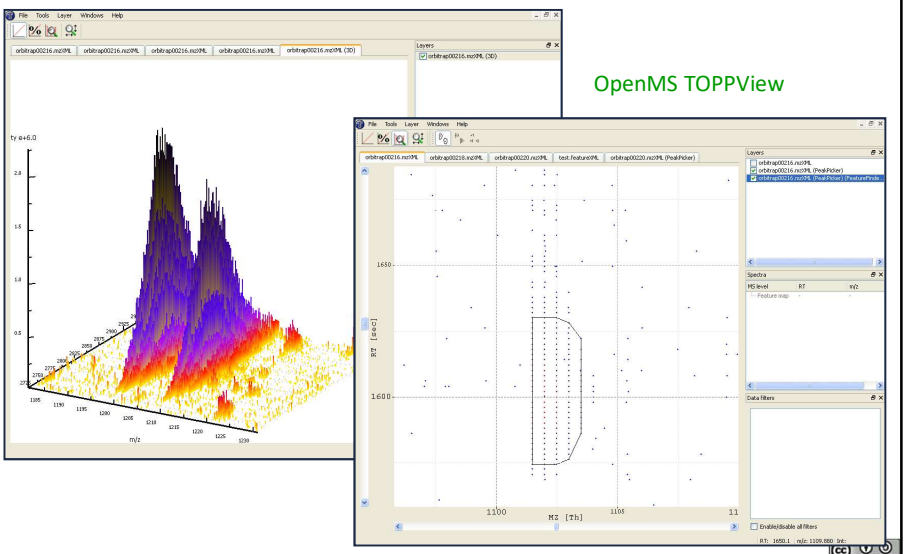
Decon2LS



Vaudel, Proteomics, 2010

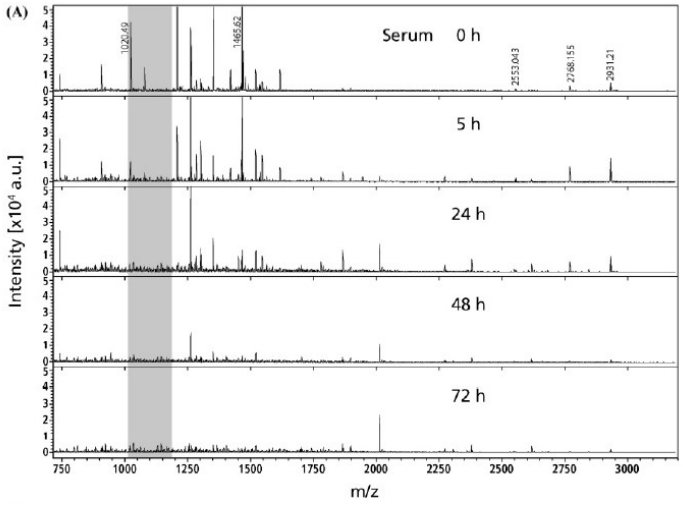
There's actually more to a peak than just m/z

OpenMS TOPPView



Vaudel, Proteomics, 2010

Serum proteins are degraded over time, even with the best sampling tubes



Yi, J Prot Res, 2007

Our open modification search engine ionbot shows that modifications are also an issue

Protein name	Protein accession	Number of modifications
Glyceraldehyde-3-phosphate dehydrogenase	P04406	166
Pyruvate kinase PKM	P14618	139
Fructose-bisphosphate aldolase A	P04075	122
Alpha-enolase	P06733	121
Triosephosphate isomerase	P60174	117
Phosphoglycerate kinase	P00558	111

Mods found across all six proteins, between 50 and 278 distinct peptides

carbamyl, carbamidomethyl, formyl, acetyl, oxidation, methyl, thiazolidine, amidine, dehydrated, dicarbamidomethyl, dioxidation, succinyl, ammonia-loss, ethyl, carboxymethyl, guanidinyl, gg, cation:fe[iii]

https://ionbot.cloud
Source data presented to ionbot from Kim *et al.*, Nature, 2014

Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

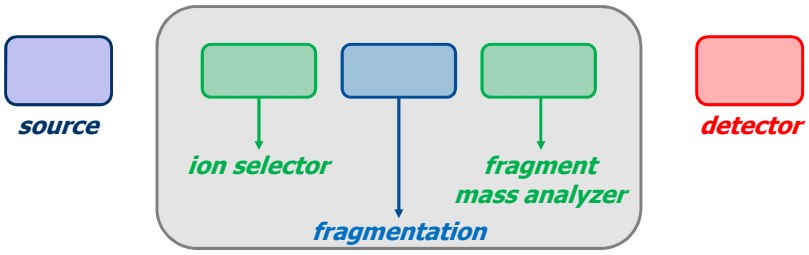
Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

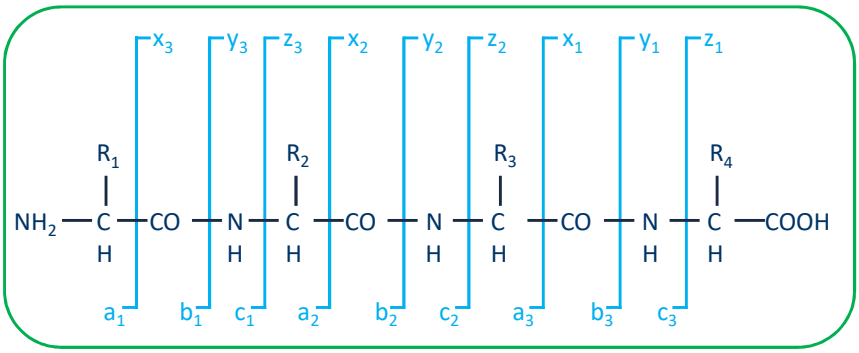
Protein inference: bad, ugly, and not so good

Identification relies on fragmentation



Tandem-MS is accomplished by using two mass analyzers in series (tandem). A single ion trap can also perform tandem-MS. The first mass analyser performs the function of ion selector, by selectively allowing only ions of a given *m/z* to pass through. The second mass analyzer is situated after fragmentation is triggered (see next slides) and is used in its normal capacity as a mass analyzer for the fragments.

Peptides subjected to fragmentation analysis can yield several types of fragment ions

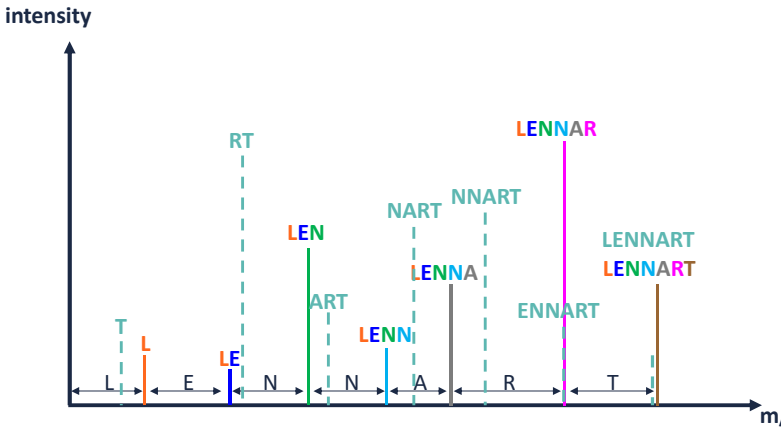


There are several other ion types that can be annotated, as well as 'internal fragments'. The latter are fragments that no longer contain an intact terminus. These are harder to use for 'ladder sequencing', but can still be interpreted.

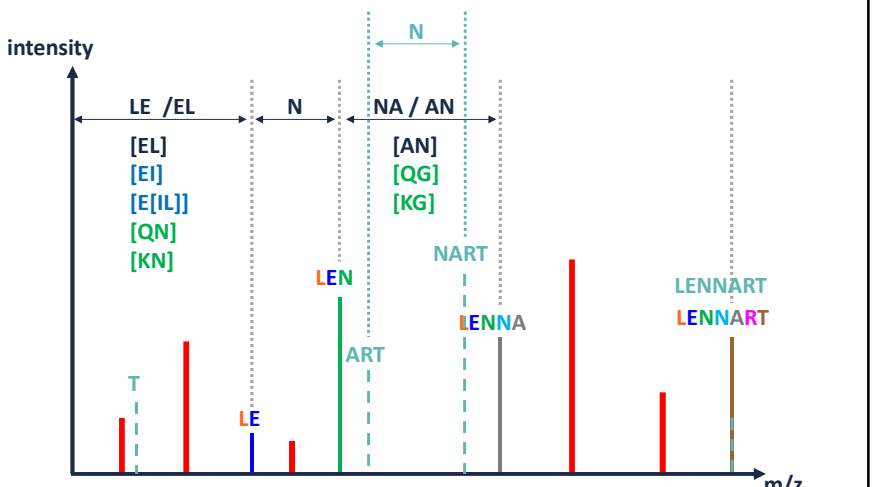
This nomenclature was coined by **Roepstorff and Fohlmann** *(Biomed. Mass Spec.*, 1984) and **Klaus Biemann** *(Biomed. Environ. Mass Spec.*, 1988) and is commonly referred to as 'Biemann nomenclature'. Note the link with the Roman alphabet.

In an ideal world, the peptide sequence will produce directly interpretable ion ladders

L E N N A R T



Real spectra usually look quite a bit worse, which introduces ambiguity in interpretation



Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

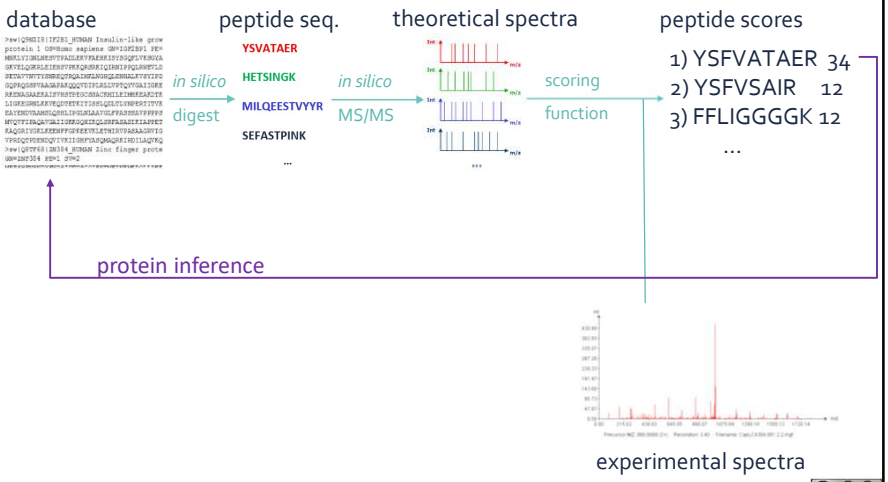
Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

Protein inference: bad, ugly, and not so good

Database search engines match experimental spectra to known peptide sequences



Three popular algorithms illustrate the three types of scoring systems

SEQUEST (UWashington, Thermo Fisher Scientific)
Intensity-based scoring system

MASCOT (Matrix Science) / Andromeda (Jürgen Cox)
Peak counting-based scoring system

X!Tandem (The Global Proteome Machine Organization)
Hybrid scoring system

SEQUEST is the original search engine, and is based on ion intensity matching

Can be used for MS/MS (PFF) identifications

Based on a cross-correlation score (includes peak height)

Published core algorithm (patented, licensed to Thermo), Eng, *JASMS* 1994

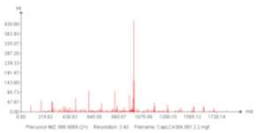
Provides preliminary (Sp) score, rank, cross-correlation score (XCorr),
and score difference between the top tow ranks (deltaCn, ΔCn)

Thresholding is up to the user, and is commonly done *per* charge state

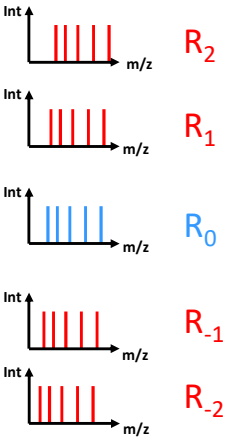
Many extensions exist to perform a more automatic validation of results

The correlation score (R_i) is calculated as the matched ion intensity

$$\sum_{i=-7}^{+75} Ri$$



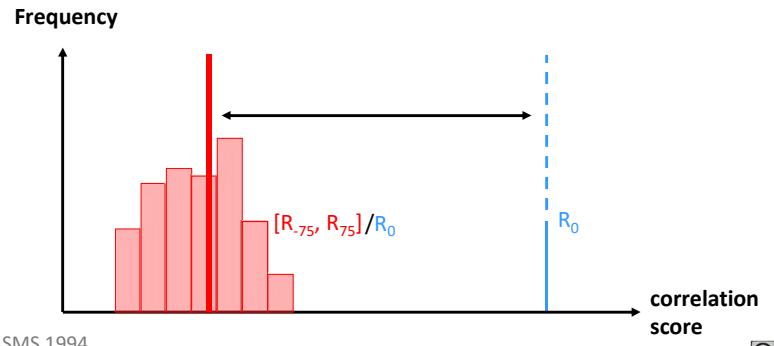
$$R_i = \sum_{j=1}^{n} x_j \cdot y_{(j+i)}$$



Eng, JASMS 1994
Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

The cross-correlation score (*Xcorr*) is R_0 calibrated by the average random correlation

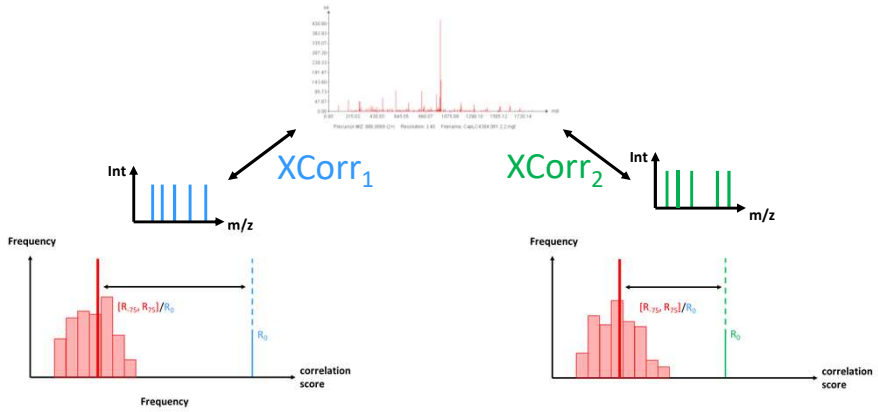
$$\text{XCorr} = R_0 - \frac{1}{150}\left(\sum_{i=-7}^{+75} Ri\right)_{/R_0}$$



Eng, JASMS 1994
Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

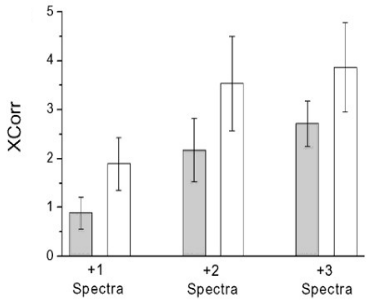
The best theoretical match is then compared to the second-best theoretical match

$$\text{deltaCn} = \frac{XCorr_1 - XCorr_2}{XCorr_1}$$

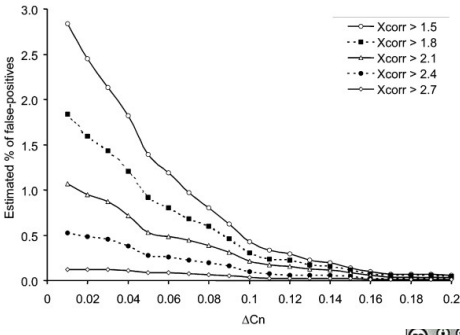


Eng, JASMS 1994
Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

But the advent of high-throughput proteomics showed issues with user-defined thresholding



MacCoss et al., Anal. Chem. 2002



Peng et al., J. Prot. Res.. 2002

Mascot is an equally recognized search engine, but is based on peak counting

Very well established search engine, Perkins, *Electrophoresis* 1999

Can do MS (PMF) and MS/MS (PFF) identifications

Based on the MOWSE score,

Unpublished core algorithm (trade secret)

Predicts an *a priori* threshold score that identifications need to pass

From version 2.2, Mascot allows integrated decoy searches

Provides rank, score, threshold and expectation value per identification

Customizable confidence level for the threshold score

Through Andromeda, we understand MASCOT

$$s = -10 \times \log_{10} \sum_{j=k}^{n} \left[\binom{n}{j} (p)^{j} (1-p)^{n-j} \right]$$

n = number of theoretical peaks
k = number of matched peaks (within a given fragment tolerance)
p = probability of finding a single, matched peak by chance
p is calculated by dividing the number of highest intensity peaks (q) by a mass-window size (100 Da)
q is limited by a maximum value, and is optimized for maximum *s*

based on **peak counting** instead of intensity sums

Cox, J Prot Res, 2011
Yılmaz, Proteome Bioinformatics (MMB), Springer, 2017

X!Tandem introduces a hybrid score, based on both peak counting and ion intensity

A successful open source search engine, Craig and Beavis, *RCMS* 2003

Can be used for MS/MS (PFF) identifications

Based on a hyperscore (*Pi* is either 0 or 1): $HyperScore = \left(\sum_{i=0}^{n} I_i * P_i \right) * N_b! * N_y!$

Relies on a hypergeometric distribution (hence hyperscore)

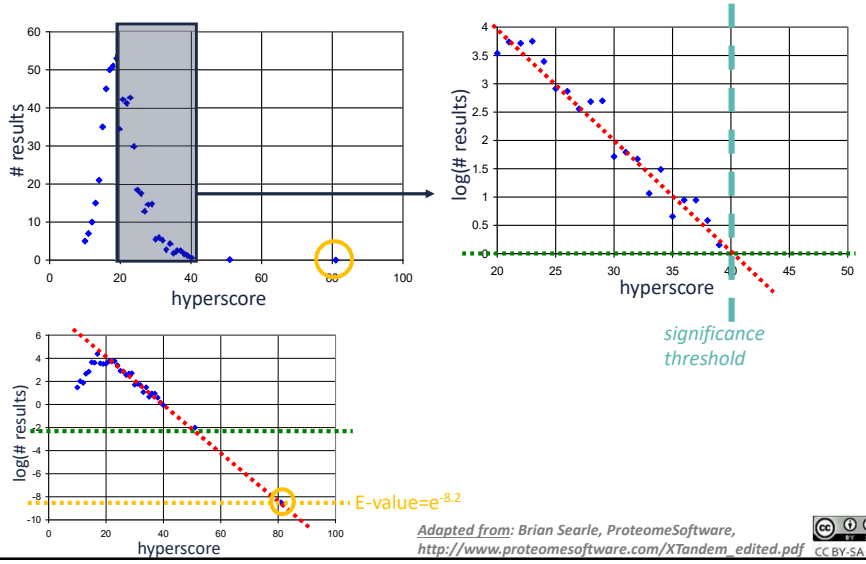
Published core algorithm, and is freely available

Provides hyperscore and expectancy score (the discriminating one)

X!Tandem is fast and can handle modifications in an iterative fashion

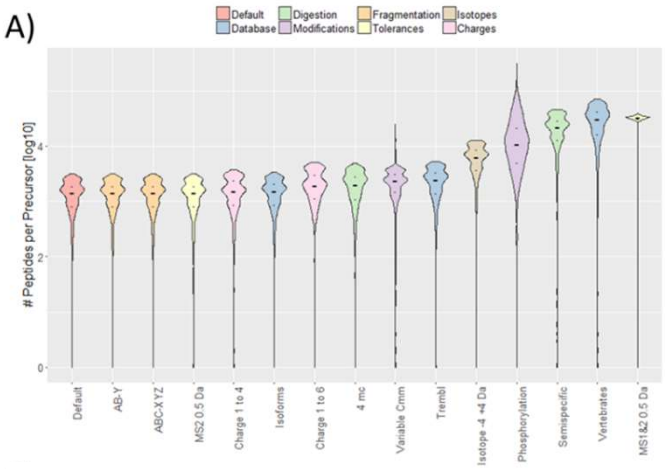
Has rapidly gained popularity as (auxiliary) search engine

X!Tandem's significance calculation for scores can be seen as a general template



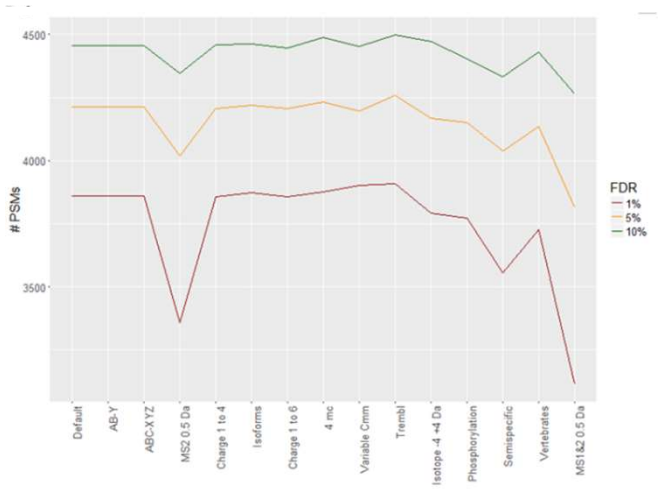
Adapted from: Brian Searle, ProteomeSoftware, http://www.proteomesoftware.com/XTandem_edited.pdf

The influence of various parameter changes on database size is clearly visible



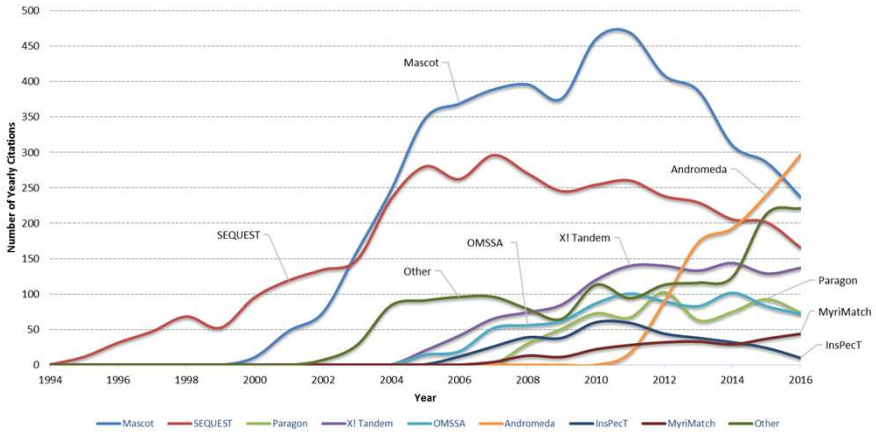
Verheggen, Mass Spec Reviews, 2017

And the effect on identification rate is correspondingly obvious



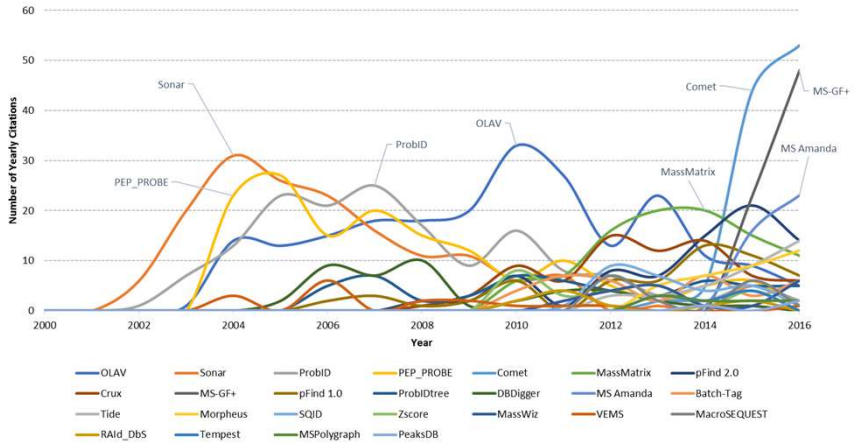
Verheggen, Mass Spec Reviews, 2017

The main search engines in use are Mascot, Andromeda, SEQUEST and X!Tandem



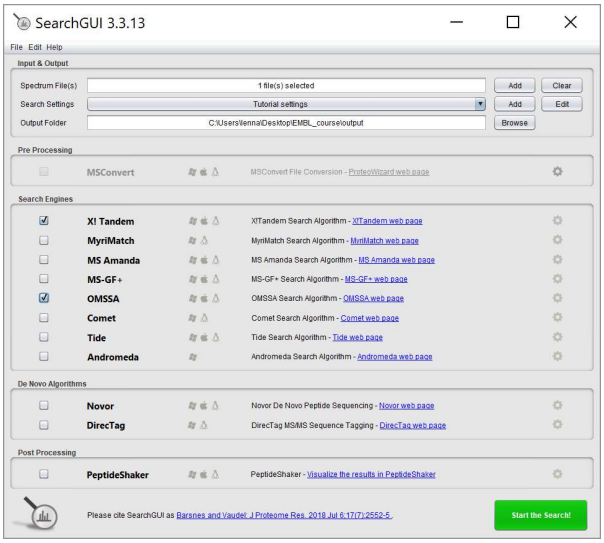
Verheggen, Mass Spec Reviews, 2017

Among the up-and-coming engines, Comet, MS-GF+ and MS-Amanda are most notable



Verheggen, Mass Spec Reviews, 2017

SearchGUI makes it very easy for you to run multiple free search engines



Vaudel, Proteomics, 2011

PeptideShaker is your gateway to the results

Vaudel, Nature Biotechnology, 2015

Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

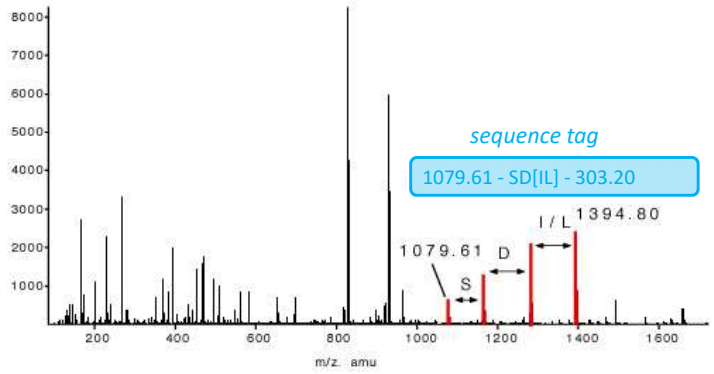
Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

Protein inference: bad, ugly, and not so good

Sequence tags are as old as SEQUEST, and still have a role to play today



The concept of sequence tags was introduced by Mann and Wilm

Mann, Analytical Chemistry, 1994

GutenTag, DirecTag, TagRecon

Tabb, *Anal. Chem.* 2003, Tabb, *JPR* 2008, Dasari, *JPR* 2010

Recent implementations of the sequence tag approach

Refine hits by peak mapping in a second stage to resolve ambiguities

Rely on a empirical fragmentation model

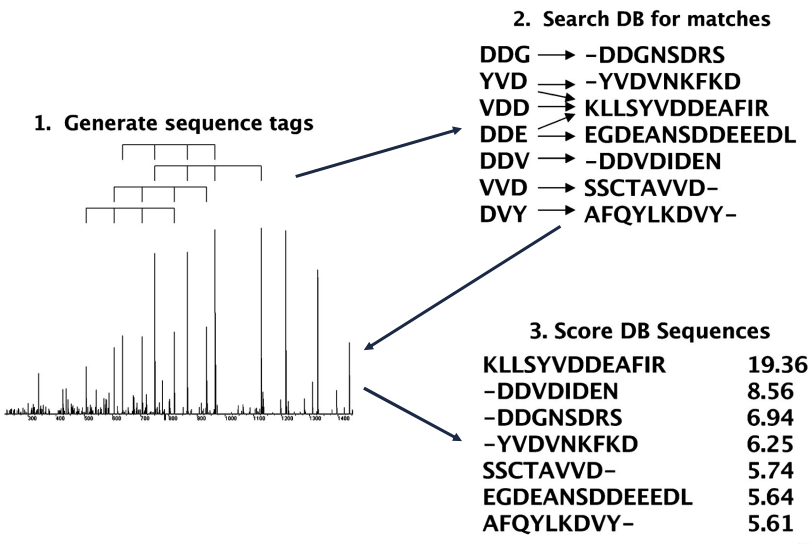
Published core algorithms, DirecTag and TagRecon freely available

GutenTag/DirecTag extracts tags, TagRecon matches tags to database

Very useful to retrieve unexpected peptides (modifications, variations)

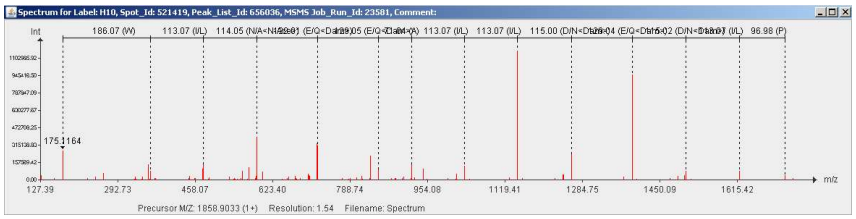
Entire workflows exist (e.g., combination with IDPicker)

GutenTag: two stage, hybrid tag searching



Tabb, Analytical Chemistry, 2003

De novo sequencing tries to read the entire peptide sequence from the spectrum



Example of a manual de novo of an MS/MS spectrum

No more database necessary to extract a sequence!

Algorithms	*References*
Lutefisk	Dancik 1999, Taylor 2000
Sherenga	Fernandez-de-Cossio 2000
PEAKS	Ma 2003, Zhang 2004
PepNovo	Frank 2005, Grossmann 2005
RapidNovor	Ma 2015
…	…

Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

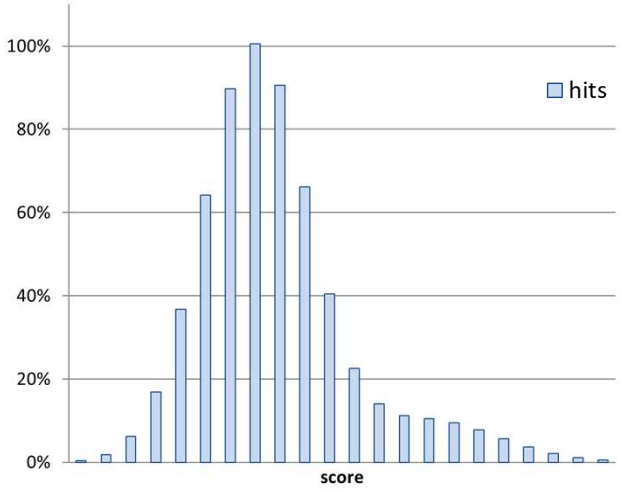
Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

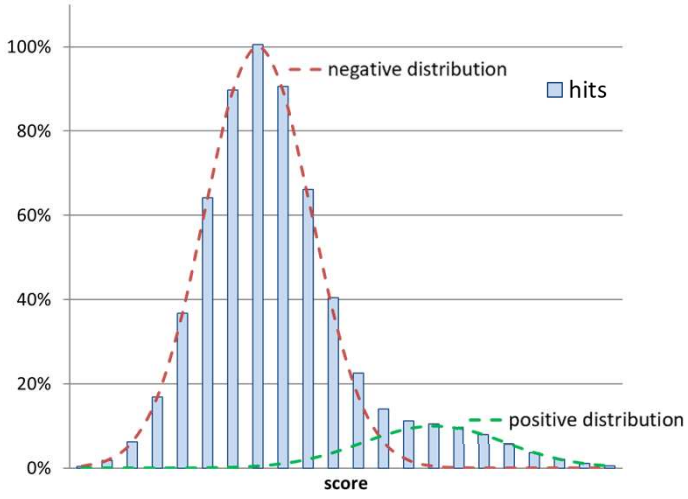
Protein inference: bad, ugly, and not so good

All hits, good and bad together, form a distribution of scores

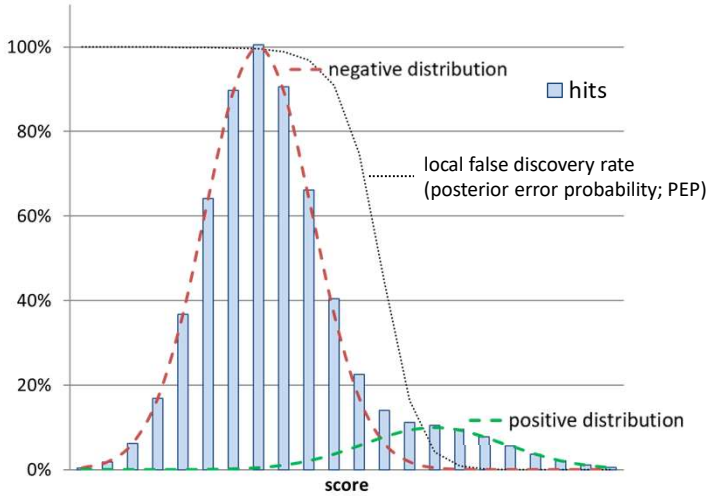


Nesvizhskii, J Proteomics, 2010

If we know how scores for bad hits distribute, we can distinguish good from bad by score



The separation is not perfect, which leads to the calculation of a local false discovery rate



Decoy databases are false positive factories, assumed to deliver representative bad hits

Three main types of decoy DB's are used:

- Reversed databases (*easy*)

  LENNARTMARTENS → SNETRAMTRANNEL

- Shuffled databases (*slightly more difficult*)

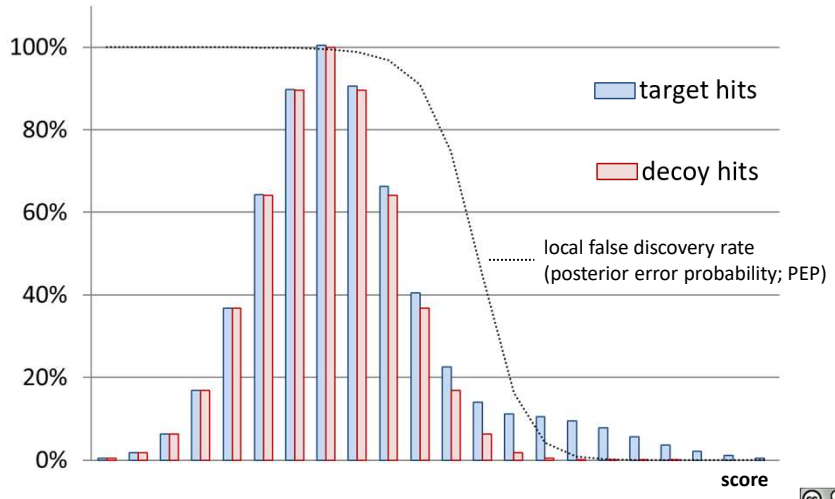
  LENNARTMARTENS → NMERLANATERTTN *(for instance)*

- Randomized databases (*as difficult as you want it to be*)

  LENNARTMARTENS → ~~GFVLAEPHSEAITK~~ *(for instance)*

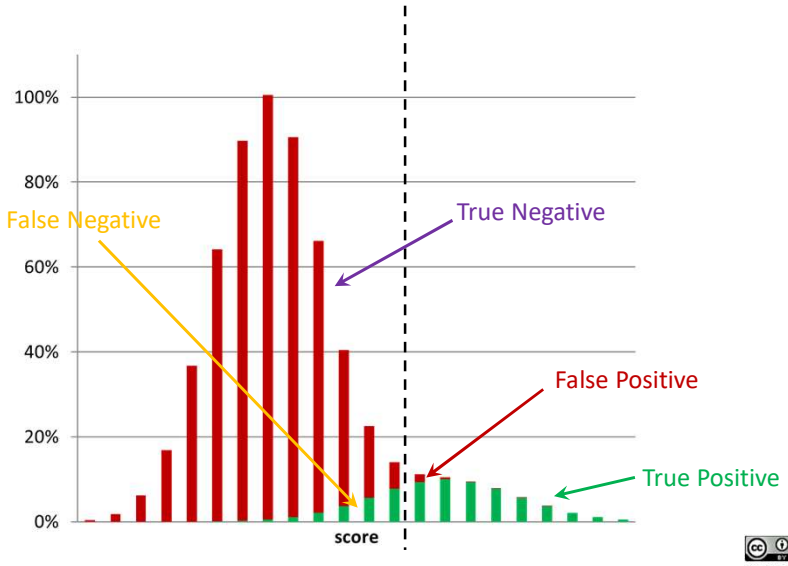
The concept is that each peptide identified from the decoy database is an incorrect identification. By counting the number of decoy hits, we can estimate the number of false positives in the original database, **provided that the decoys have similar properties as the forward sequences**.

With the help of the scores of decoy hits, we can assess the score distribution of bad hits



Käll, Journal of Proteome Research, 2008

Setting a threshold classifies all hits as either bad or good, which inevitably leads to errors



Amino acids, peptides, and proteins

Mass spectrometry basics

MS/MS spectra and identification

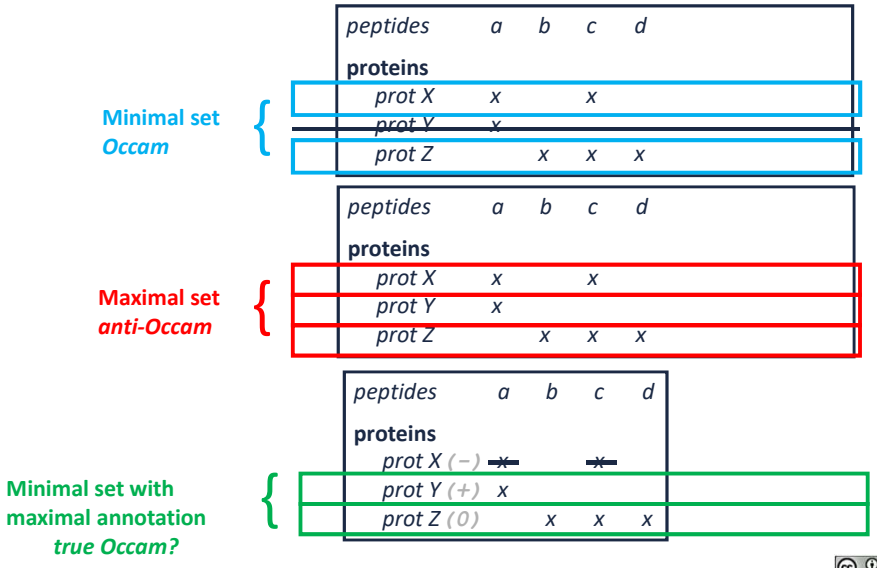
Database search algorithms in three phases

Sequencial search algorithms

Decoys and false discovery rate calculation

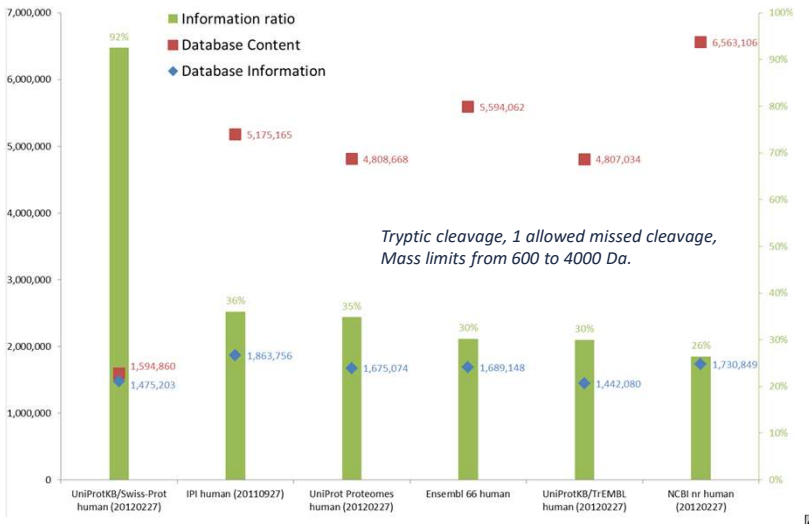
Protein inference: bad, ugly, and not so good

Protein inference is a question of conviction



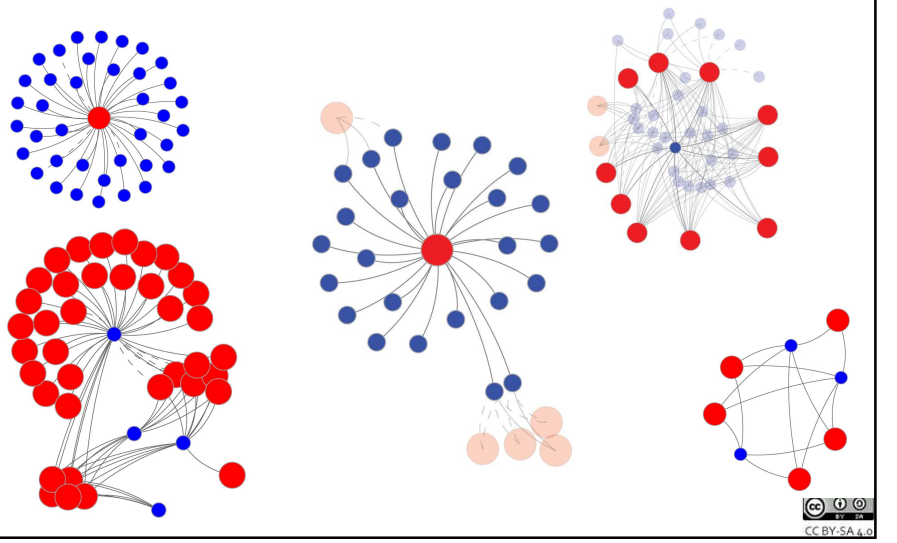
Martens, Molecular Biosystems, 2007

The complexity of protein inference is linked to the information ratio of a database

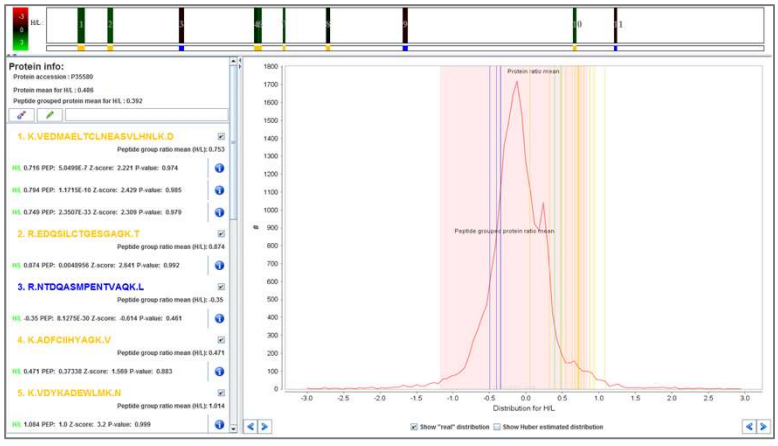


Barsnes, Amino Acids, 2013

In real life, protein inference issues will be mainly bad, often ugly, and occasionally good



Protein inference can create issues in quantification due to degenerate peptides



A nice example of the mess of degenerate peptides in quantification

Colaert, Proteomics, 2010