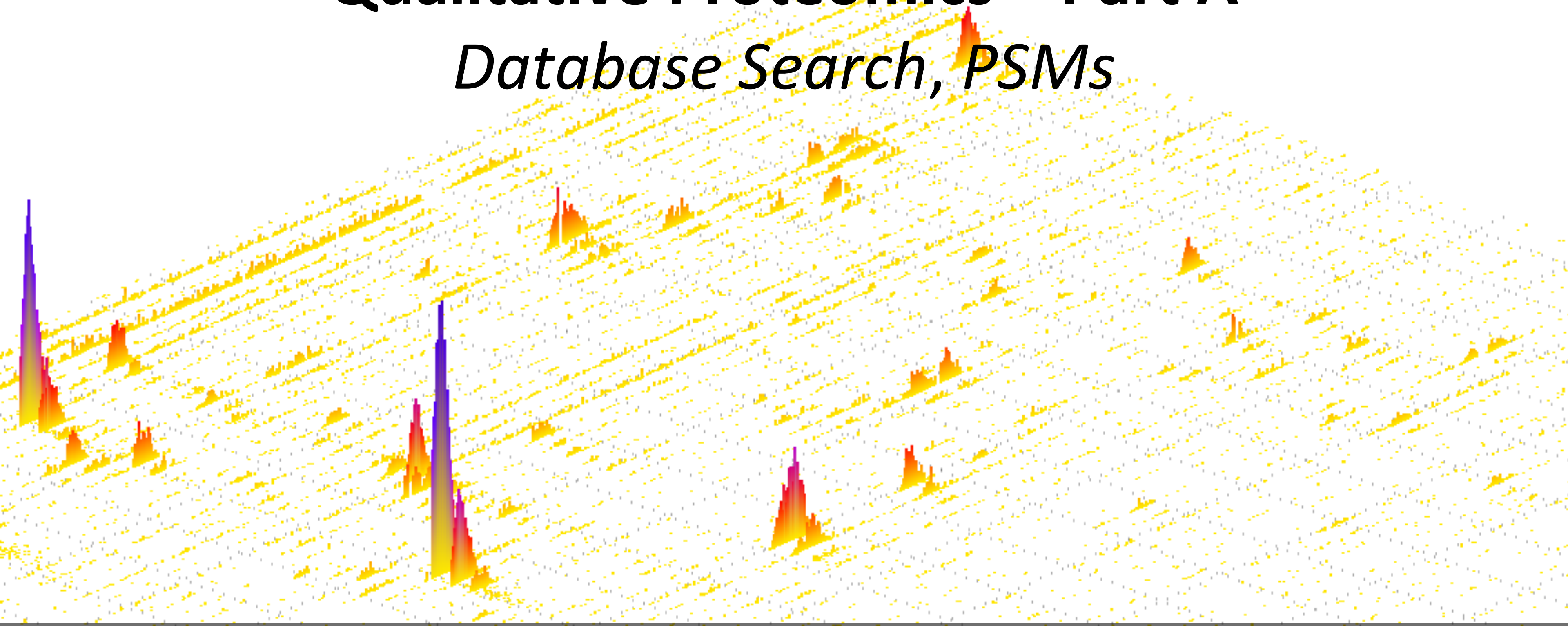
May Institute 2017

Computation and statistics for mass spectrometry and proteomics

Qualitative Proteomics – Part A

Database Search, PSMs



Oliver Kohlbacher

University of Tübingen and
MPI for Developmental Biology

KohlbacherLab.org | @okohlbacher

Today‘s Schedule

Monday 5/1/2017	**Proteomics and metabolomics with OpenMS**
8:00 AM	Registration
9:00 AM	Fundamentals of non-targeted proteomics and metabolomics
10:30 AM	**Refreshments**
11:00 AM	Hands-on: Tutorial: Introduction to OpenMS and KNIME
12:30 PM	**Lunch Break**
1:30 PM	Lecture: Database search, peptide-spectrum matches.
2:00 PM	Hands-on: Peptide and protein identification by database search.
3:00 PM	**Refreshments**
3:30 PM	Lecture: FDR estimation, protein inference, quality control.
4:00 PM	Hand-on: Peptide and protein identification by database search.
5:00 PM	Improvised poster session
6:00 PM	**Adjourn**

Overview

- **Concepts of Database Search**
  - Fundamentals of peptide fragmentation, ion series
  - Database search: key ideas
- **Database Search Engines**
  - X!Tandem
  - Scoring function and underlying statistics

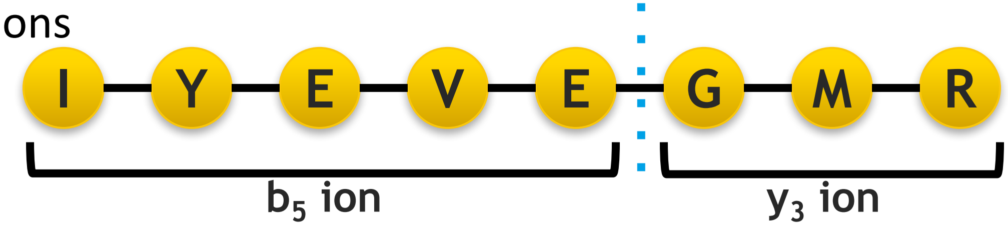
PEPTIDE DATABASE SEARCH

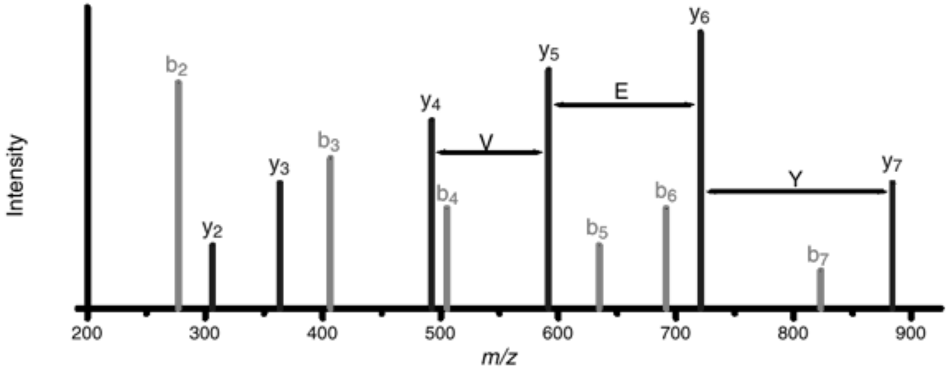
- Peptide fragmentation
- Database search concepts

Peptide Identification

Why can we identify peptides from tandem MS spectra?

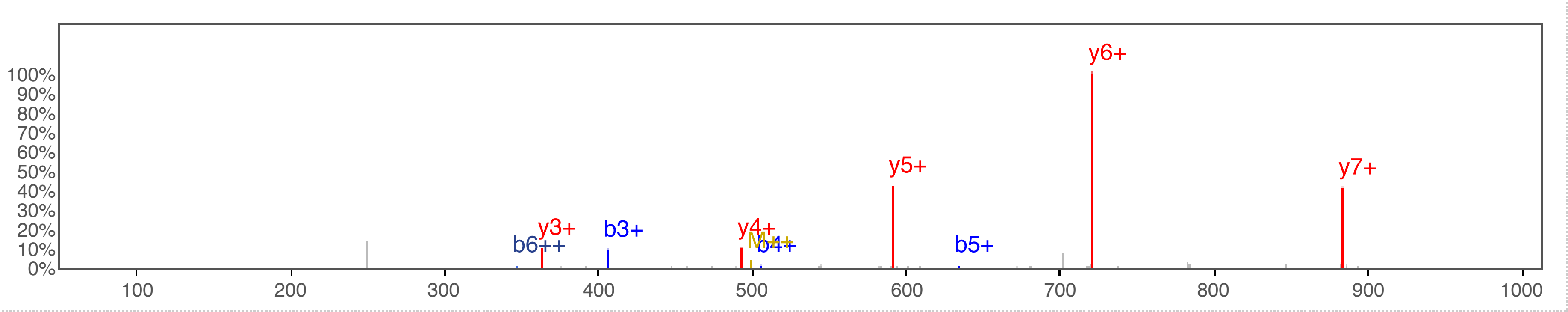
- **Goal: identify sequence**
- Tandem MS
  - Sequence consists of the **same 20 building blocks** (amino acids)
  - CID: peptide breaks preferentially along the **backbone**
  - Peptide **fragment ions correspond to prefixes and suffixes** of the whole peptide sequences
  - Complete ion series (ladders) reveal the sequence via mass differences of adjacent fragment ions





Peptide Identification

- **Issues**
  - Spectra are incomplete – ions are missing
  - Missing information makes it very hard to reconstruct full sequence

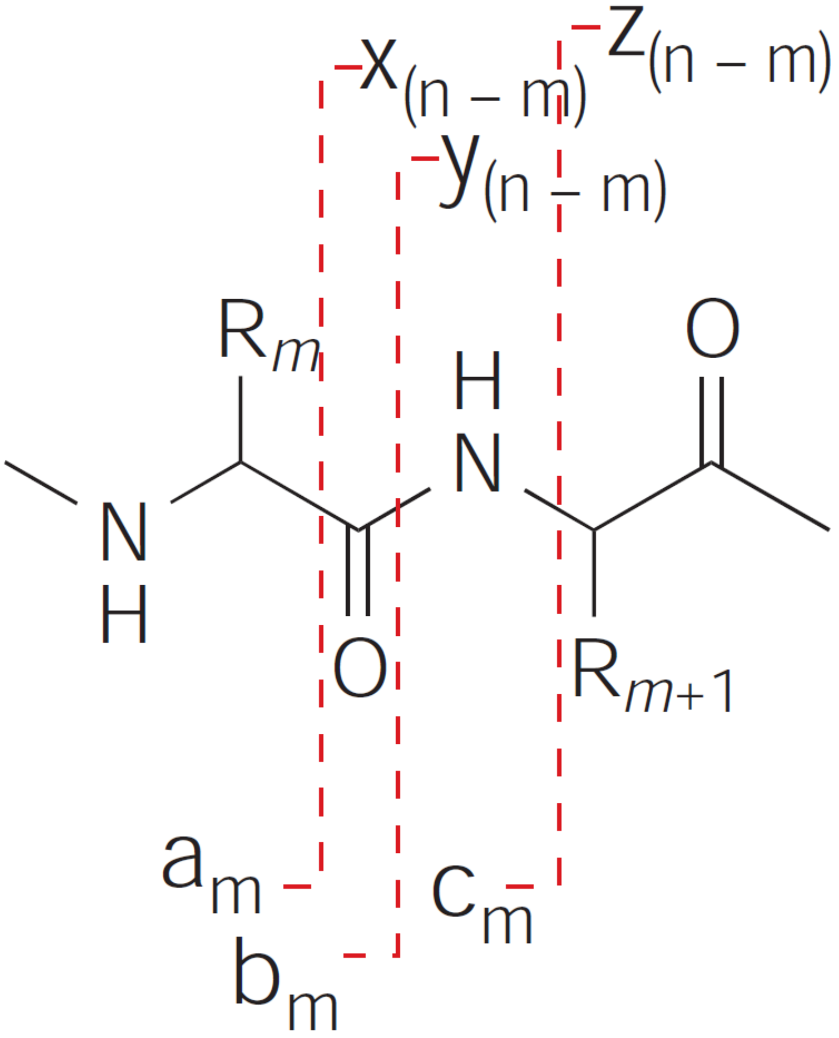


- **Database search**
  - Not all sequences occur in a proteome – only a fraction of sequence space is used
  - Try to find those sequences that match the ions present in the spectrum

Consensus spectrum: PeptideAtlas id 829036

Product ion generation

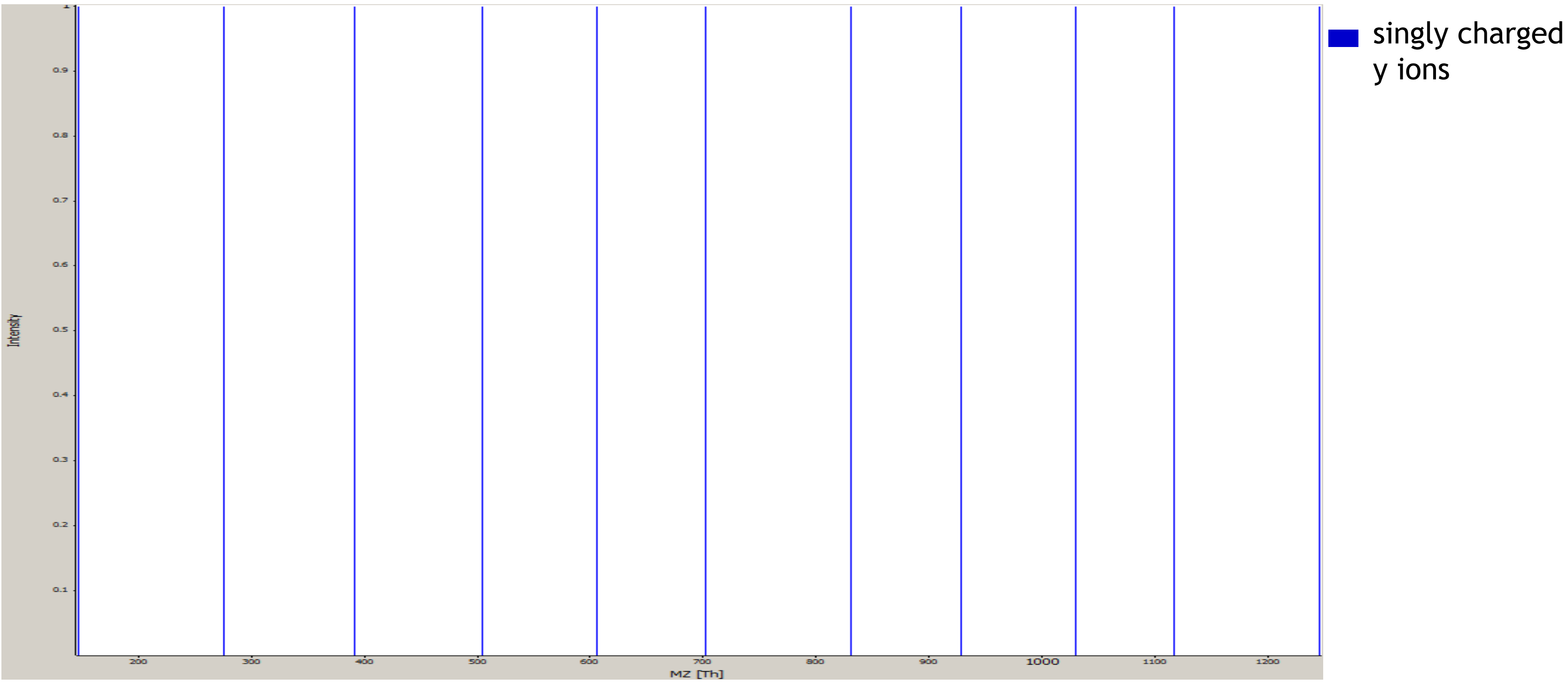
- A peptide of length n can potentially give rise to a,b,c and x,y,z ions. This example shows the fragments that can be produced between amino acids R_m and R_{m+1}
- This nomenclature for fragment ions was first proposed by Roepstorff and Fohlman in 1984
  (Roepstorff and Fohlman, Biological Mass Spectrometry, Volume 11, Issue 11, page 601, November 1984)



Steen and Mann. Nature Reviews, Molecular Cell Biology, Vol. 5 2004

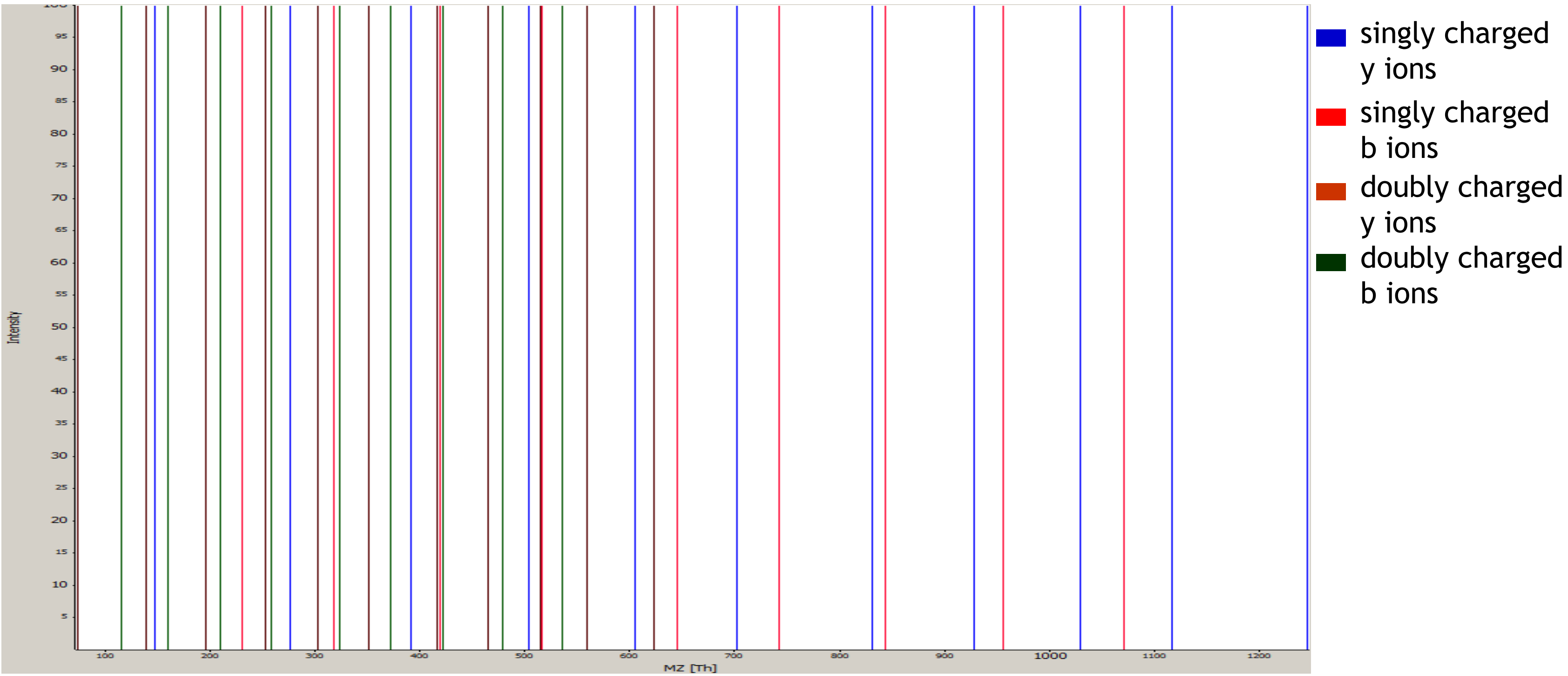
Ion Series - Example

- For simplicity we will consider theoretical spectra for the artificial (tryptic) peptide TESTPEPTIDEK
- For singly charged ion fragments, only one of the sister fragments will be observed



Ion Series - Example

- If the same peptide was multiply charged, the charges are usually distributed across the product ions
- Tandem spectrum then usually contains both sister ions and also doubly charged product ions



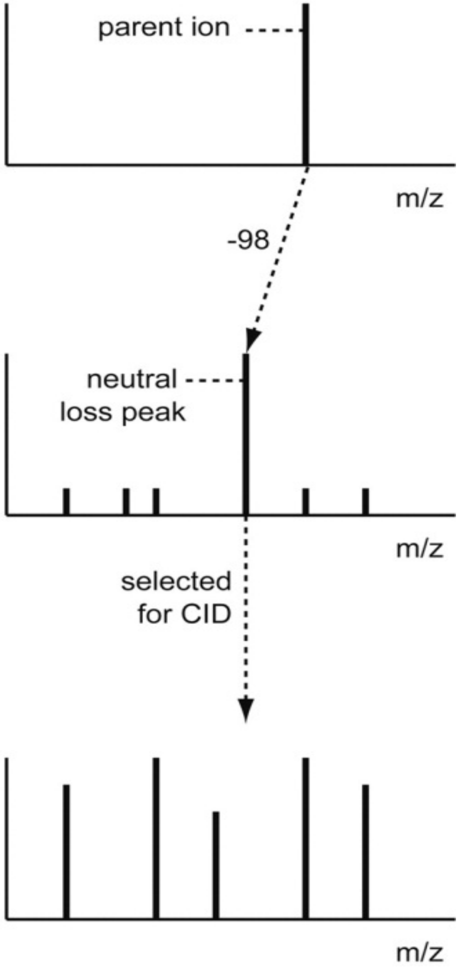
Neutral Losses

- Besides backbone ions, we also observe the precursor ions and precursor ions with **neutral losses**

- *Neutral losses* most frequently occur as
  - **water loss** (H_2O: -18.011 Da) on S, T, D and E
  - **ammonia loss** (NH_3: - 17.027 Da) on R, K, N and Q
  - **loss of phosphoric acid** (H_3PO_4:-98 Da) on S, T and Y

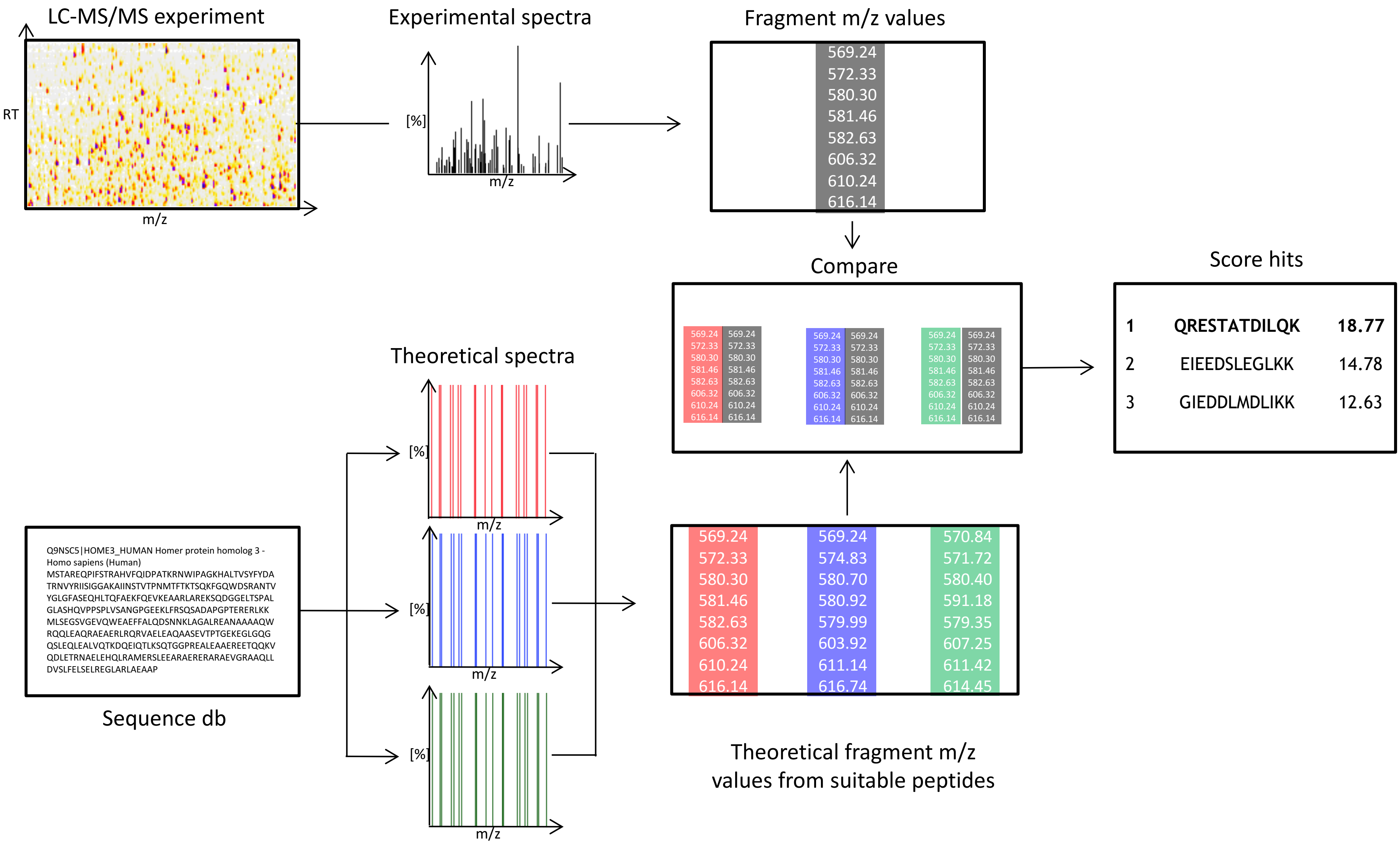
- Neutral losses are uncharged fragments, but result in an additional charged ion with $mass_{ion} – mass_{neutral}$

- The problem of very intense ions resulting from neutral losses of precursor ions can be overcome by triggering an additional fragmentation



Hoffert J D et al. PNAS 2006;103:7159-7164

Database Search – Overview



LC-MS/MS experiment

Experimental spectra

Fragment m/z values

569.24
572.33
580.30
581.46
582.63
606.32
610.24
616.14

Compare

569.24	569.24	569.24	569.24	569.24	569.24
572.33	572.33	572.33	572.33	572.33	572.33
580.30	580.30	580.30	580.30	580.30	580.30
581.46	581.46	581.46	581.46	581.46	581.46
582.63	582.63	582.63	582.63	582.63	582.63
606.32	606.32	606.32	606.32	606.32	606.32
610.24	610.24	610.24	610.24	610.24	610.24
616.14	616.14	616.14	616.14	616.14	616.14

Score hits

1	**QRESTATDILQK**	**18.77**
2	EIEEDSLEGLKK	14.78
3	GIEDDLMDLIKK	12.63

Theoretical spectra

Q9NSC5|HOME3_HUMAN Homer protein homolog 3 -
Homo sapiens (Human)
MSTAREQPIFSTRAHVFQIDPATKRNWIPAGKHALTVSYFYDA
TRNVYRIISIGGAKAIINSTVTPNMTFTKTSQKFGQWDSRANTV
YGLGFASEQHLTQFAEKFQEVKEAARLAREKSQDGGELTSPAL
GLASHQVPPSPLVSANGPGEEKLFRSQSADAPGPTERERLKK
MLSEGSVGEVQWEAEFFALQDSNNKLAGALREANAAAAQW
RQQLEAQRAEAERLRQRVAELEAQAASEVTPTGEKEGLGQG
QSLEQLEALVQTKDQEIQTLKSQTGGPREALEAAEREETQQKV
QDLETRNAELEHQLRAMERSLEEARAERERARAEVGRAAQLL
DVSLFELSELREGLARLAEAAP

Sequence db

569.24	569.24	570.84
572.33	574.83	571.72
580.30	580.70	580.40
581.46	580.92	591.18
582.63	579.99	579.35
606.32	603.92	607.25
610.24	611.14	611.42
616.14	616.74	614.45

Theoretical fragment m/z values from suitable peptides

Database Search – Key Steps

1. **Extract all sequence candidates** (usually tryptic) from the database matching the precursor mass of the MS^2 spectrum with a given error tolerance

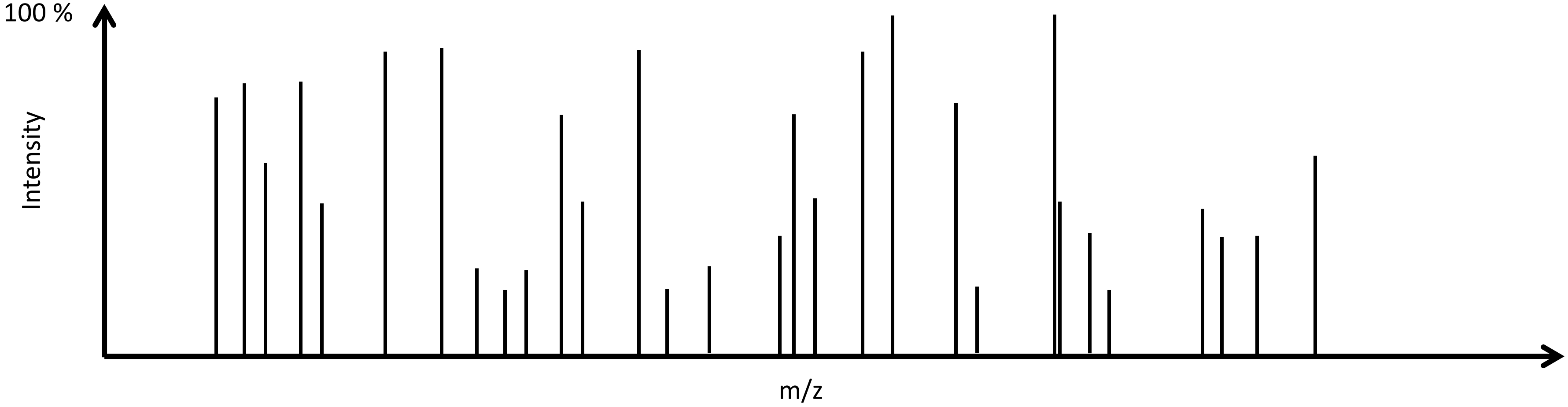
2. **Generate theoretical spectrum** for each of the candidate sequences

3. **Align** the theoretical spectra to the experimental spectrum

4. **Score** the alignment

5. **Report all peptide-spectrum matches** above a certain score threshold

Step 1. Generate Candidates



- Given: Experimental spectrum *S*
- Task: Identify the correct sequence for *S* from a given protein database

1. Define the search space for *S* for a given mass tolerance *d:*
   - m_{prec} is the mass of the precursor ion of spectrum *S*
   - From the database, extract all peptide sequences with mass m_{cand} given that

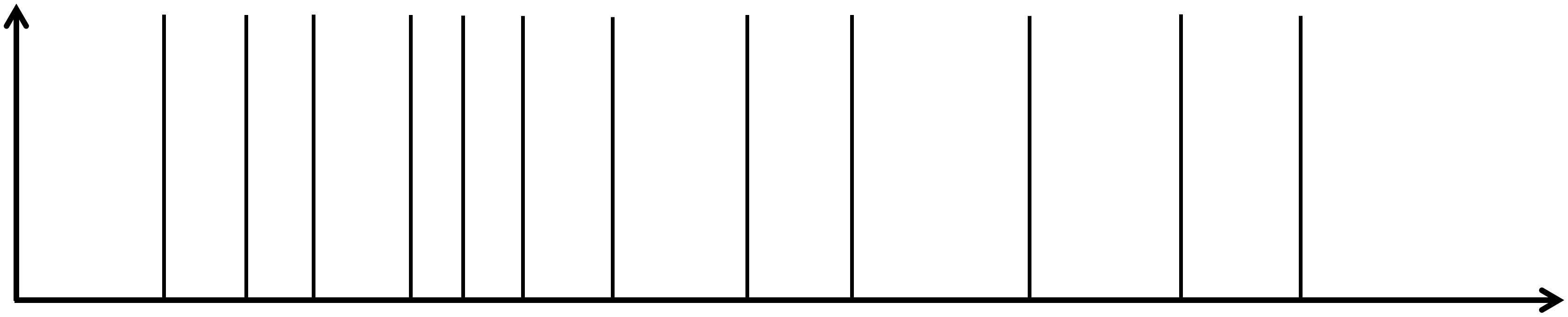
$$|m_{prec} - m_{cand}| \leq d$$

   - This set of candidates is defined as the search space for spectrum *S* and denoted as

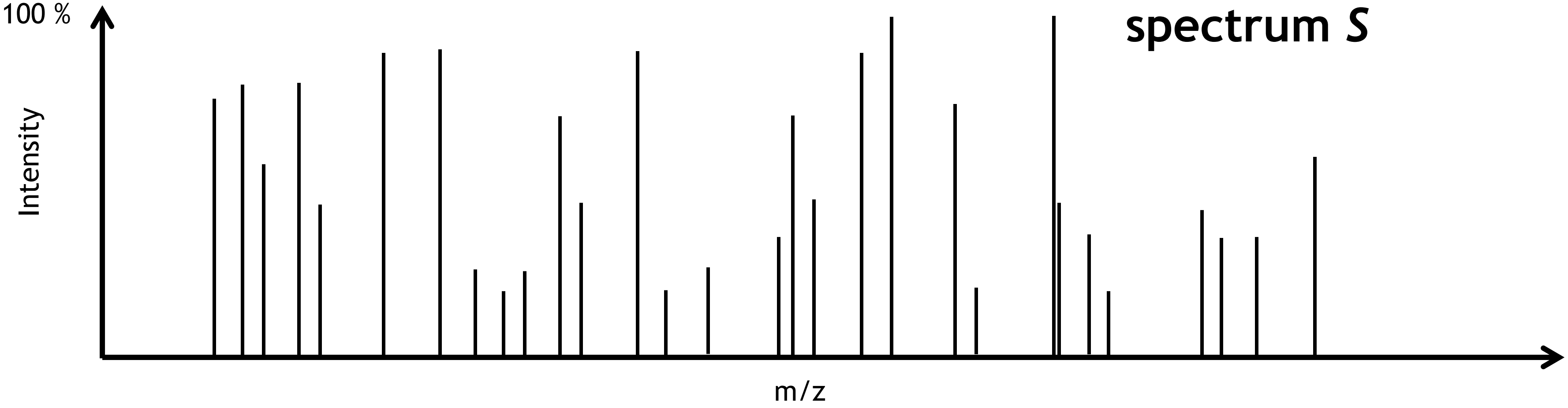
$$\Omega_S$$

Step 2: Generate Theoretical Spectra

- 1st option: extract all masses from the MS^2 spectrum
- 2nd option: try to model fragment ion intensities

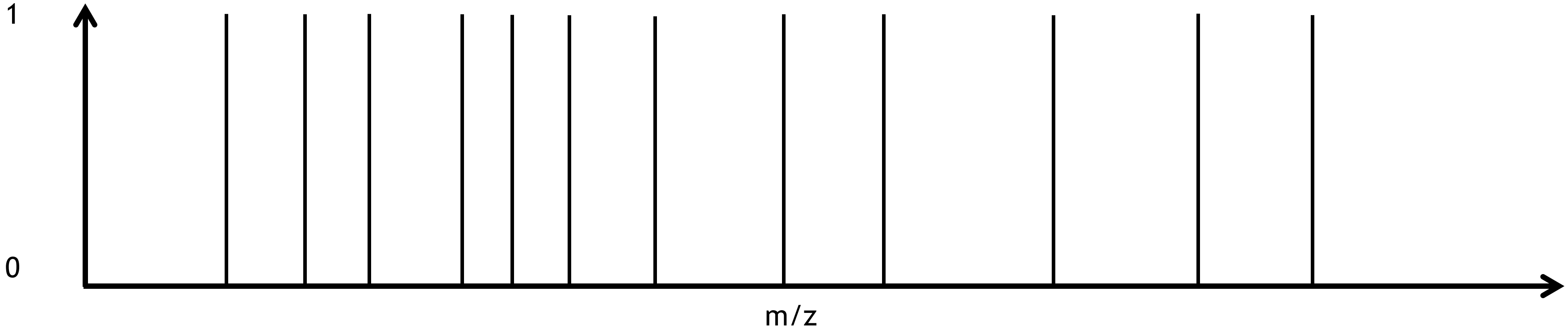


Step 3. Align Spectra

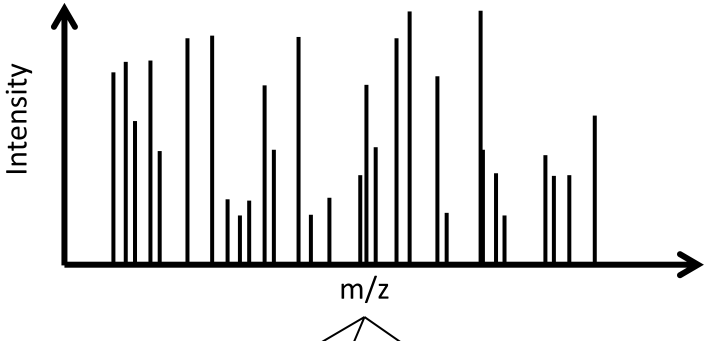


spectrum ***S***

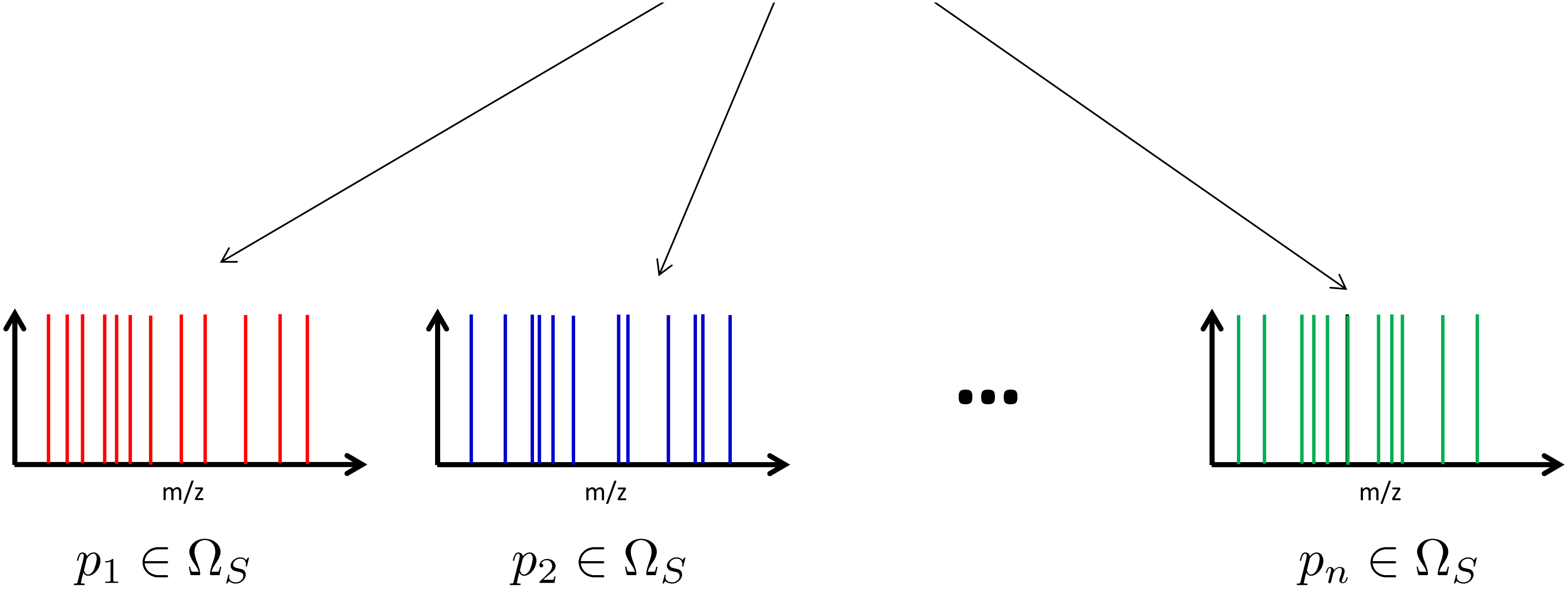
Theoretical spectrum *T*, generated from a sequence $p_i \in \Omega_S$



Step 3: Align Spectra



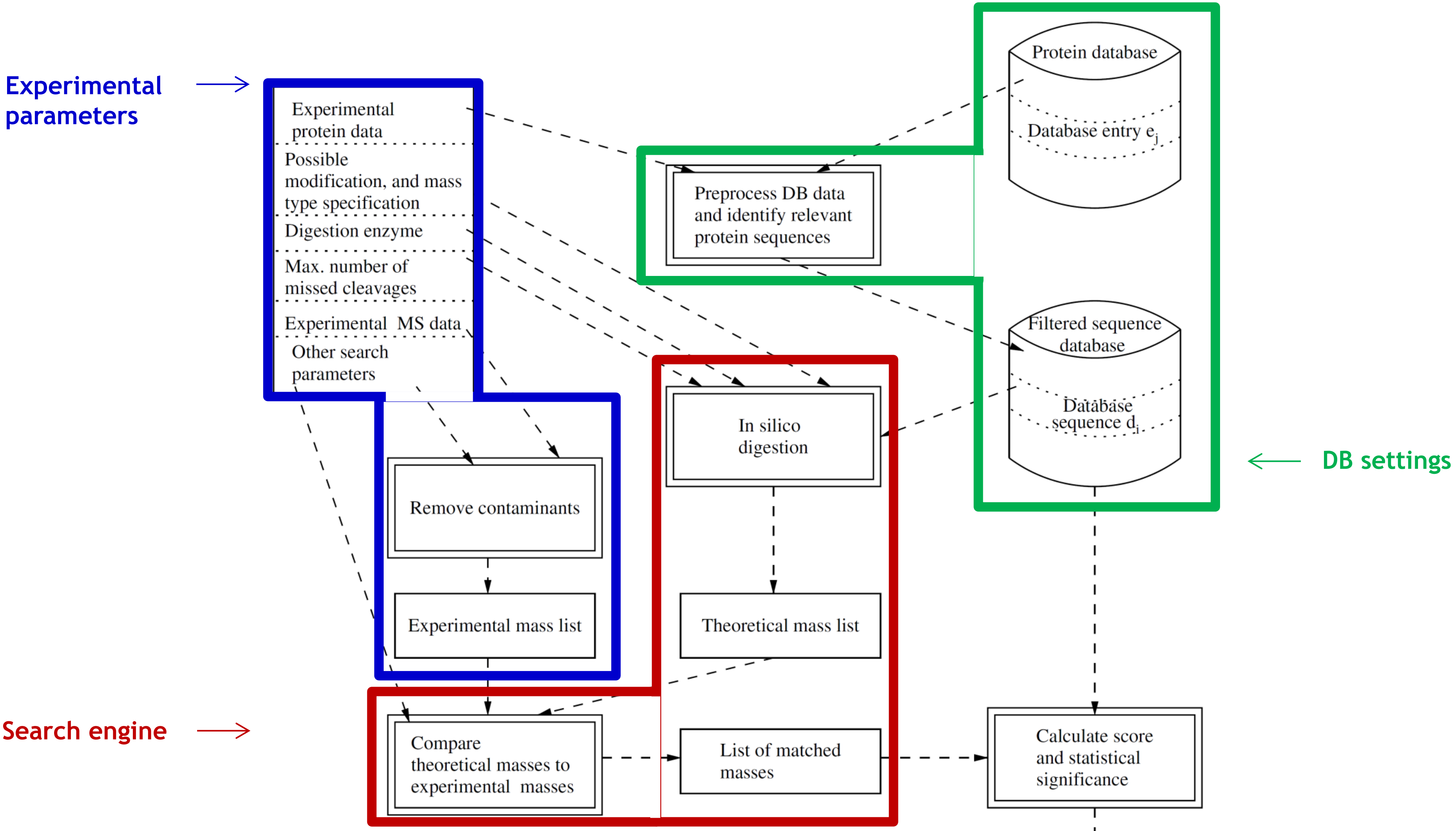
2. Compare theoretical spectra for all $p_i \in \Omega_S$ to the experimental spectrum *S*



Step 4: Scoring of peptide candidates

- There are numerous tools for the comparison of theoretical and experimental candidate peptides
- The main difference of search engines is the implementation of the scoring schemes (resulting in differences in runtime and performance)
- However, conceptually all search engine algorithms are based on fragment ion comparison

More Complex Workflow


Experimental parameters
Experimental protein data
Possible modification, and mass type specification
Digestion enzyme
Max. number of missed cleavages
Experimental MS data
Other search parameters
Protein database
Database entry e_j
Preprocess DB data and identify relevant protein sequences
Filtered sequence database
Database sequence d_i
DB settings
Remove contaminants
Experimental mass list
In silico digestion
Theoretical mass list
Search engine
Compare theoretical masses to experimental masses
List of matched masses
Calculate score and statistical significance
Presentation

Real-life implementations are more complex and often require additional pre-/post-processing and depend on a large number of parameters

SEARCH ENGINES

- X!Tandem
- Sequest
- Other search engines

Database Search Engines

- Dozens of different database search tools are currently being used
- Tools differ with respect to
  - Spectrum pre-processing
  - Scoring of peptide-spectrum matches
  - Post-processing of peptide-spectrum matches
  - Score statistics
  - Speed
- Results for the same dataset will differ between search engines!

X!Tandem

- X!Tandem
  - is a popular open-source database search engine
  - is fast
  - has been published in various versions including multiple refinements to the core algorithm sketched here (latest version: X!Tandem Sledgehammer, 2013)

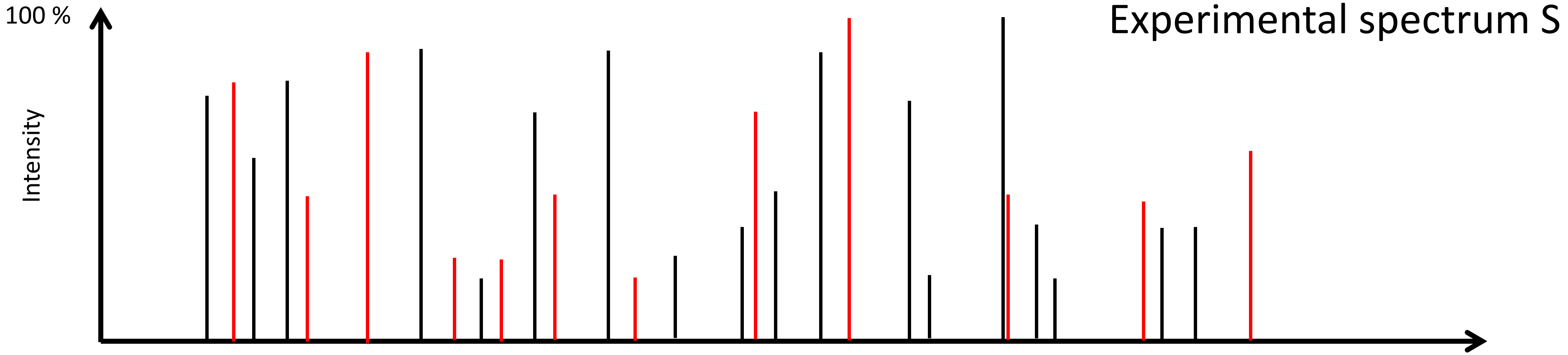
Original reference

- Craig,R. and Beavis,R.C. (2003) *Rapid Commun. Mass Spectrom.*, **17**, 2310–2316.

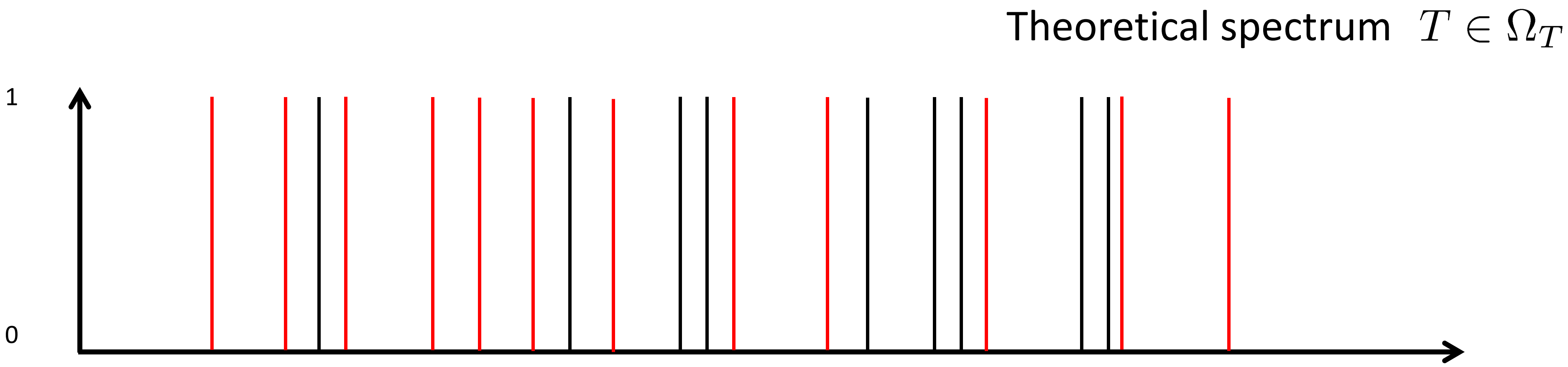
- http://www.thegpm.org/tandem/instructions.html

Find overlapping masses

To find overlapping masses, a maximal **fragment mass tolerance** window needs to be set (for ion traps this is usually 0.5 Da)

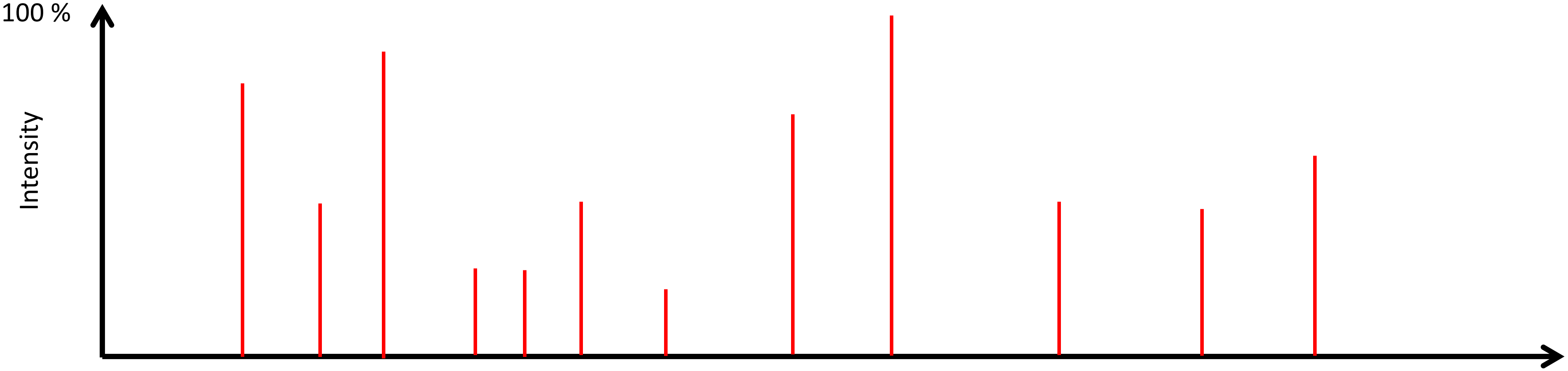


Experimental spectrum S



Theoretical spectrum $T \in \Omega_T$

X!Tandem's dot product



- Reduce the experimental spectrum to only those peaks that match peaks in the theoretical spectrum

- Calculate dot product (dp) (using ion intensities and the number of matching ions)

$$dp = \sum_{i=0}^{n} I_i \, P_i$$

Intensities from experimental spectrum

I_i ... fragment ion intensities

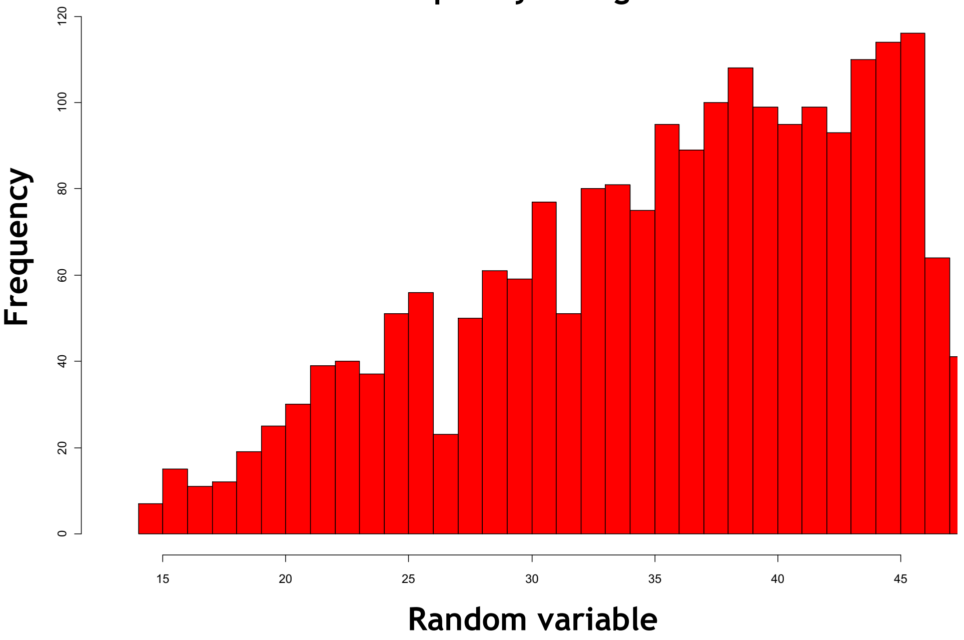
Predicted or not in theoretical spectrum

$P_i \in \{0, 1\}$

Survival function and e-value

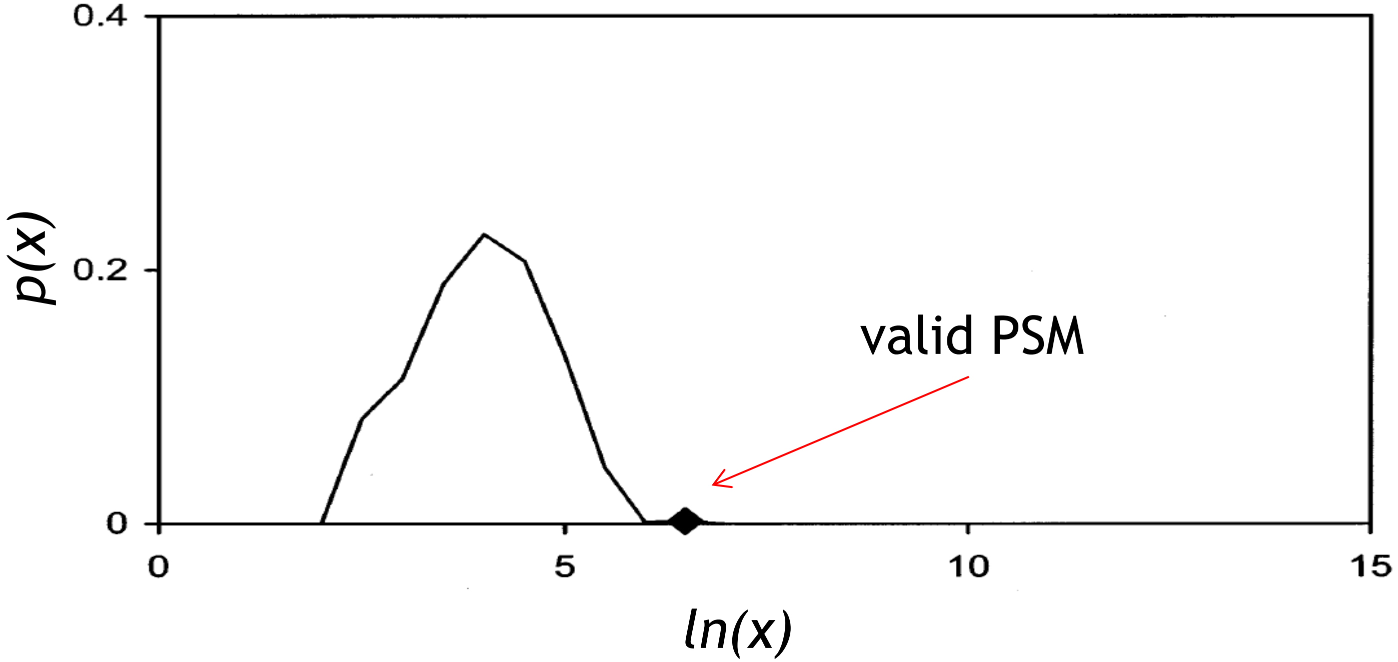
- Let x represent the dot product score for the experimental spectrum *S* and the theoretical spectrum $T \in \Omega$.
- *p*(x) is calculated from the frequency histograms (counts of PSMs per score bin).
- With f(x), the number of PSMs that are given the score x, p(x) is calculated as $p(x) = f(x)/N$ with N being the total number of PSMs

Example of a frequency histogram



Fenyö and Beavis, Anal. Chem.2003, 75, 768-774

Survival function and e-value



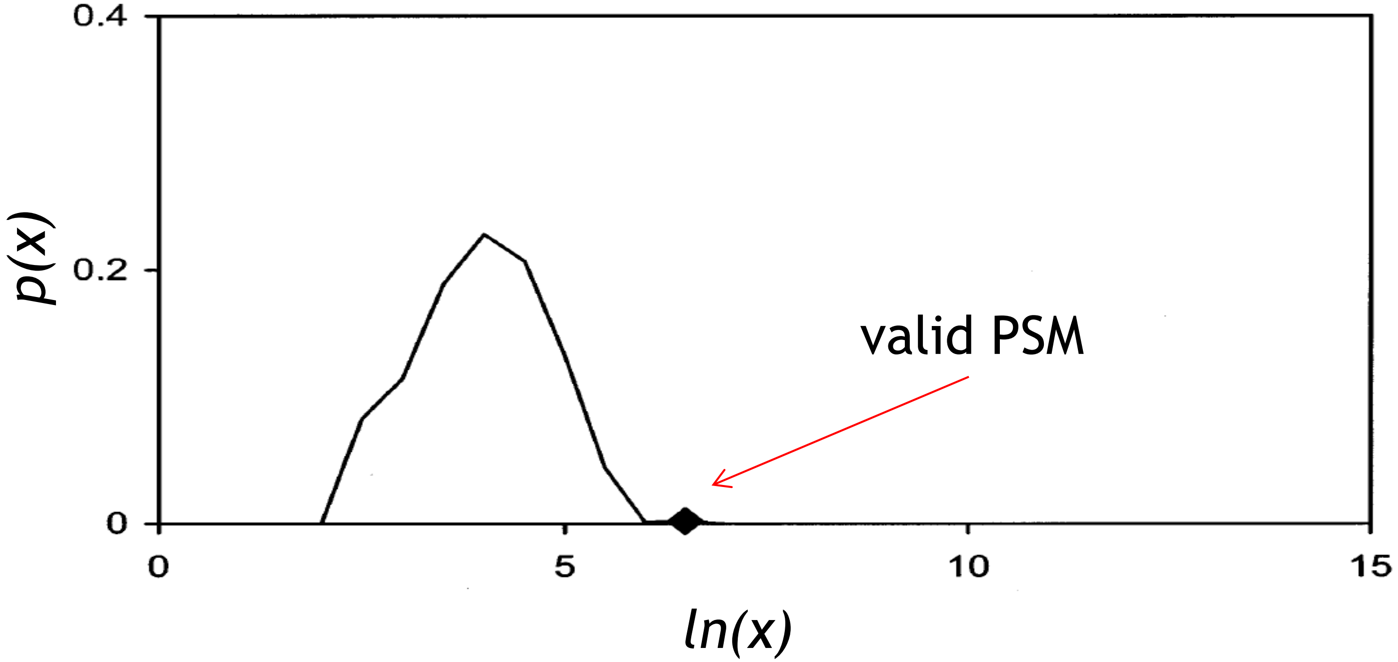
- *The survival function, s(x), for a discrete stochastic score probability distribution, p(x) is defined as*

$$s(x) = P(X > x) = \sum_{X>x} p(x)$$

where *P(X > x)* is the probability to have a greater value than *x* by random matches in a database.

Fenyö and Beavis, Anal. Chem.2003, 75, 768-774

Survival function and e-value



- With the survival function *s(x),* we can calculate the E-value *e(x)*, indicating the number of PSMs that are expected to have scores of *x* or better

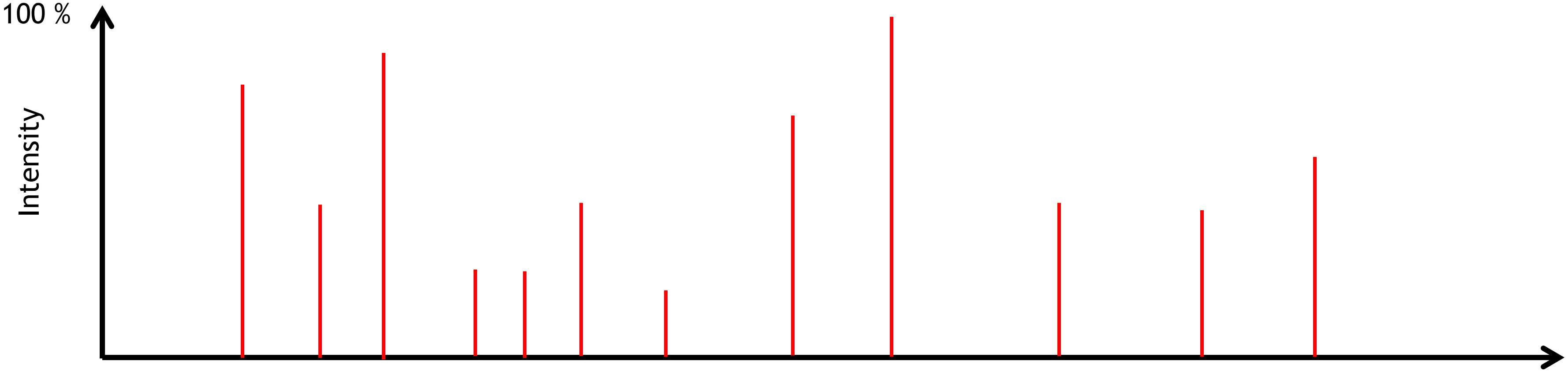
$$e(x) = ns(x)$$

where *n* is the number of sequences in Ω_S

- Now, each PSM can be ranked accoring to *e(x)*

Fenyö and Beavis, Anal. Chem.2003, 75, 768-774

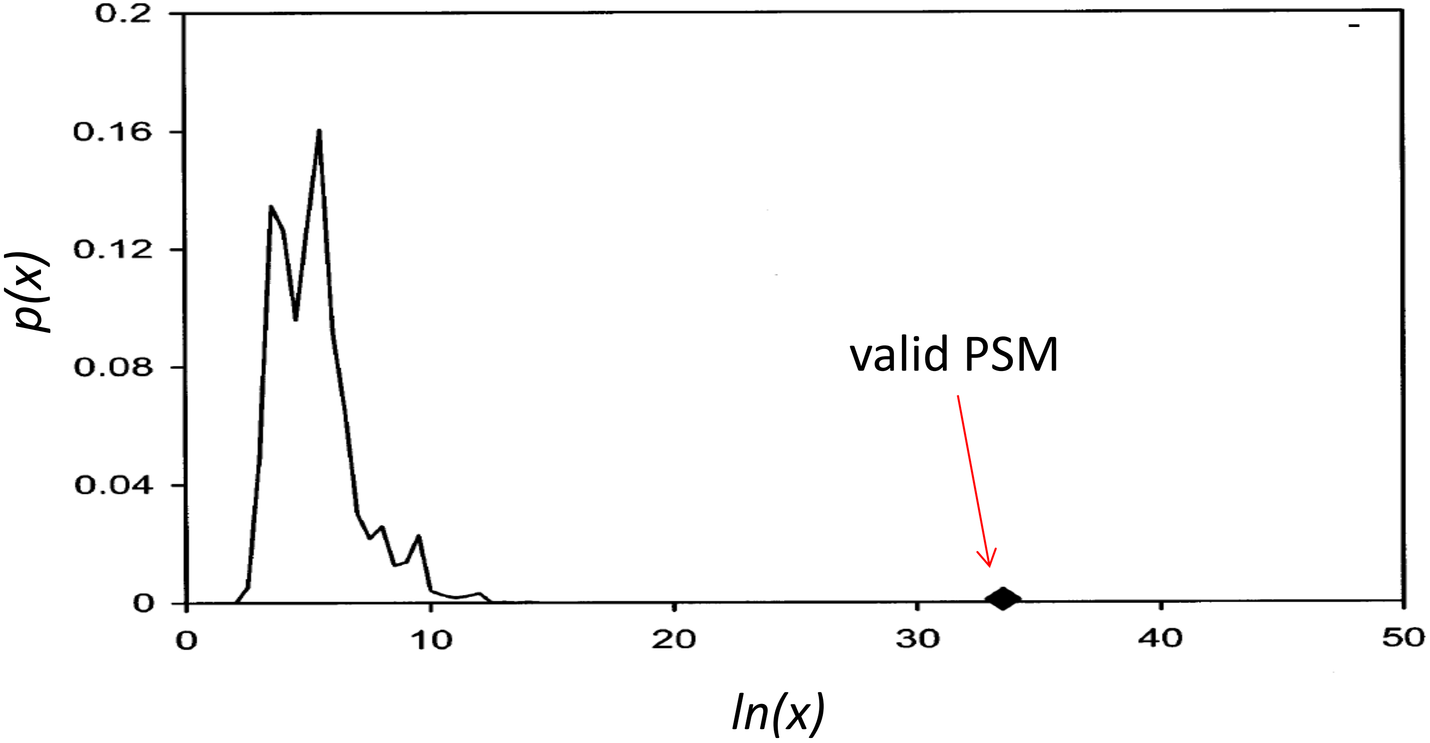
X!Tandem Hyperscore



- The hyperscore (HS) is calculated by multiplying with factorials of the number of assigned b and y ions.
- The use of the factorials is based on the hypergeometric distribution that is assumed for matches of product ions

Fenyö and Beavis, Anal. Chem.2003, 75, 768-774

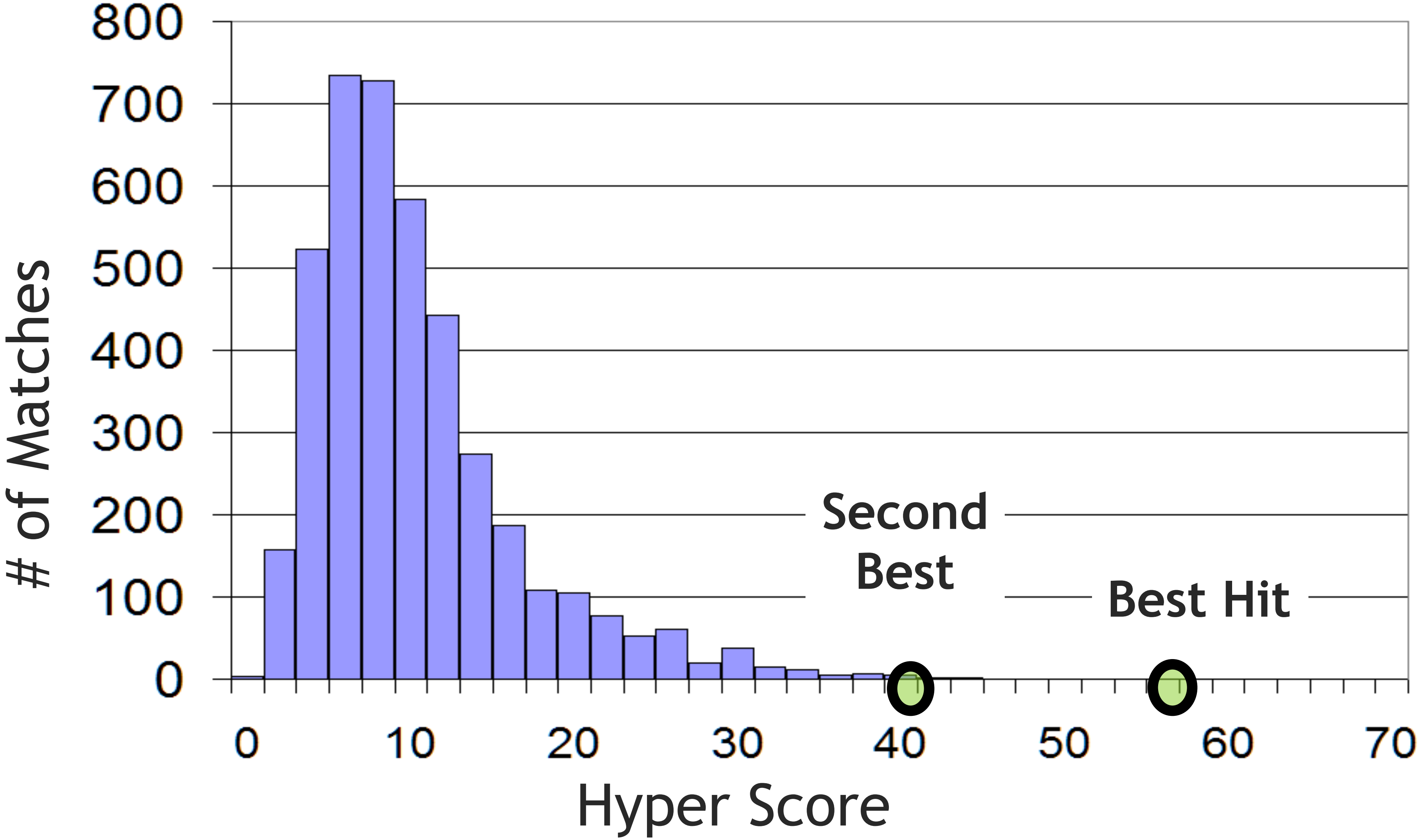
X!Tandem Hyperscore



- If p(x) is now plotted as a function of their log(hyperscores), the valid PSM is much better separated from the bulk of incorrect assignments

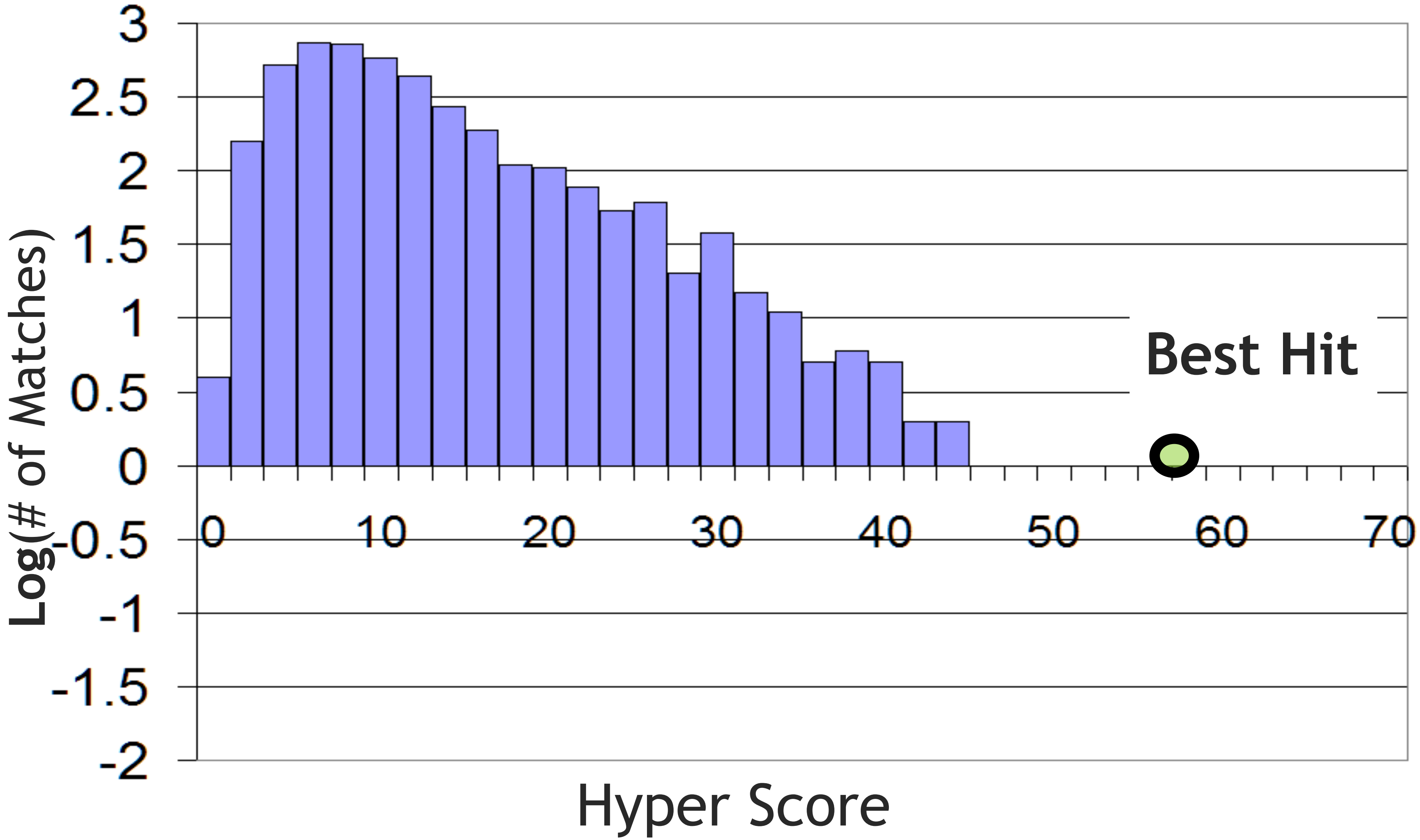
http://www.proteomesoftware.com/pdf_files/XTandem_edited.pdf

Distribution of “Incorrect” Hits



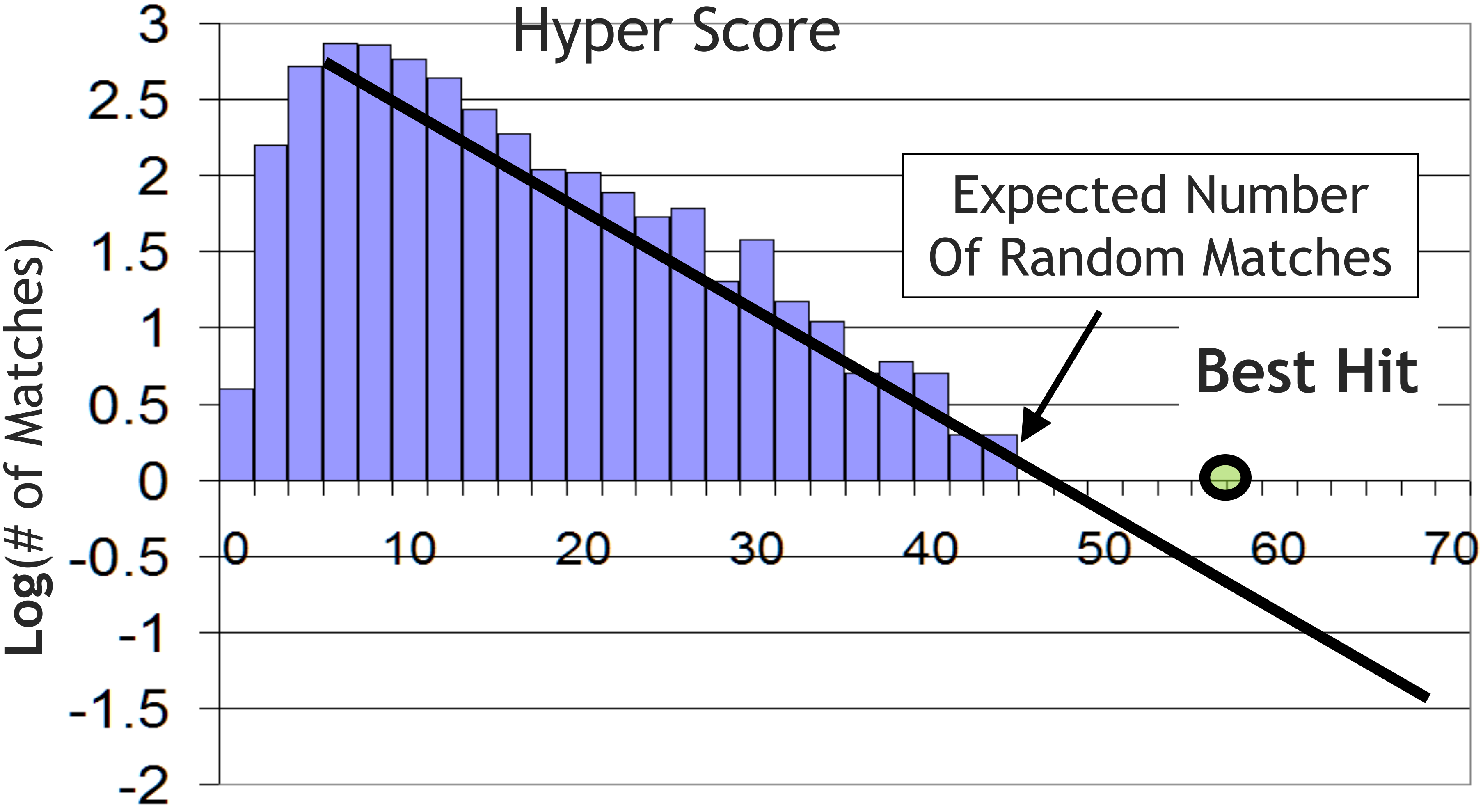
Adapted from Interpreting MS/MS Proteomics Results by Brian Searls

Estimate Likelihood (E-Value)



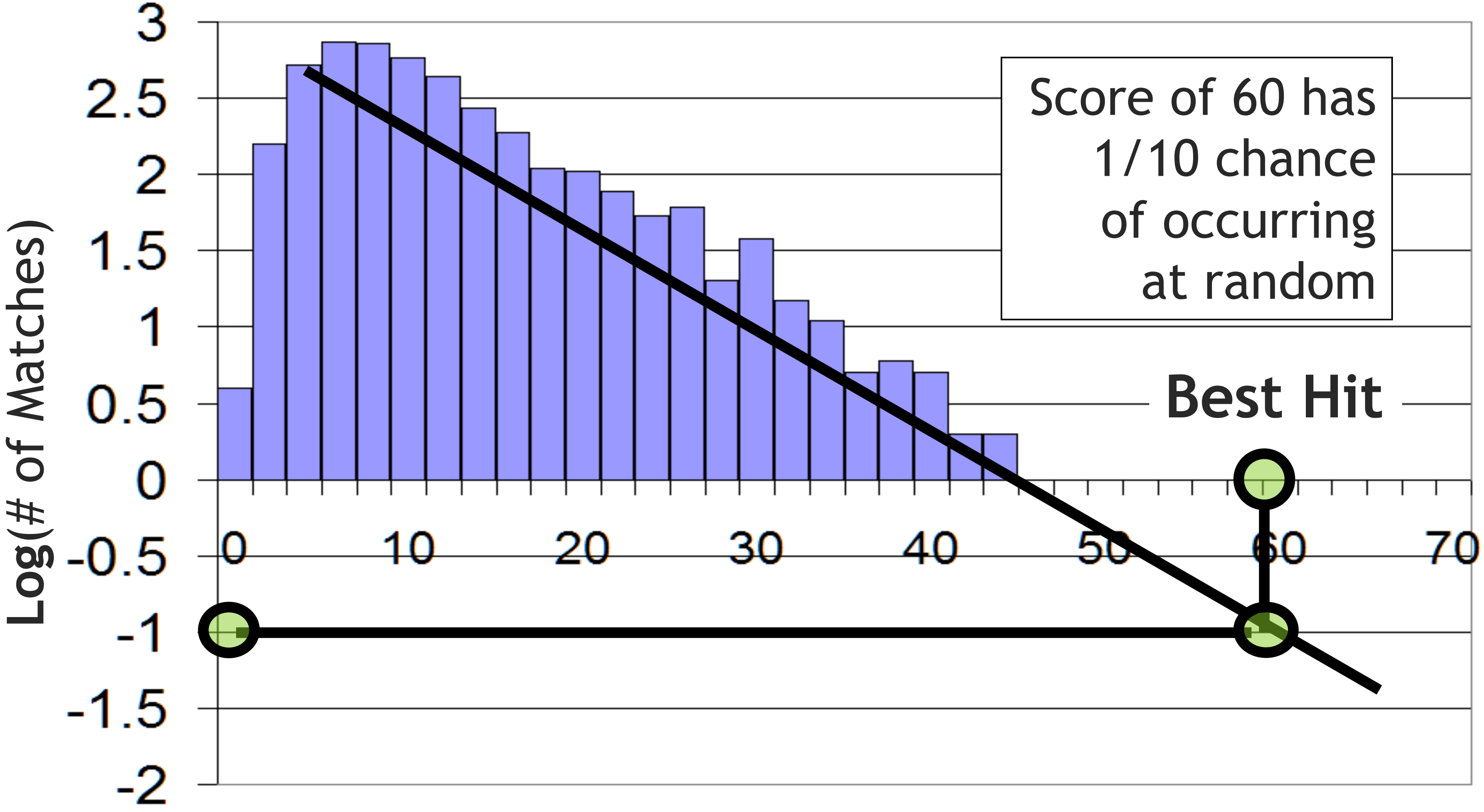
Adapted from Interpreting MS/MS Proteomics Results by Brian Searls

Estimate Likelihood (E-Value)



Adapted from Interpreting MS/MS Proteomics Results by Brian Searls

Estimate Likelihood (E-Value)



Score of 60 has 1/10 chance of occurring at random

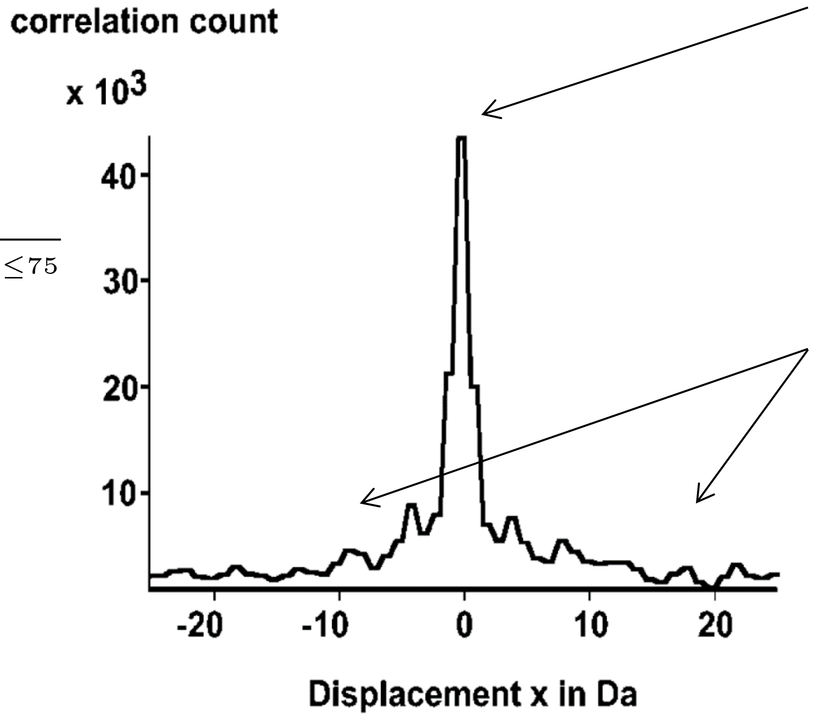
Adapted from Interpreting MS/MS Proteomics Results by Brian Searls

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* [Da] $\in \{-75, 75\}$



Displacement x = 0, denotes the cross correlation

Displacement x != 0, denotes the auto-correlation

Grenzel et al, Proteomics. 2003(3):1597-1610.

Other Search Engines

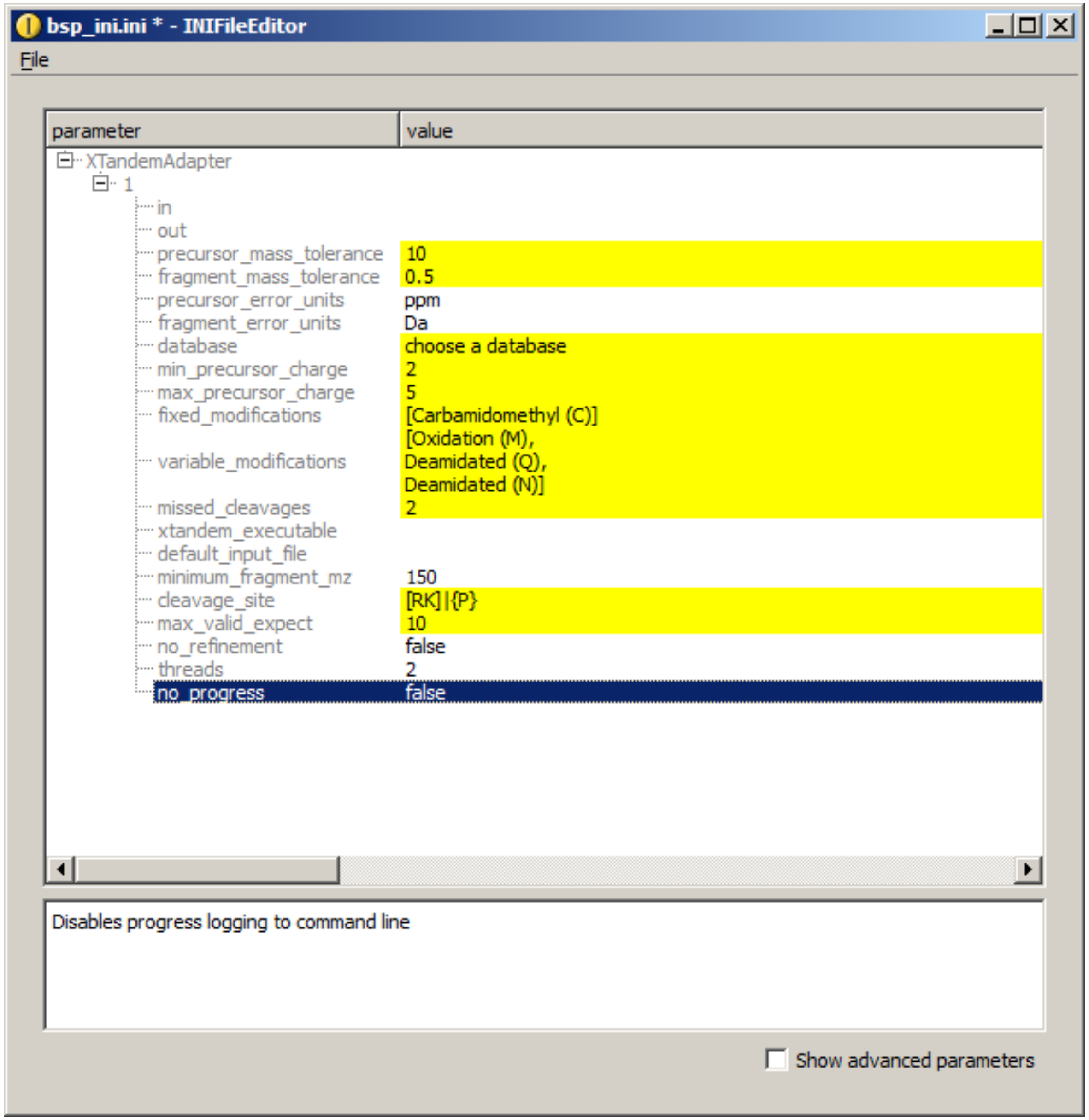
- Mascot from Matrix Science (http://www.matrixscience.com/)
  - Mascot is one of the most popular search engines
  - Commercial software
  - Algorithmic details have never been published
  - Mascot calculats *p*-values for all candidates in the search space and ranks the output according to these p-values
- Phenyx
  - Commercial software
  - Colinge et al., Proteomics. Vol. 3, No. 8, August 2003, pp. 1454-1463.
- InsPecT
  - Very fast open-source search engine
  - Designed for the identification of posttranslational modification
  - Tanner et al., J Proteome Res. 2005 Jul-Aug;4(4):1287-95.
- Myrimatch
  - Open source
  - Tabb et al., J Proteome Res. 6(2) 654-61. 2007 Feb

Search Settings

- OpenMS offers TOPP tools for the most common search engines

- .ini files allow to adjust the parameters

- This is an example for X!Tandem settings for analyzing LTQ-Orbitrap data



parameter	value
XTandemAdapter	
1	
in	
out	
precursor_mass_tolerance	10
fragment_mass_tolerance	0.5
precursor_error_units	ppm
fragment_error_units	Da
database	choose a database
min_precursor_charge	2
max_precursor_charge	5
fixed_modifications	[Carbamidomethyl (C)]
variable_modifications	[Oxidation (M), Deamidated (Q), Deamidated (N)]
missed_cleavages	2
xtandem_executable	
default_input_file	
minimum_fragment_mz	150
cleavage_site	[RK]|{P}
max_valid_expect	10
no_refinement	false
threads	2
no_progress	false

Disables progress logging to command line

☐ Show advanced parameters

Mass Tolerance Settings

- Mass tolerance settings:
  - Easy to estimate when knowing the instrument, calibration runs
  - Precursor tolerance determines search space
    - should be stringent, but broad enough to have several entries per search space (e.g., for E-value calculation)
    - 5-10 ppm is commonly used for data acquired on well-calibrated Orbitrap instruments
  - Product (or fragment) tolerance determines the number of theoretical fragment ions that can be matched to the experimental spectrum
    - again, should be stringent, but also provide enough flexibility for statistical assessment (e.g., drawing the Poisson distribution in the OMSSA algorithm)
    - 0.5 Da is commonly used for data recorded by ion traps (e.g. LTQ)

Charge State and Missed Cleavages

Charge state

- Frequently, the mass spectrometer is set to only fragment features with charge > 1
- If you know your data is restricted to several charge states (e.g., for your mass spectrometric settings), you can save time by not looking at these

Missed cleavages

- Sometimes, proteases don’t cleave perfectly
- 1 or 2 missed cleavages should be allowed, but be careful since the number of missed cleavages increases your search space sizes!

Modifications

The modification settings mostly depend on the sample preparation

Fixed modifications

- **Carbamidomethylation of cysteins** is used as fixed modification in most experiments, since proteins are usually subjected to a DL-Dithiothreitol (DTT) treatment to reduce disulfide bonds built by cysteins. To protect the liberated –SH the samples are treated with iodoacetamide. This leads to a stable modification of cysteins

$$\text{-CH}_2\text{-SH} \quad + \quad \text{I-CH}_2\text{-CONH}_2 \quad \text{->} \quad \text{-CH}_2\text{- S-CH}_2\text{-CONH}_2$$

CYS residue Iodacetamide S-caramidomethylated CYS

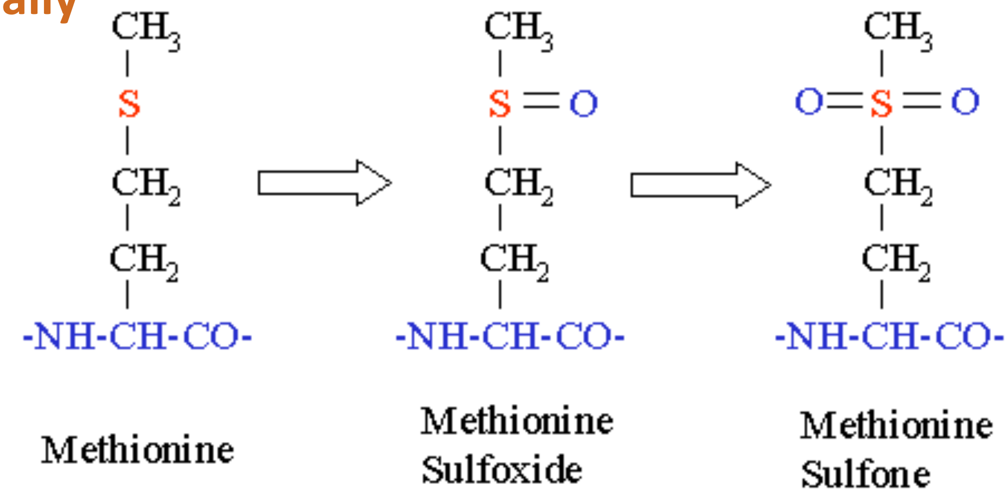
- A fixed modification on amino acid X replaces the original amino acid X during database search

Modifications

The modification settings mostly depend on the sample preparation

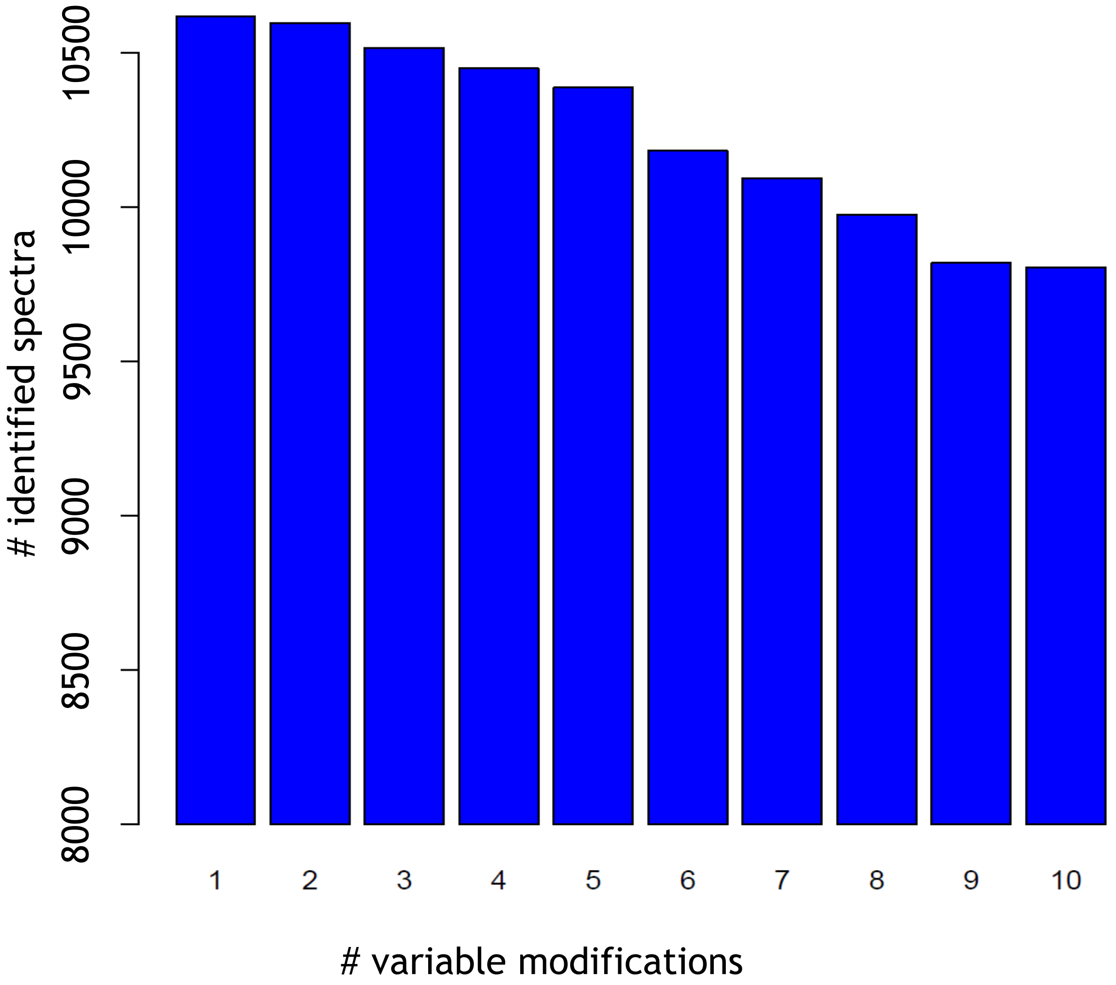
Variable modifications

- Variable modifications should be set if you know that a subset of the amino acids are modified. Routinely oxidation of methionine should be set as variable modifications. During the electrospray ionization Met residues frequently react with the oxygen in the ionization source environment
- Note that variable modifications are considered as additional amino acids and **impact search space size drastically**



http://ionsource.com/Card/MetOx/metox.1.gif

Variable Modifications



Intuitively…

- More variable modifications should discover more peptides
- Large parts of the proteome are modified

However…

- More 'amino acids': increase the search space (combinatorial explosion)
- Loss in sensitivity
- Variable modifications need to carefully chosen

Materials

- Online Materials
  - Learning Unit 3B (statistics for FDR)
  - Learning Unit 7A, B
- Slides on peptide ID by Brian Searle
  - https://proteome-software.wikispaces.com/file/view/interpreting-MS-MS-proteomics-results.ppt

References

- Eidhammer et al., Computational Methods for Mass Spectrometry Proteomics. Wiley. 2007.
- Freitas and Xu, BMC Bioinformatics. 2010, 11:436
- Roepstorff and Fohlman, Biological Mass Spectrometry, Volume 11, Issue 11, page 601, November 1984
- Steen and Mann. Nature Reviews, Molecular Cell Biology, Vol. 5 2004
- Johnson et al. Anal. Chem 1987;59:2621-2625
- Hoffert J D et al. PNAS 2006;103:7159-7164
- Craig,R. and Beavis,R.C. (2003) Rapid Commun.
  Mass Spectrom., 17, 2310–2316
- Geer et al. (2004) J Proteome Res. 2004 Sep-Oct;3(5):958-64.
- Eng et al., *J. Am. Soc. Mass Spectrom*. 1994, 5, 976-989.
- Fenyö and Beavis, Anal. Chem.2003, 75, 768-774
- http://www.proteomesoftware.com/pdf_files/XTandem_edited.pdf
- Grenzel et al, Proteomics. 2003(3):1597-1610.
- Elias and Gygi, Nature Methods. Vol. 4, No. 3, March 2007
- Searle et al., Journal of Proteome Research. *2008, 7, 245–253* **245**