bioinformatics for proteomics

lennart martens

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

www.compomics.com
@compomics

Introduction: MS/MS spectra and identification

Database search algorithms

Sequencial search algorithms

Notable caveats and painful disasters

Identification validation

Protein inference: bad, ugly, and not so good

Introduction: MS/MS spectra and identification

Database search algorithms

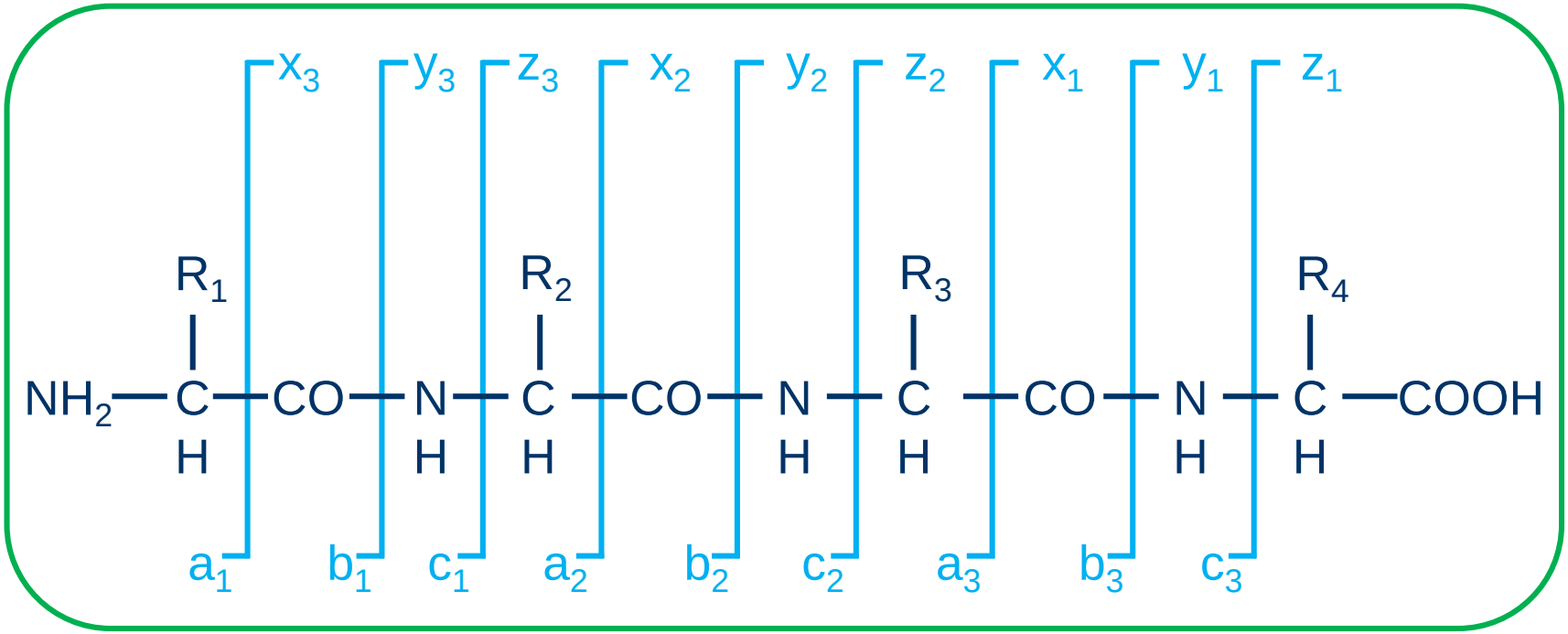
Sequencial search algorithms

Notable caveats and painful disasters

Identification validation

Protein inference: bad, ugly, and not so good

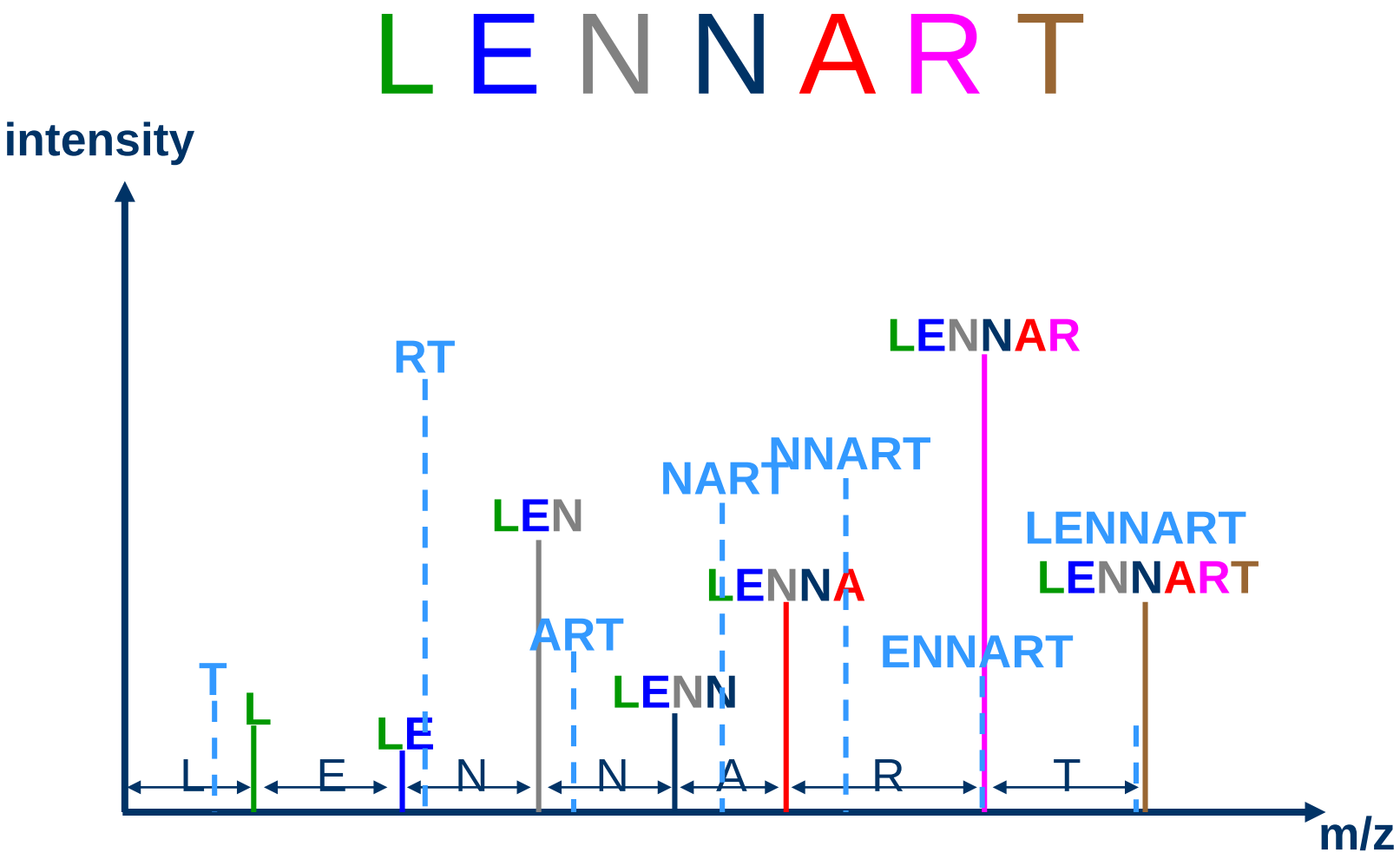
Peptides subjected to fragmentation analysis can yield several types of fragment ions



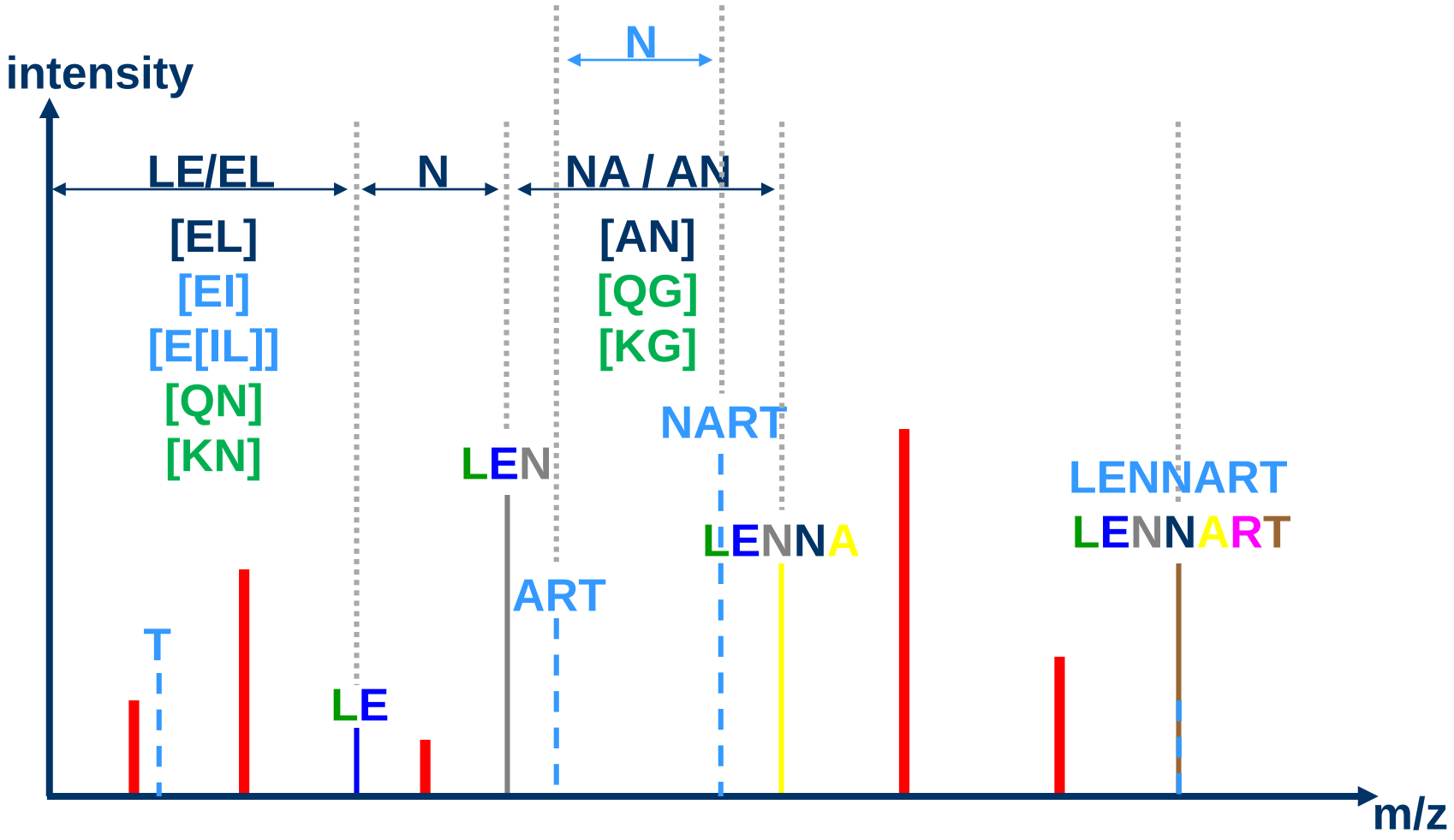
There are several other ion types that can be annotated, as well as 'internal fragments'. The latter are fragments that no longer contain an intact terminus. These are harder to use for 'ladder sequencing', but can still be interpreted.

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



Real spectra usually look quite a bit worse



We can distinguish three types of M/MS identification algorithms

Spectral comparison

Sequencial comparison

Threading comparison

Eidhammer, Wiley, 2007

Introduction: MS/MS spectra and identification

Database search algorithms

Sequencial search algorithms

Notable caveats and painful disasters

Identification validation

Protein inference: bad, ugly, and not so good

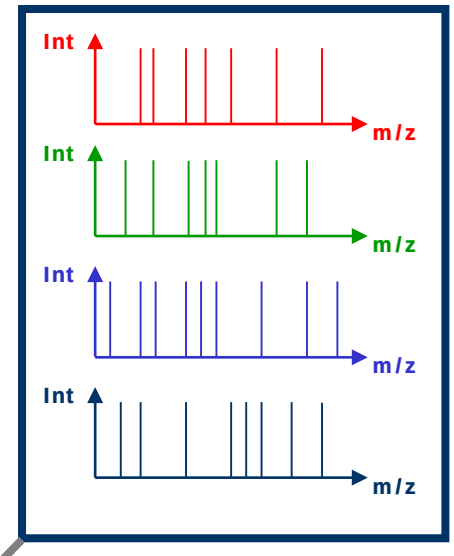
Database search engines match experimental spectra to known peptide sequences

protein sequence database

in silico digest

peptide sequences

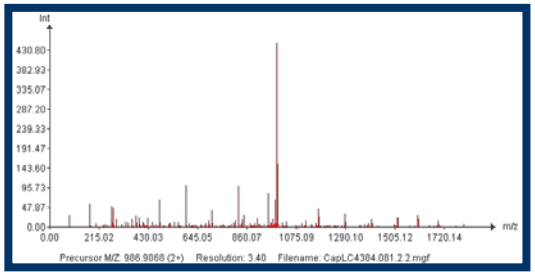
in silico MS/MS



theoretical MS/MS spectra

peptide scores

in silico matching



experimental MS/MS spectrum

Three popular algorithms can serve as templates for the large variety of tools

- SEQUEST (UWashington, Thermo Fisher Scientific)
  http://fields.scripps.edu/sequest
- MASCOT (Matrix Science)
  http://www.matrixscience.com
- X!Tandem (The Global Proteome Machine Organization)
  http://www.thegpm.org/TANDEM

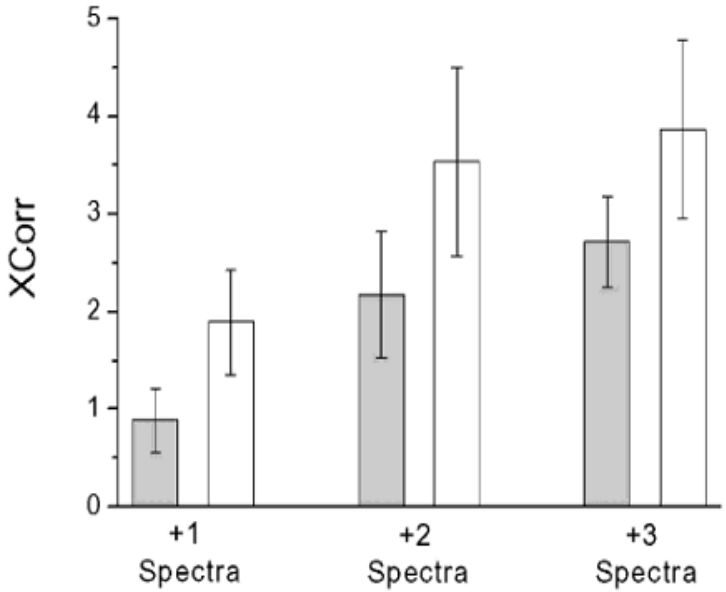
SEQUEST is the original search engine, but not that much used anymore these days

- 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

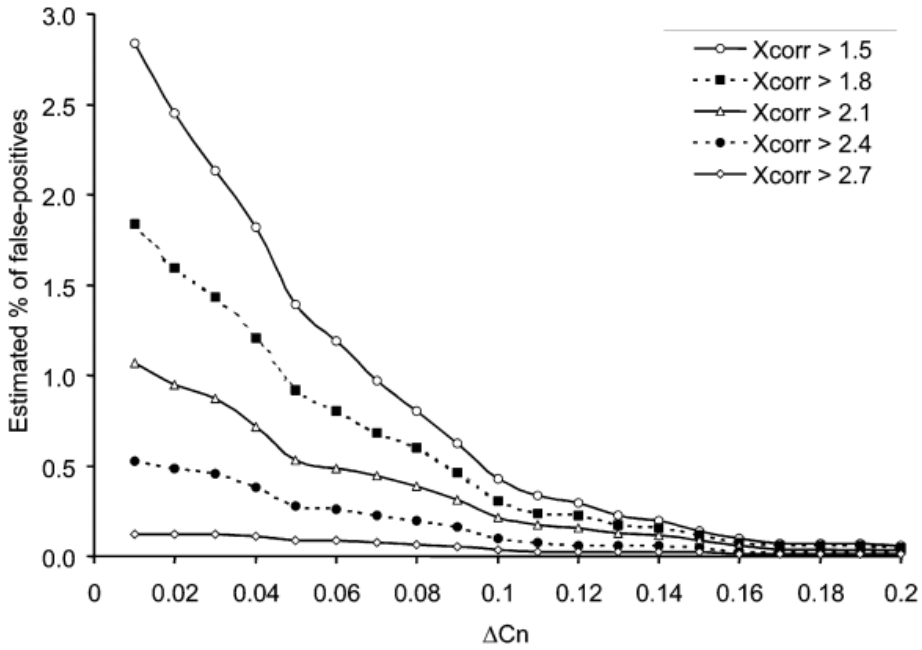
$$R_i = \sum_{j=1}^{n} x_j \cdot y_{(j+i)}$$

$$\text{XCorr} = R_0 - \frac{1}{151}\left(\sum_{i=-75}^{+75} Ri\right) \qquad \text{deltaCn} = \frac{XCorr_1 - XCorr_2}{XCorr_1}$$

SEQUEST reveals the problems with scoring different charges, and using different scores



From: MacCoss et al., Anal. Chem. 2002



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

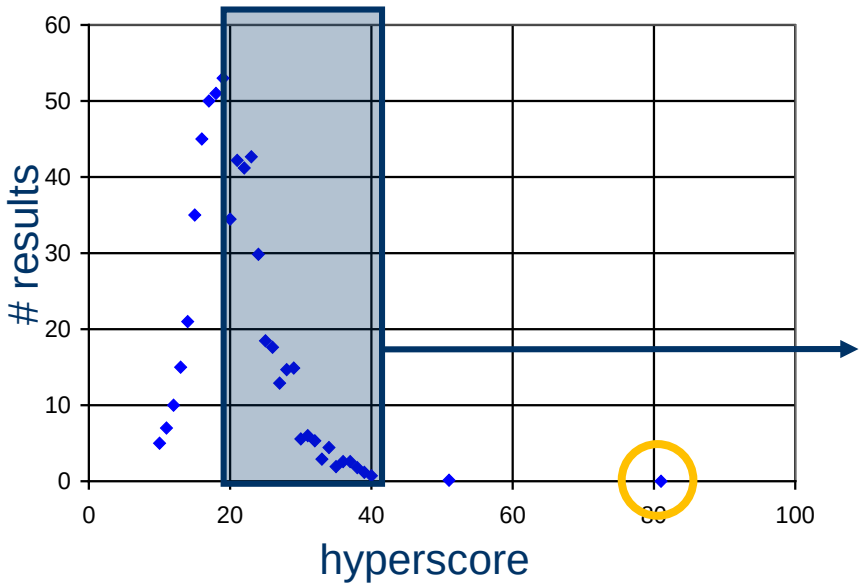
Mascot is probably the most recognized search engine, despite its secret algorithm

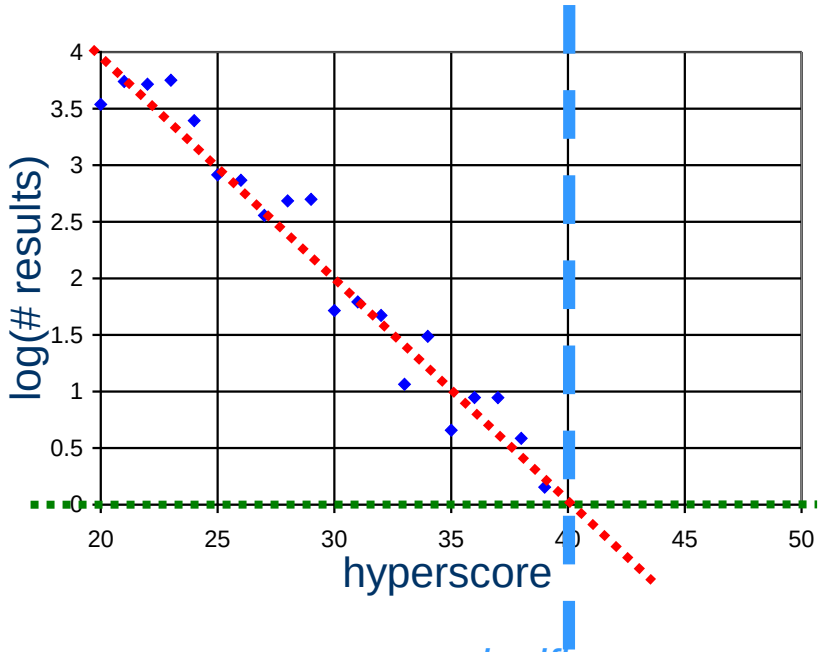
- Very well established search engine, Perkins, *Electrophoresis* 1999
- Can do MS (PMF) and MS/MS (PFF) identifications
- Based on the MOWSE score,
- Unpublished core algorithm (trade secret)
- Predicts an *a priori* threshold score that identifications need to pass
- From version 2.2, Mascot allows integrated decoy searches
- Provides rank, score, threshold and expectation value per identification
- Customizable confidence level for the threshold score

X!Tandem is a clear front-runner among open source search engines

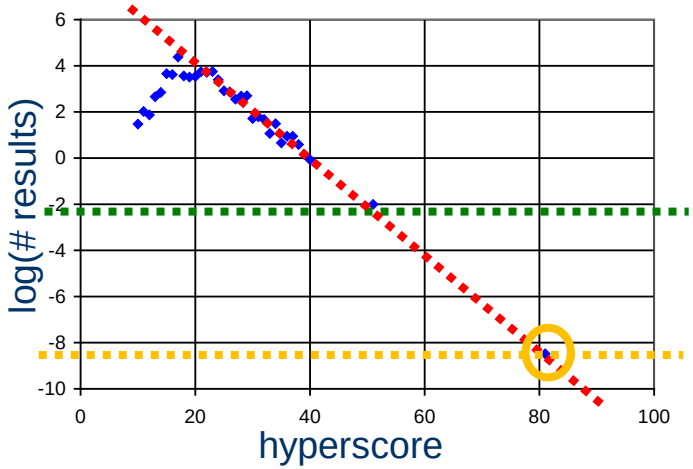
- A successful open source search engine, Craig and Beavis, *RCMS* 2003
- Can be used for MS/MS (PFF) identifications
- Based on a hyperscore (*Pi* is either 0 or 1): $HyperScore = \left(\sum_{i=0}^{n} I_i * P_i \right) * N_b! * N_y!$
- Relies on a hypergeometric distribution (hence hyperscore)
- Published 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





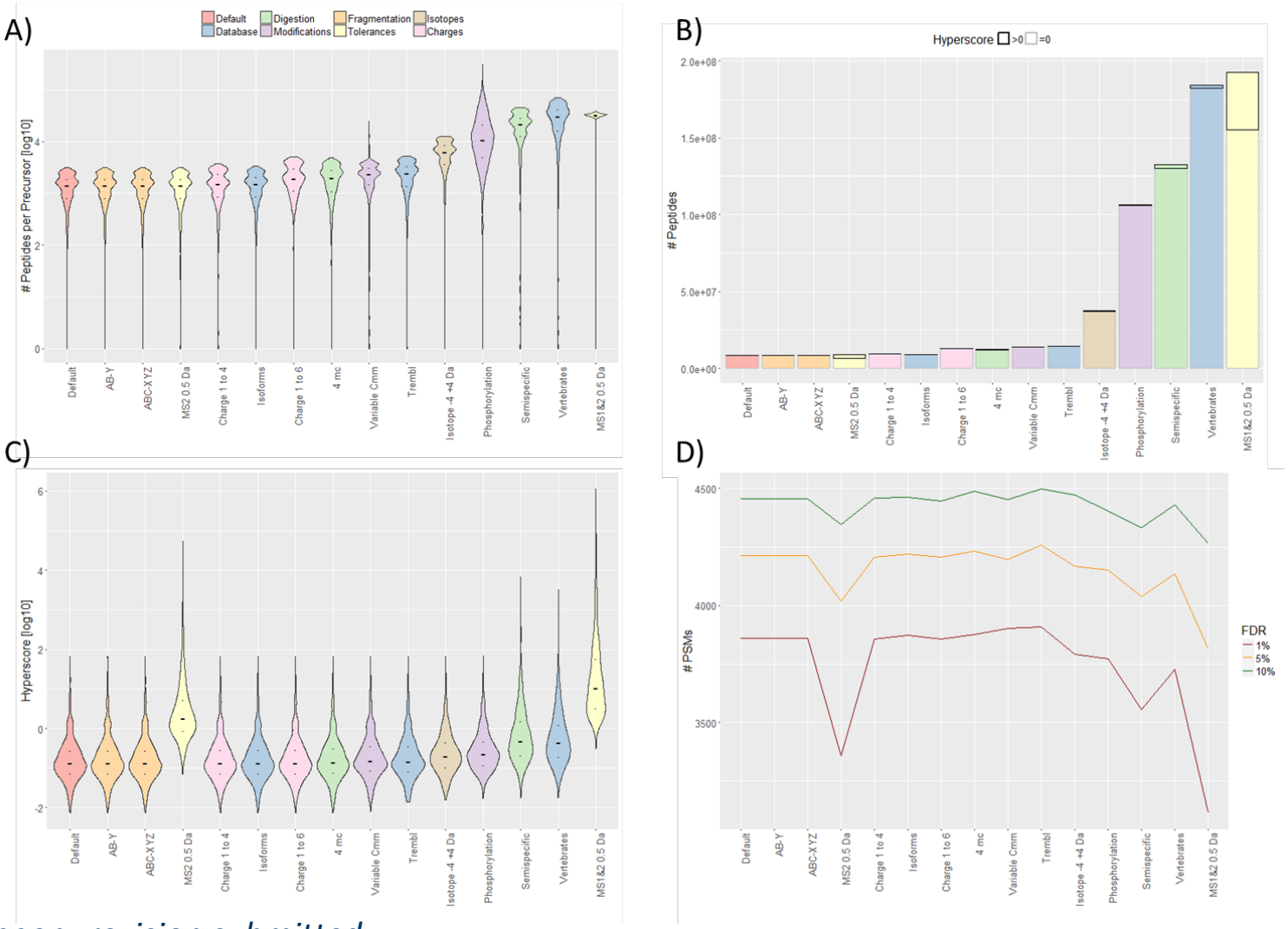
significance threshold



E-value=$e^{-8.2}$

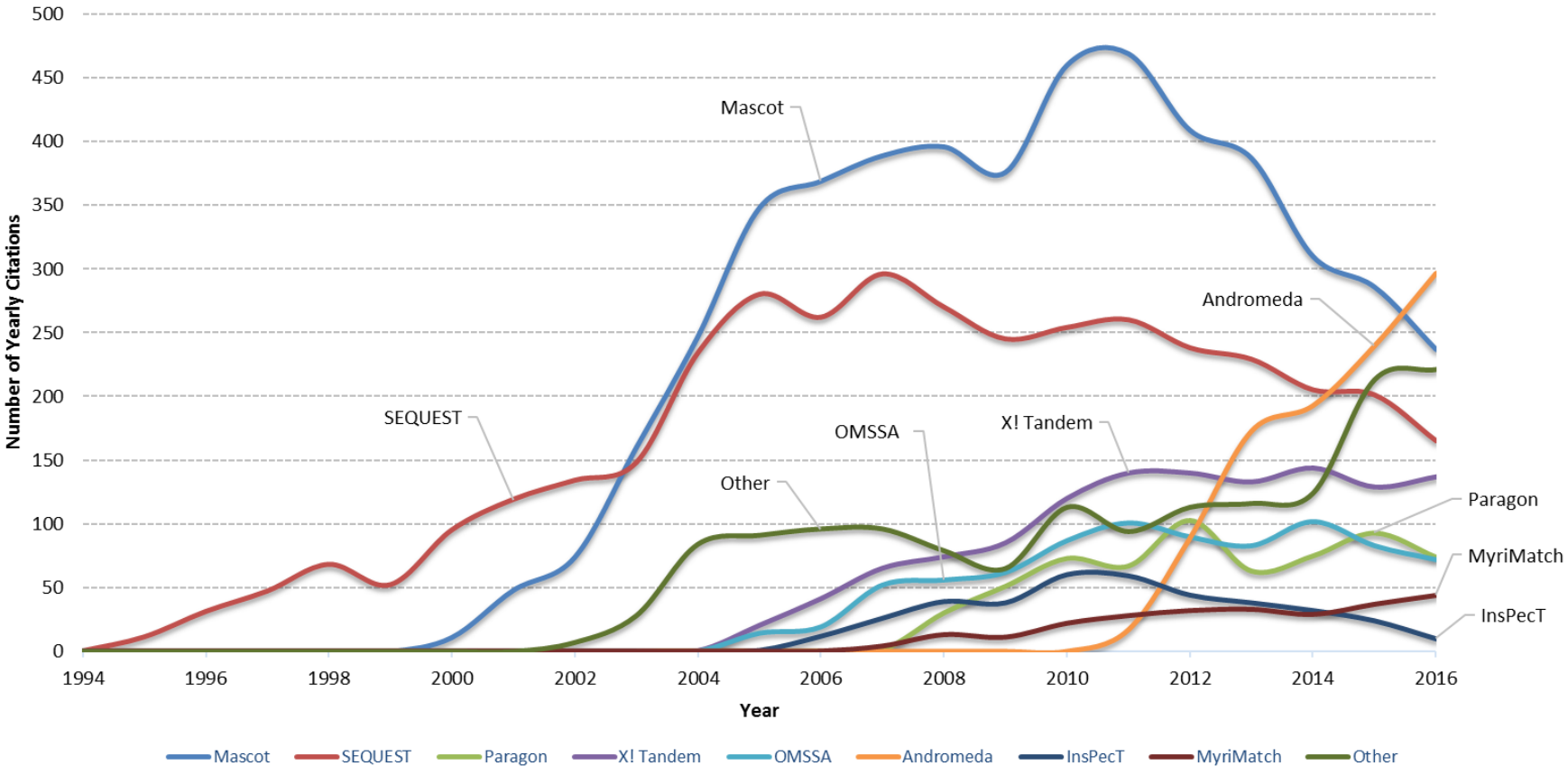
***Adapted from*: Brian Searle, ProteomeSoftware, http://www.proteomesoftware.com/XTandem_edited.pdf**

The influence of various parameter changes is clearly visible (here for X!Tandem)



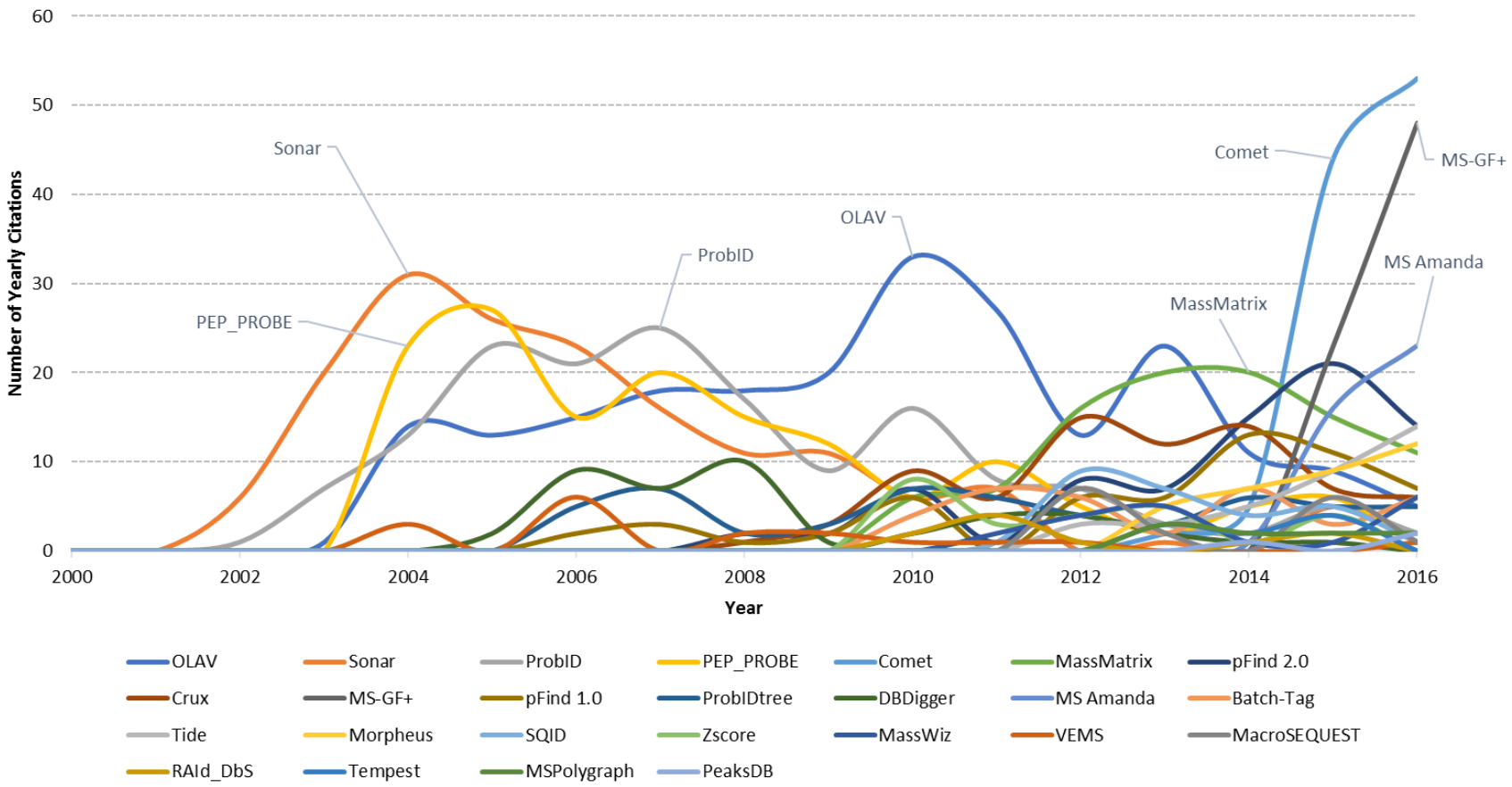
Verheggen, *revision submitted*

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



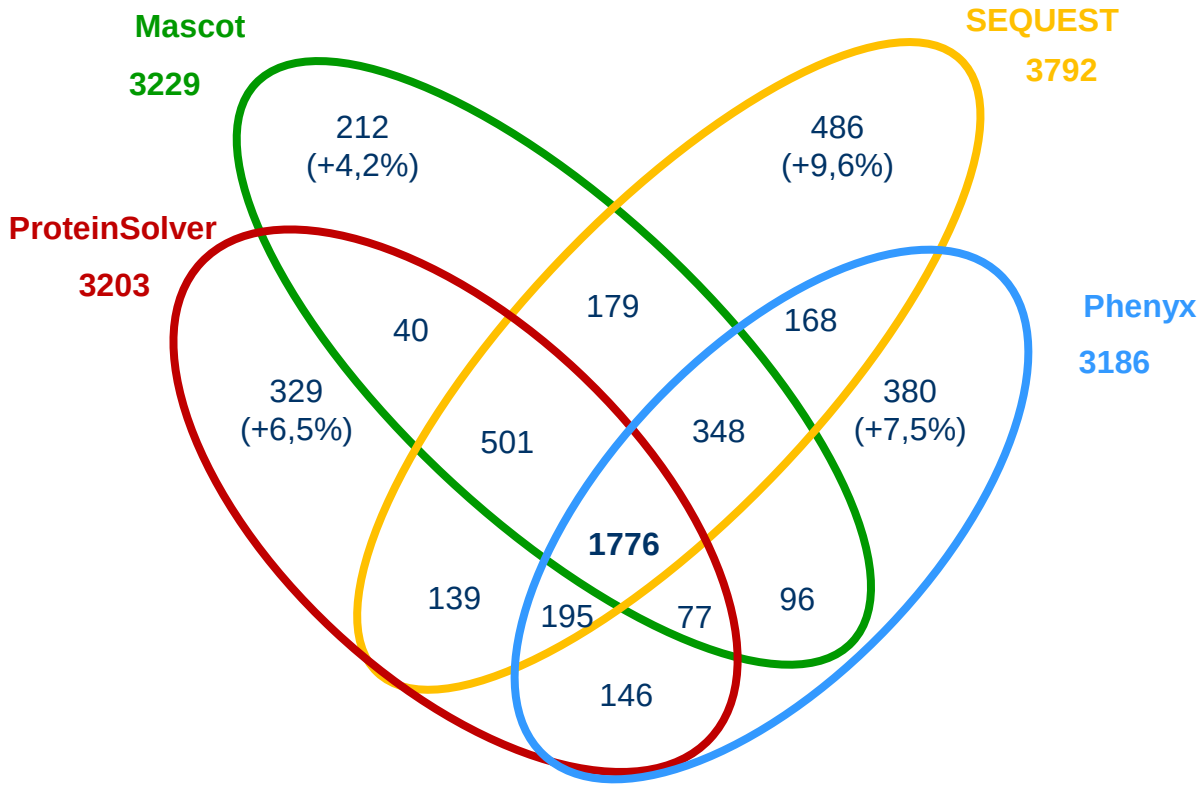
Verheggen, *revision submitted*

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



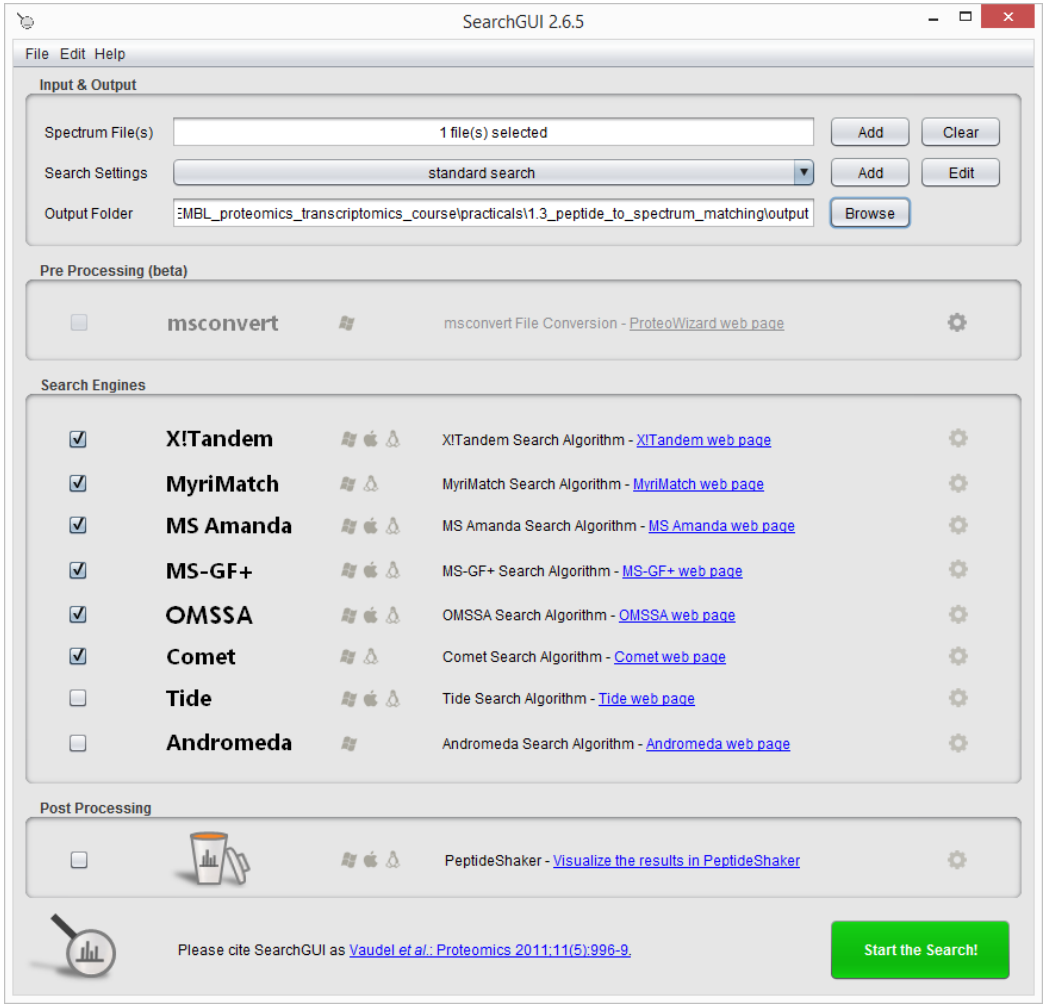
Verheggen, *revision submitted*

Because of their unique biases and sensitivity, combining search algorithms can be useful



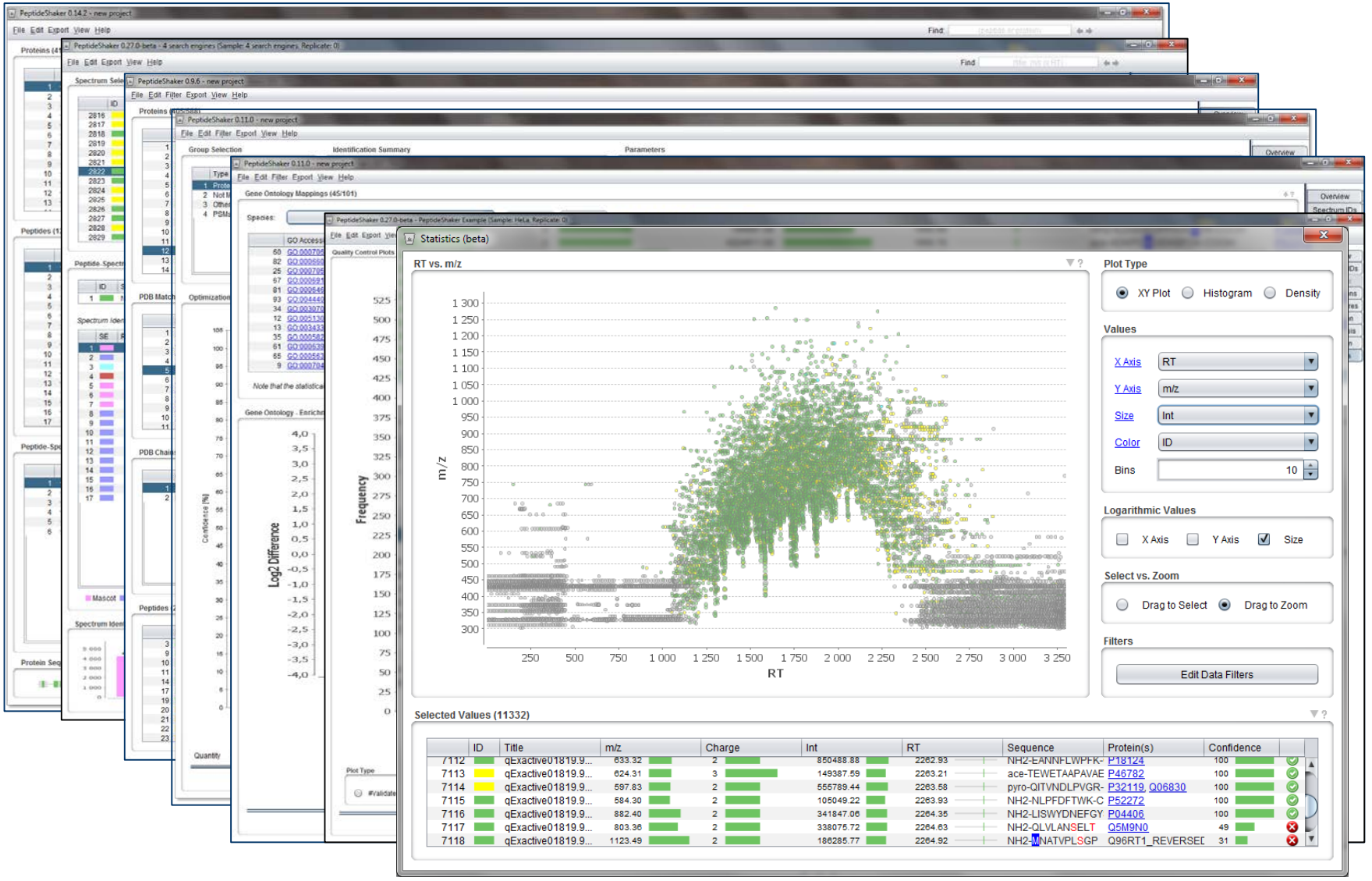
Numbers courtesy of Dr. Christian Stephan, then at Medizinisches Proteom-Center, Ruhr-Universität Bochum; Human Brain Proteome Project

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

Introduction: MS/MS spectra and identification

Database search algorithms

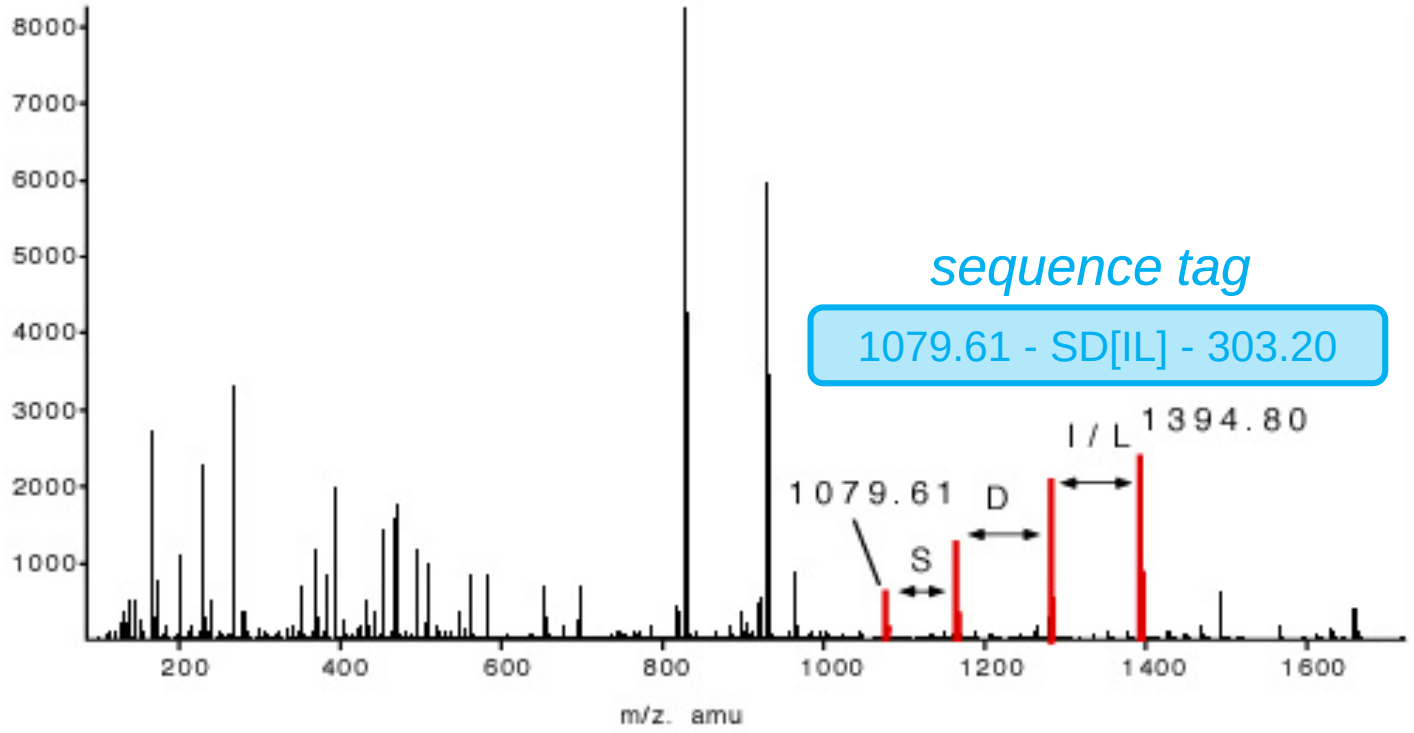
Sequencial search algorithms

Notable caveats and painful disasters

Identification validation

Protein inference: bad, ugly, and not so good

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



The concept of sequence tags was introduced by Mann and Wilm

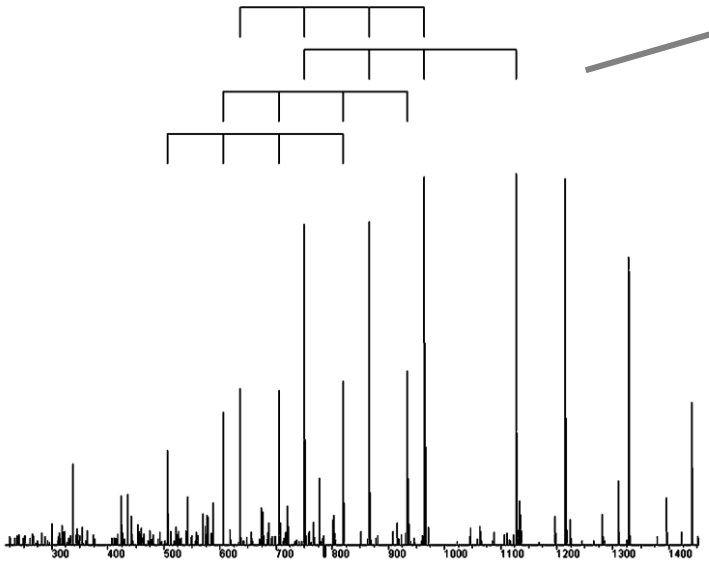
Mann, Analytical Chemistry, 1994

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 and DirecTag extracts tags,
- TagRecon matches these to the database
- Very useful to retrieve unexpected peptides (modifications, variations)
- Entire workflows exist (e.g., combination with IDPicker)

GutenTag: two stage, hybrid tag searching

1. Generate sequence tags



2. Search DB for matches

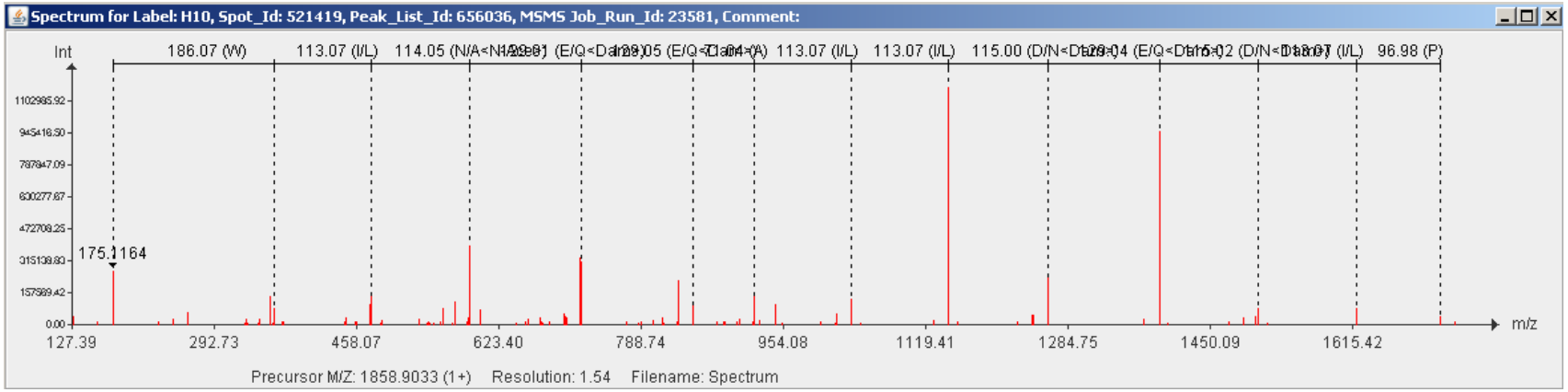
DDG → -DDGNSDRS
YVD → -YVDVNKFKD
VDD → KLLSYVDDEAFIR
DDE → EGDEANSDDEEEDL
DDV → -DDVDIDEN
VVD → SSCTAVVD-
DVY → AFQYLKDVY-

3. Score DB Sequences

KLLSYVDDEAFIR	19.36
-DDVDIDEN	8.56
-DDGNSDRS	6.94
-YVDVNKFKD	6.25
SSCTAVVD-	5.74
EGDEANSDDEEEDL	5.64
AFQYLKDVY-	5.61

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!

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

Introduction: MS/MS spectra and identification

Database search algorithms

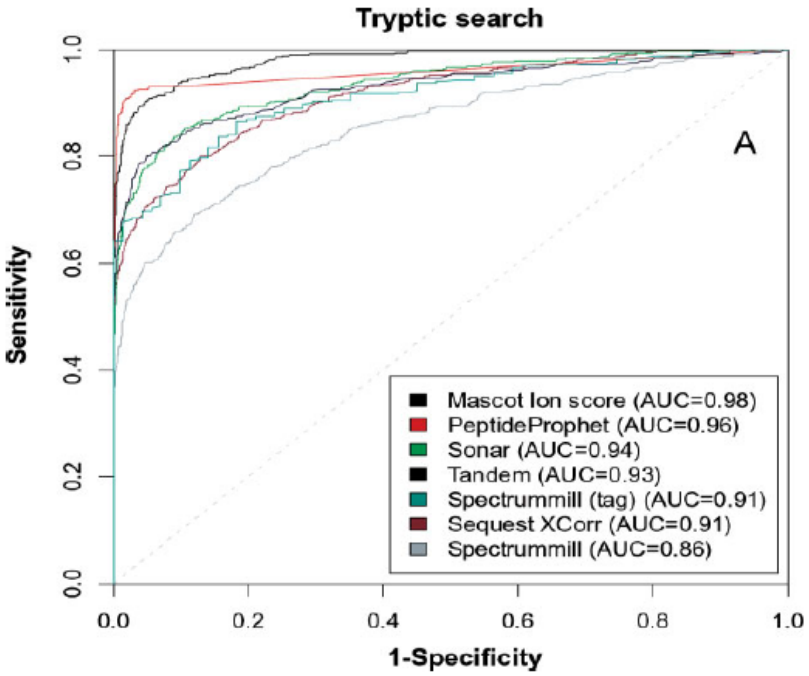
Sequencial search algorithms

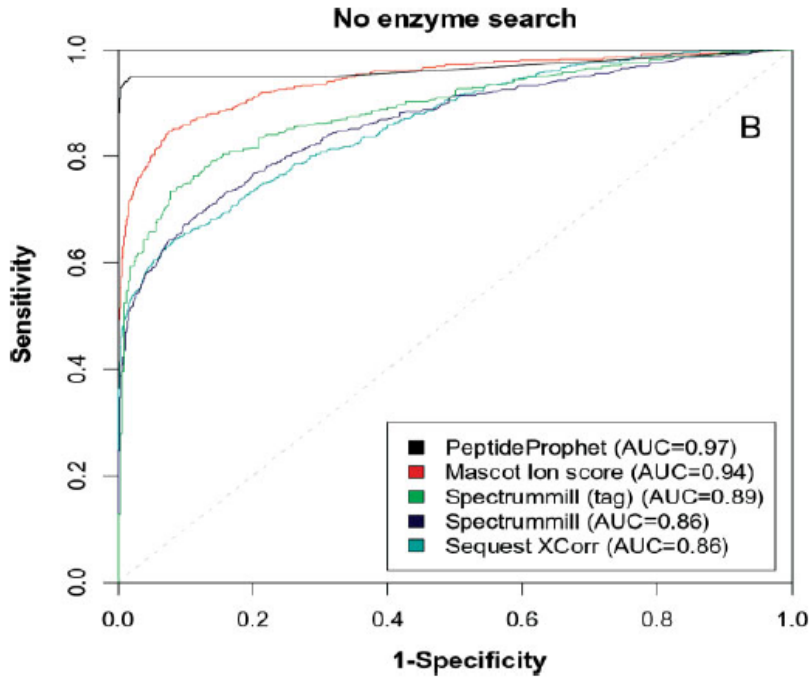
Notable caveats and painful disasters

Identification validation

Protein inference: bad, ugly, and not so good

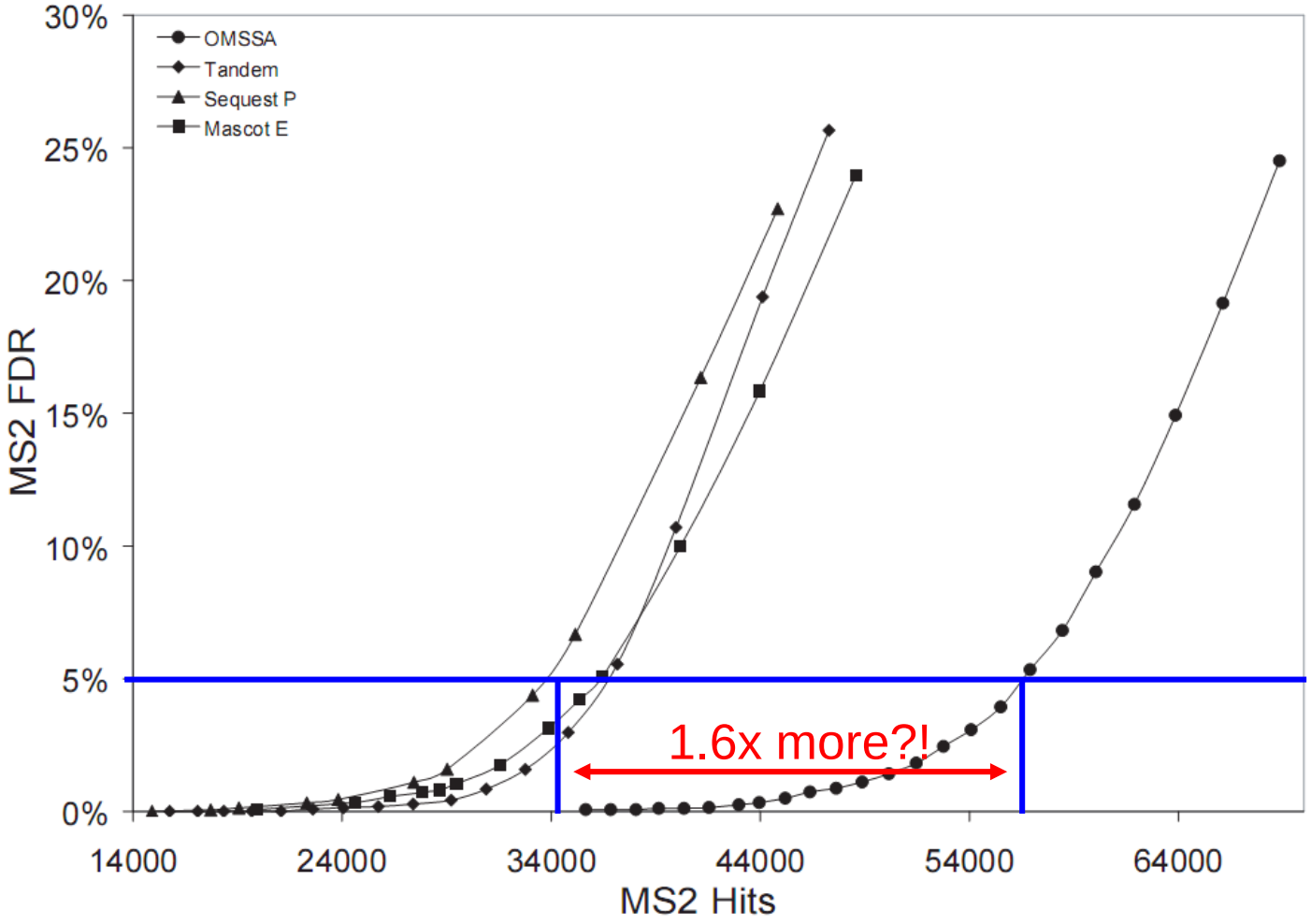
Comparison of search engines shows a difference in underlying assumptions





Kapp, Proteomics, 2005

Some comparisons are just dead wrong, regardless of where they are published



Balgley, MCP, 2007

Colony colapse disorder, soldiers, and forcing the issue (or rather: the solution)

The New York Times

Science

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION | AR

ENVIRONMENT | SPACE & COSMOS

Scientists and Soldiers Solve a Bee Mystery

Mike Albans for The New York Times

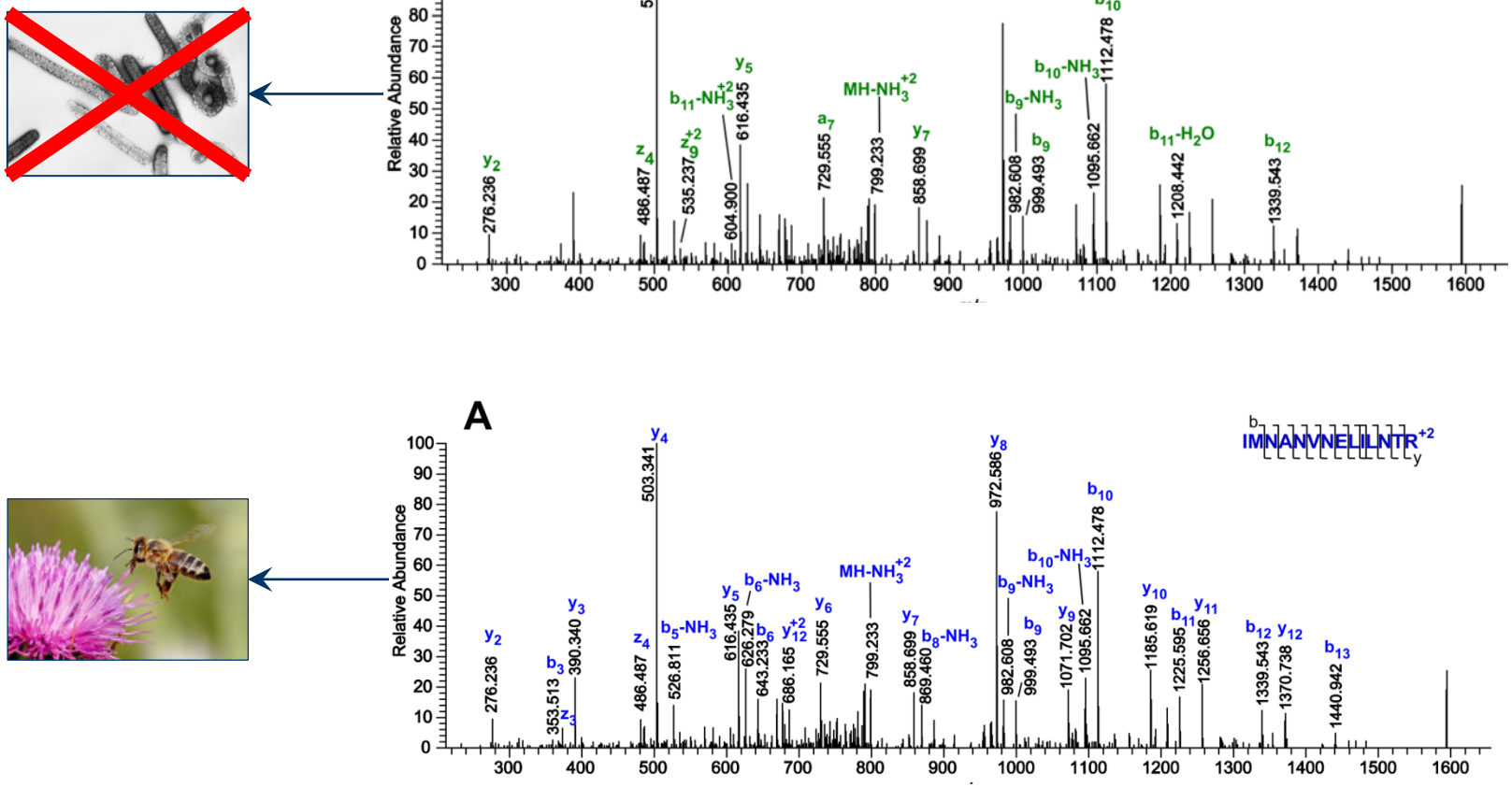
Members of a joint United States Army-University of Montana research team that located a virus that is possibly collapsing honeybee colonies scanning a healthy hive near Missoula, Mont.

By KIRK JOHNSON

Published: October 6, 2010

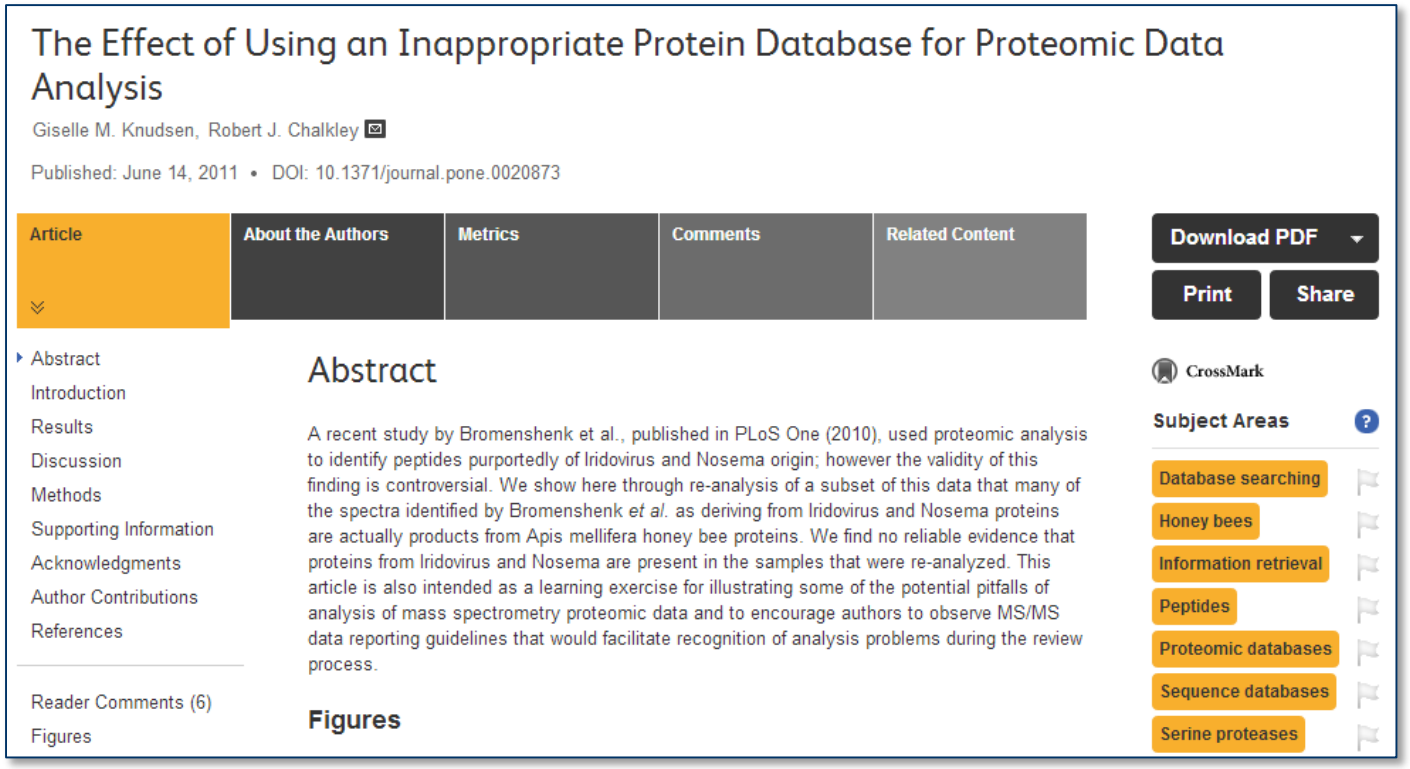
Knudsen, PLoS ONE, 2011

The identification seems reasonable, if limited in an unreasonable way



Knudsen, PLoS ONE, 2011

The end result may be that you are taken to task for mistakes in your research



The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis

Giselle M. Knudsen, Robert J. Chalkley

Published: June 14, 2011 • DOI: 10.1371/journal.pone.0020873

Article | About the Authors | Metrics | Comments | Related Content

Download PDF | Print | Share

- Abstract
- Introduction
- Results
- Discussion
- Methods
- Supporting Information
- Acknowledgments
- Author Contributions
- References

Reader Comments (6)

Figures

Abstract

A recent study by Bromenshenk et al., published in PLoS One (2010), used proteomic analysis to identify peptides purportedly of Iridovirus and Nosema origin; however the validity of this finding is controversial. We show here through re-analysis of a subset of this data that many of the spectra identified by Bromenshenk *et al.* as deriving from Iridovirus and Nosema proteins are actually products from Apis mellifera honey bee proteins. We find no reliable evidence that proteins from Iridovirus and Nosema are present in the samples that were re-analyzed. This article is also intended as a learning exercise for illustrating some of the potential pitfalls of analysis of mass spectrometry proteomic data and to encourage authors to observe MS/MS data reporting guidelines that would facilitate recognition of analysis problems during the review process.

Figures

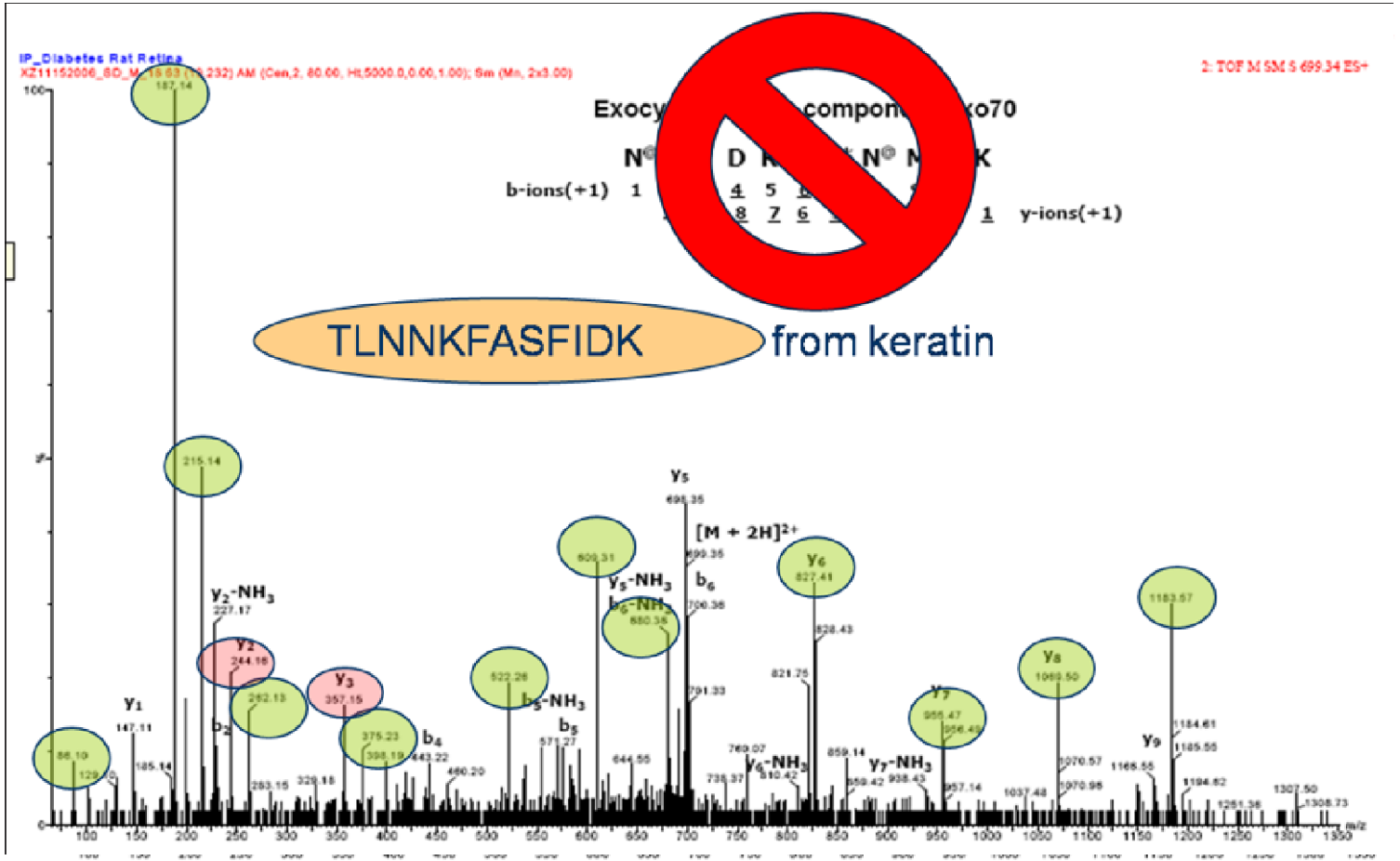
CrossMark

Subject Areas

- Database searching
- Honey bees
- Information retrieval
- Peptides
- Proteomic databases
- Sequence databases
- Serine proteases

Beware of common contaminants

Tyrosine nitrosylation



Ghesquière, Proteomics, 2010

Introduction: MS/MS spectra and identification

Database search algorithms

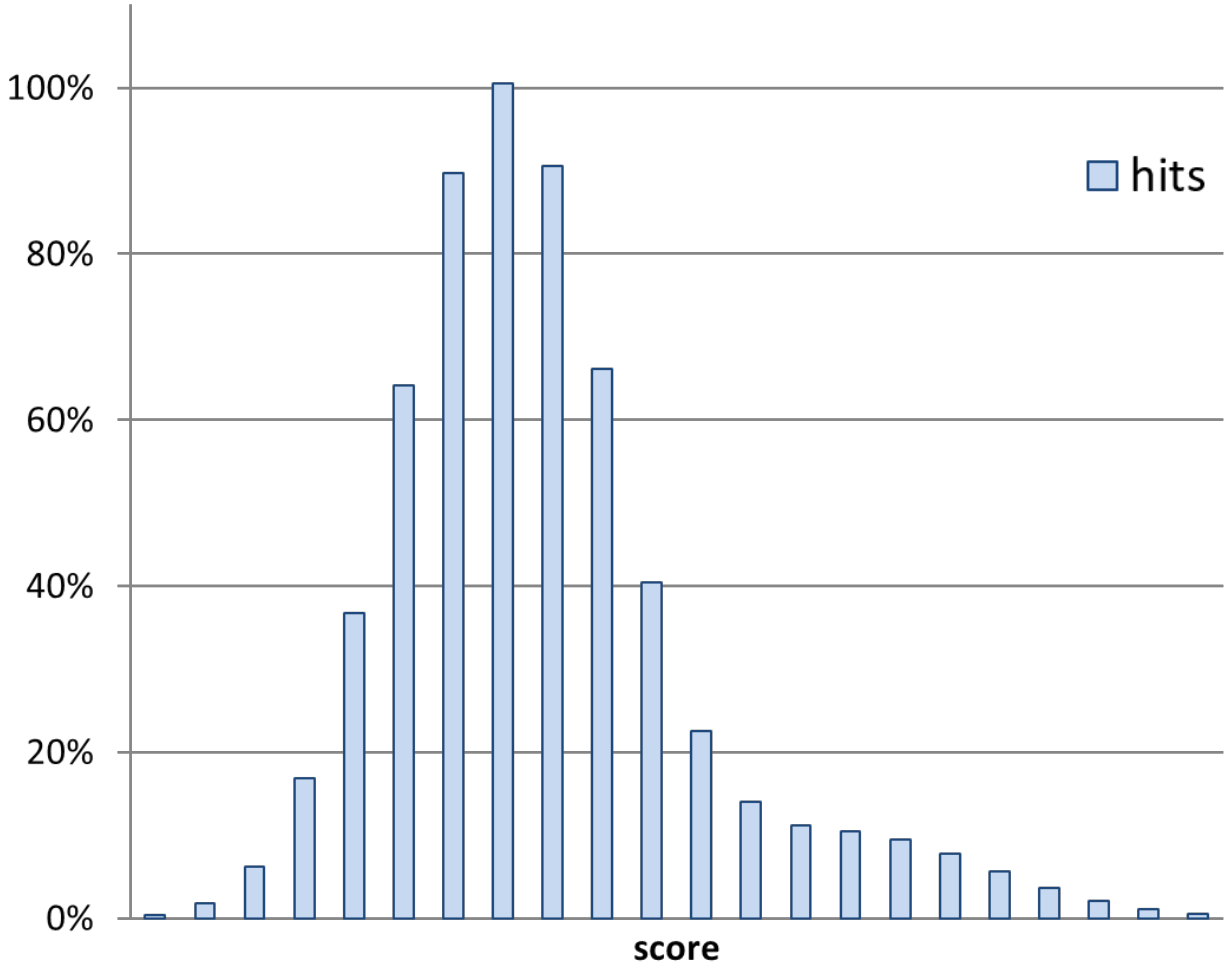
Sequencial search algorithms

Notable caveats and painful disasters

Identification validation

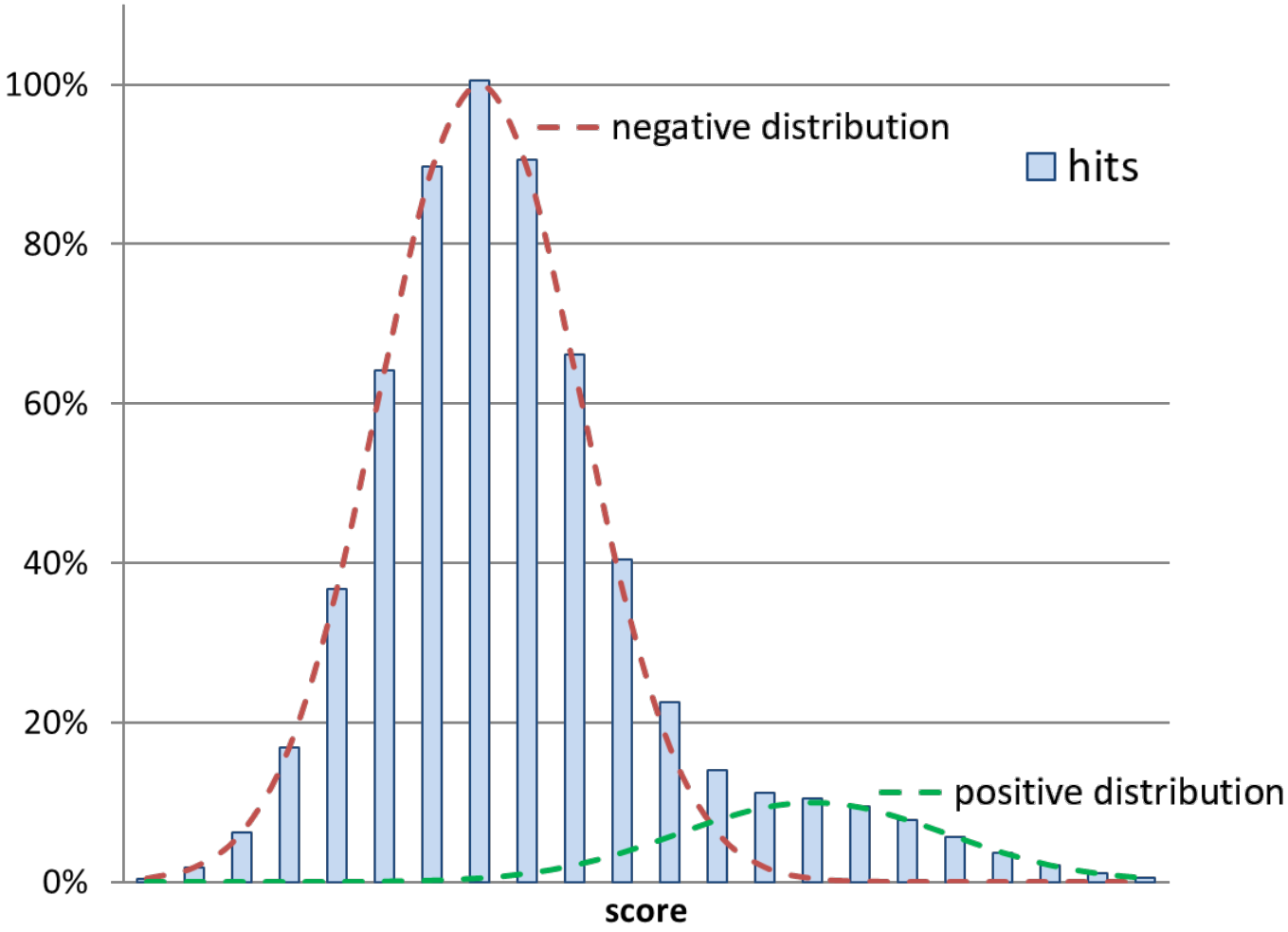
Protein inference: bad, ugly, and not so good

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

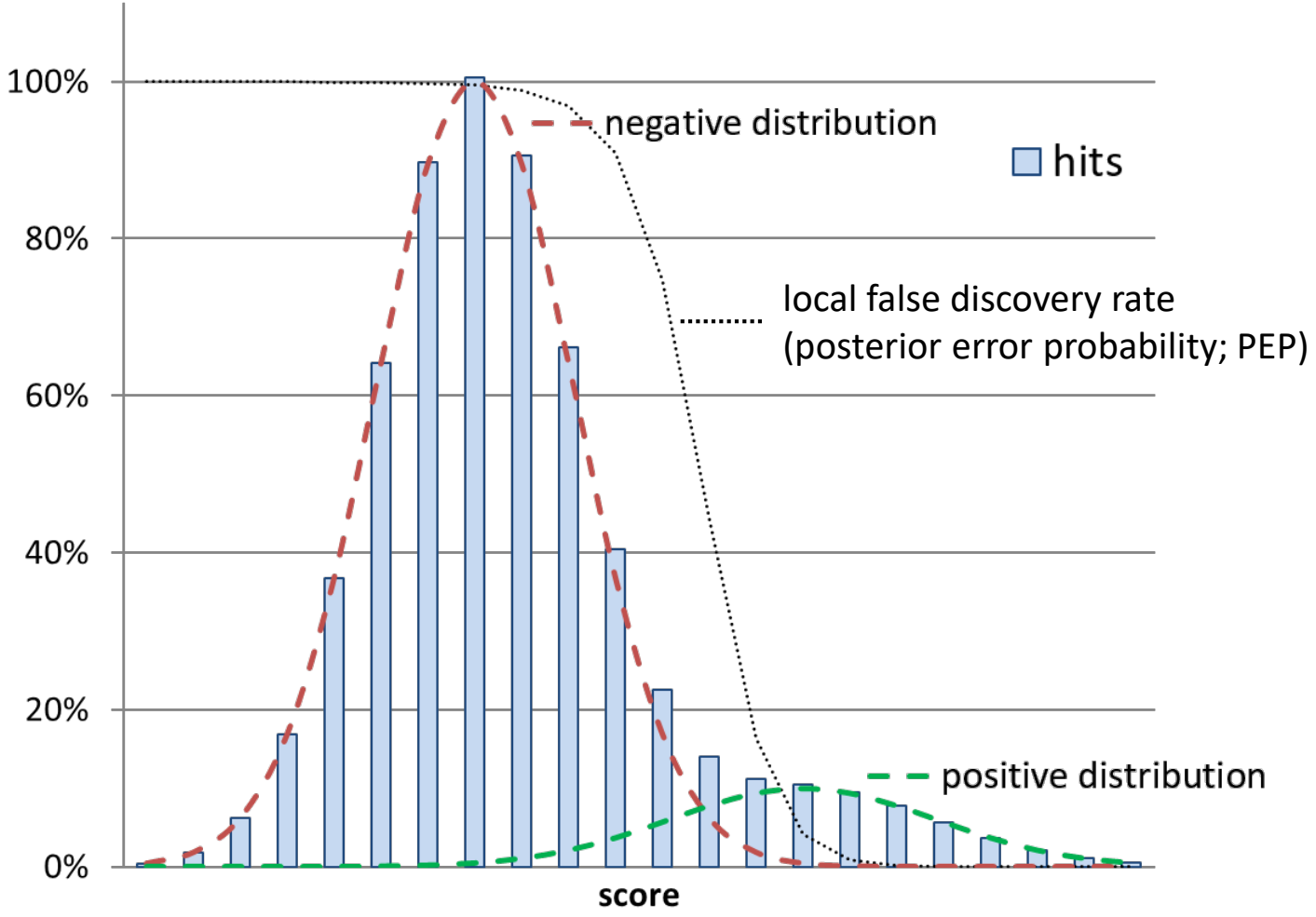


Nesvizhskii, J Proteomics, 2010

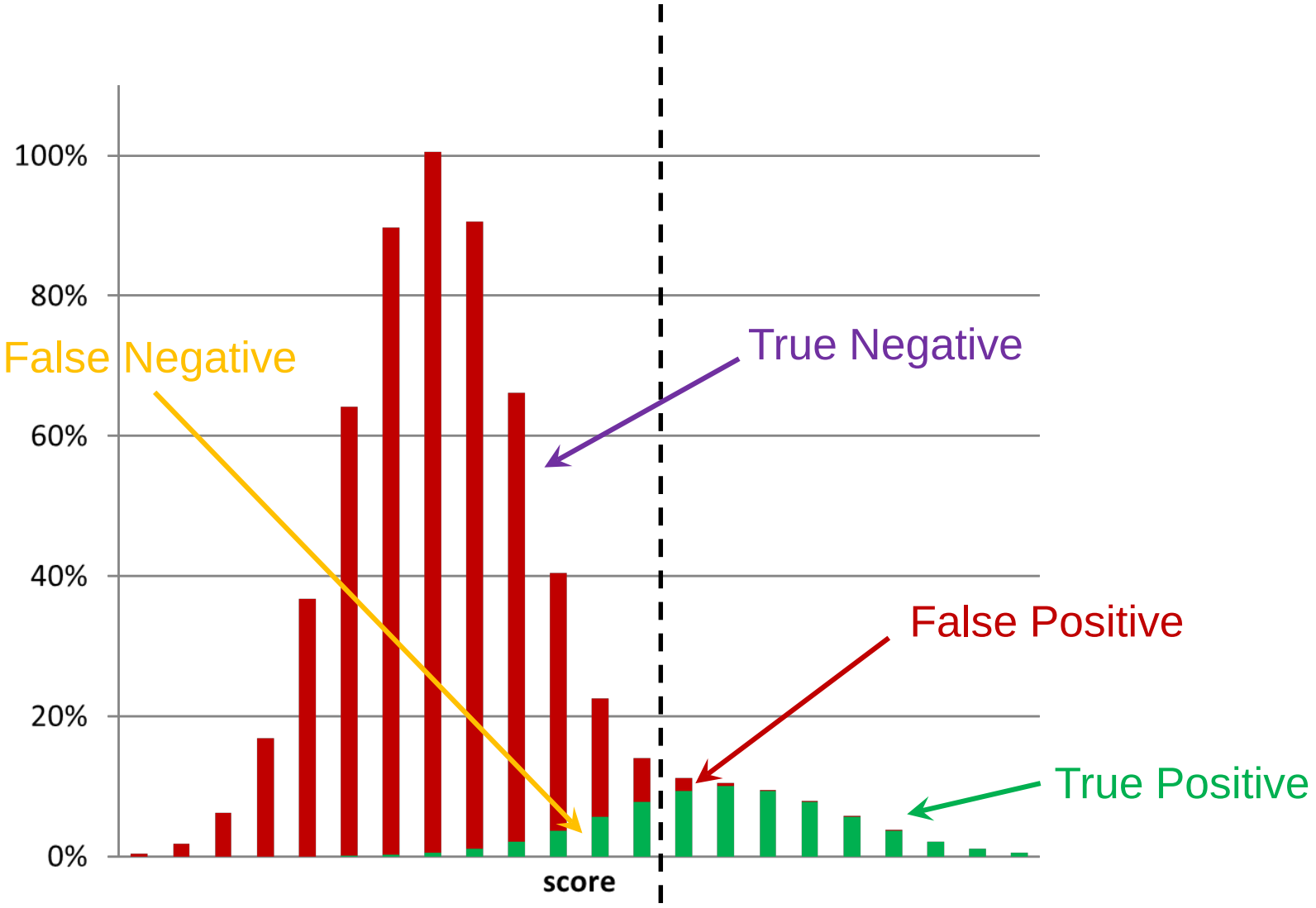
If we know how scores for bad hits distribute, we can distinguish good from bad by score



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



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



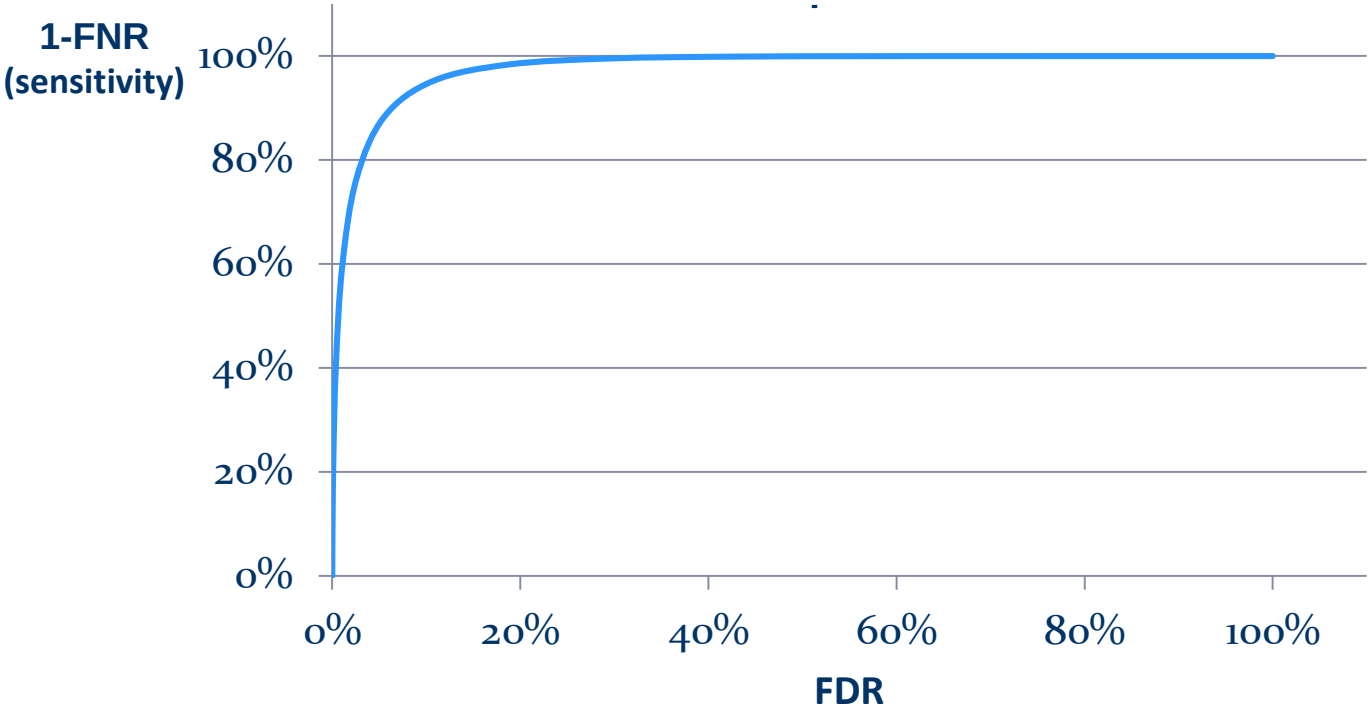
We can evaluate the effect of these errors by plotting the effect of moving the threshold

False Positive Rate ☹

$$FDR = \frac{n_{FP}}{n_{FP} + n_{TP}}$$

False Negative Rate ☹

$$FNR = \frac{n_{FN}}{n_{FN} + n_{TP}}$$



Decoy databases are false positive factories that are assumed to deliver reliably bad hits

Three main types of decoy DB's are used:

- Reversed databases (*easy*)

  LENNARTMARTENS → SNETRAMTRANNEL

- Shuffled databases (*slightly more difficult*)

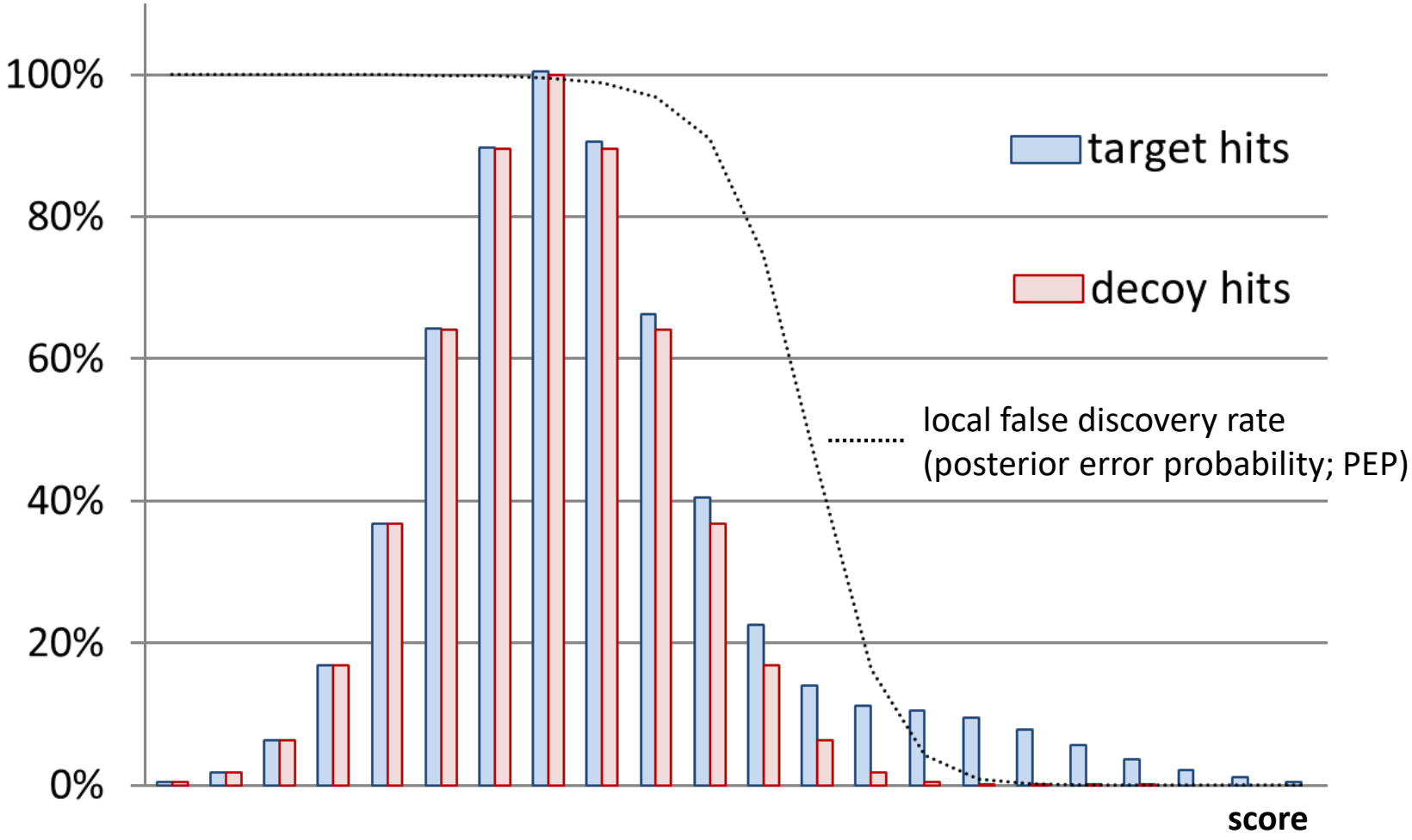
  LENNARTMARTENS → NMERLANATERTTN *(for instance)*

- 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

Introduction: MS/MS spectra and identification

Database search algorithms

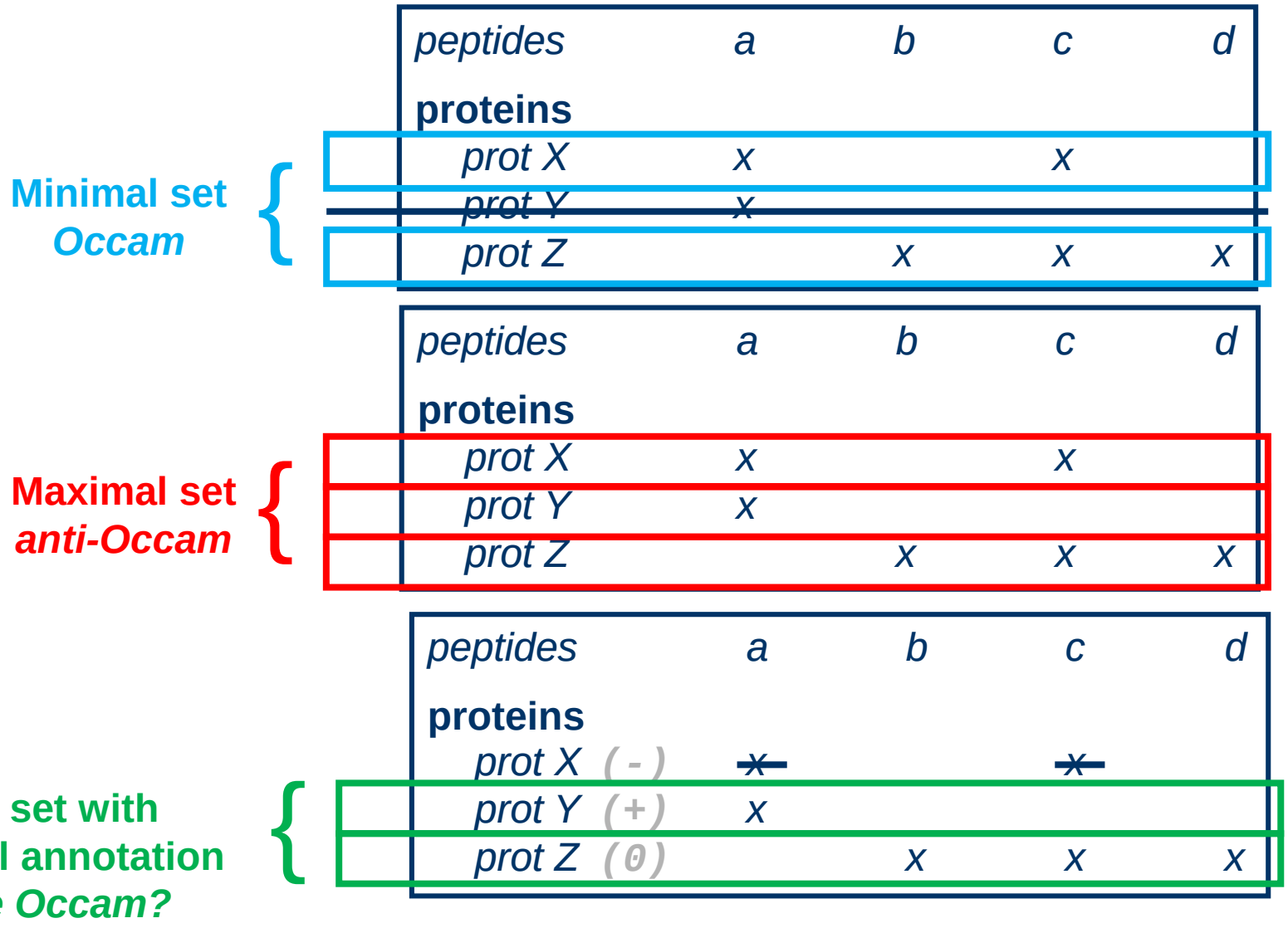
Sequencial search algorithms

Notable caveats and painful disasters

Identification validation

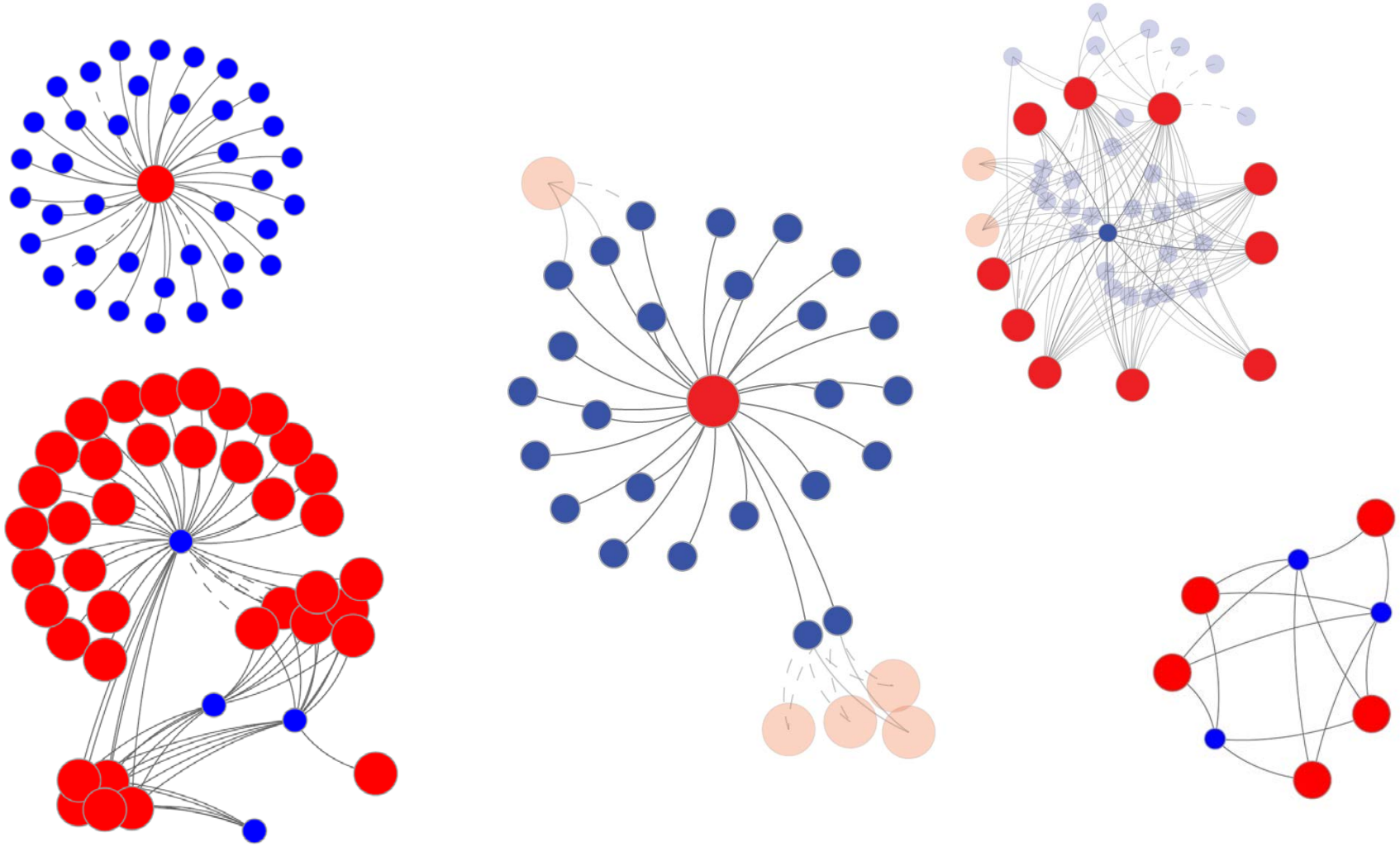
Protein inference: bad, ugly, and not so good

Protein inference is a question of conviction

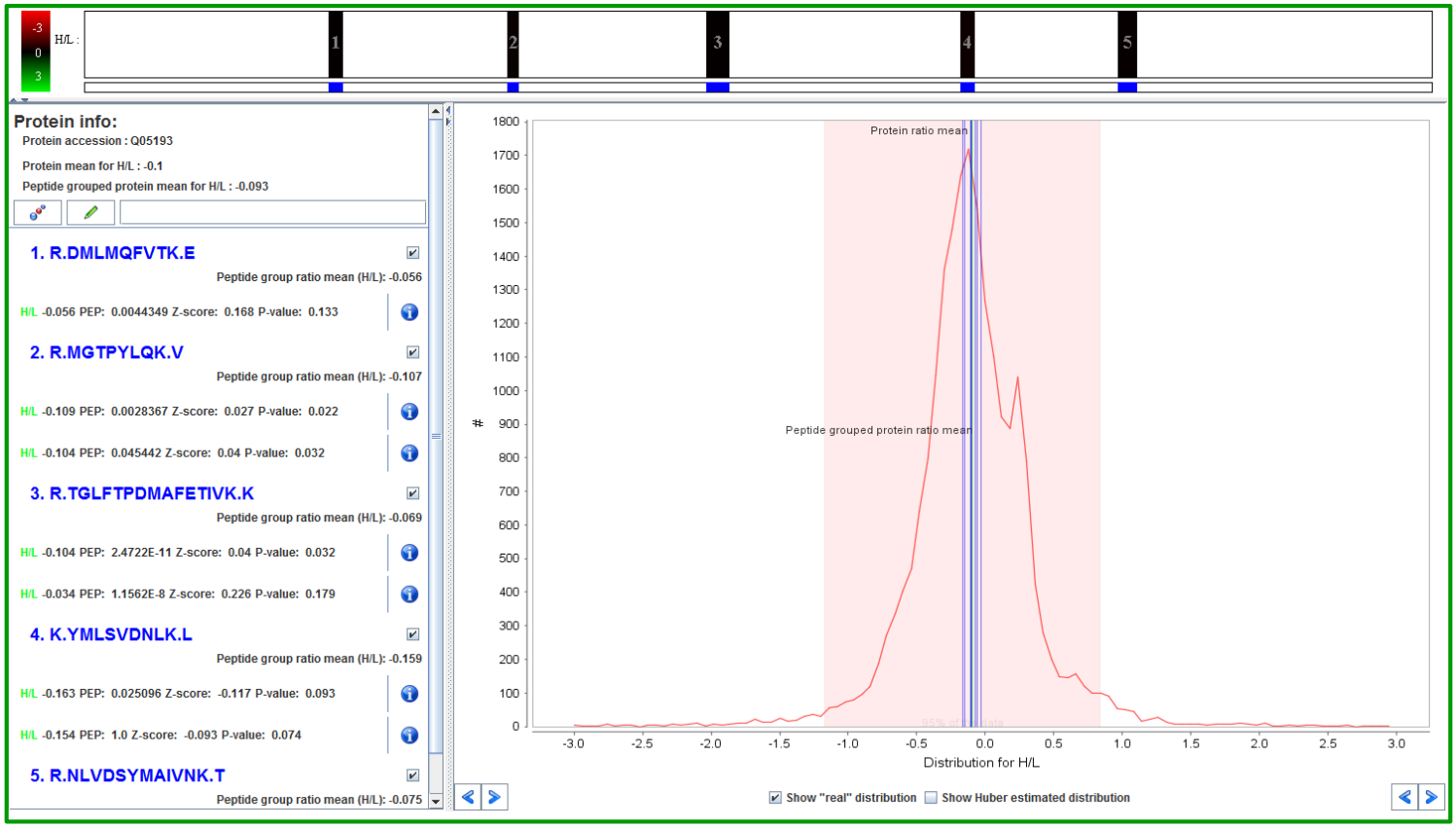


Martens, Molecular Biosystems, 2007

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



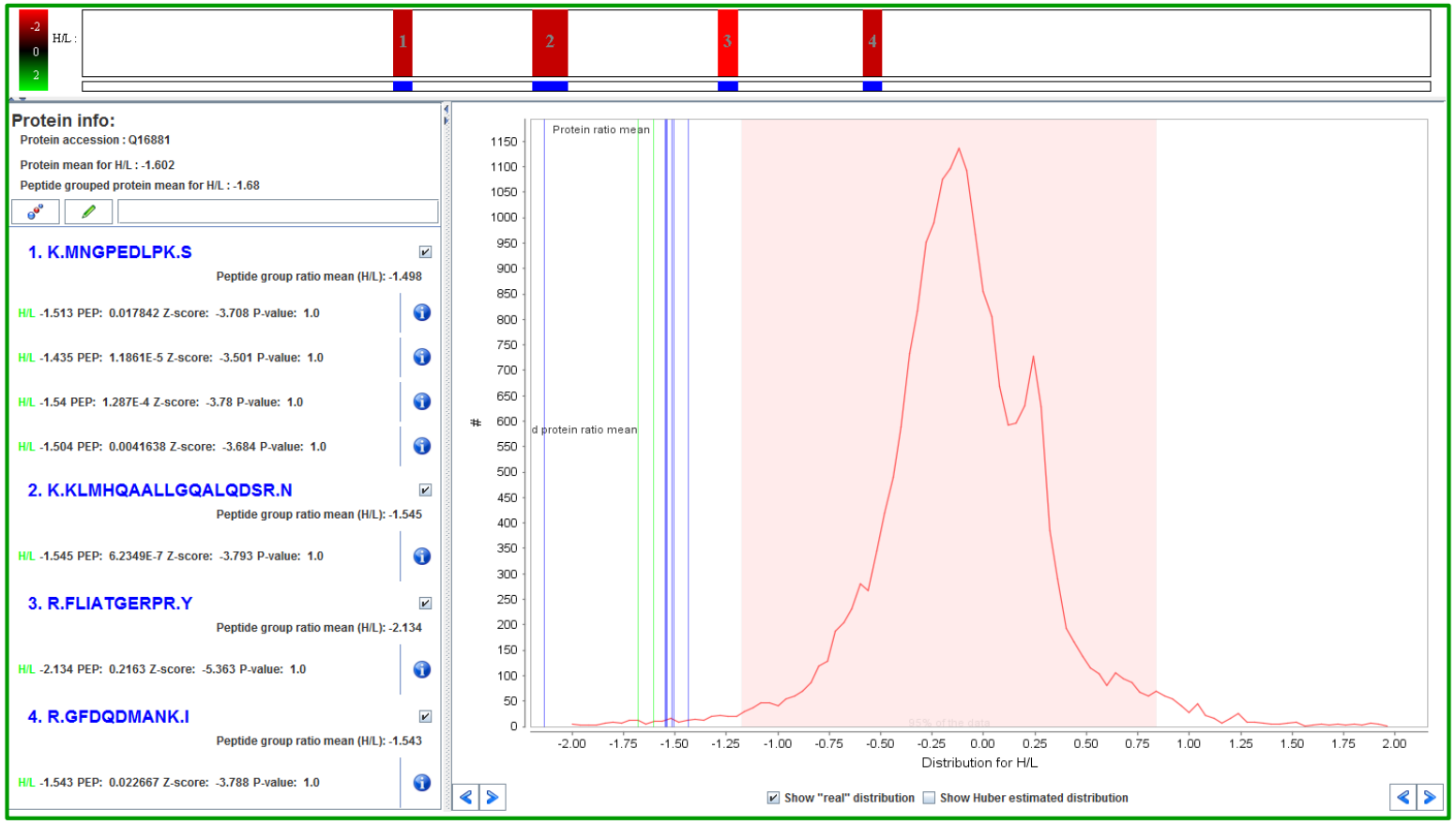
Protein inference is linked to quantification (i)



Nice and easy, 1/1, only unique peptides (blue) and narrow distribution

Colaert, Proteomics, 2010

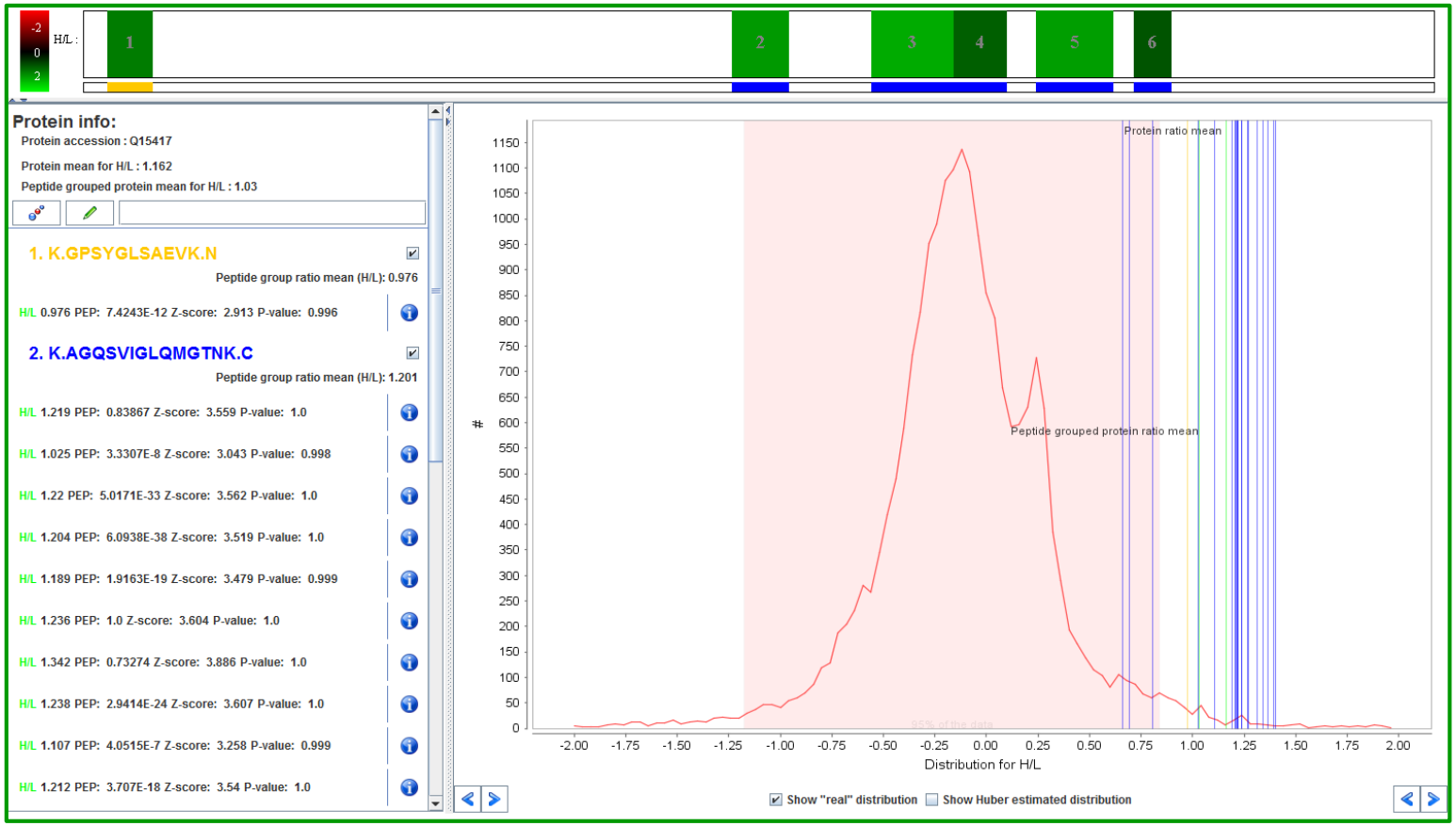
Protein inference is linked to quantification (ii)



Nice and easy, down-regulated

Colaert, Proteomics, 2010

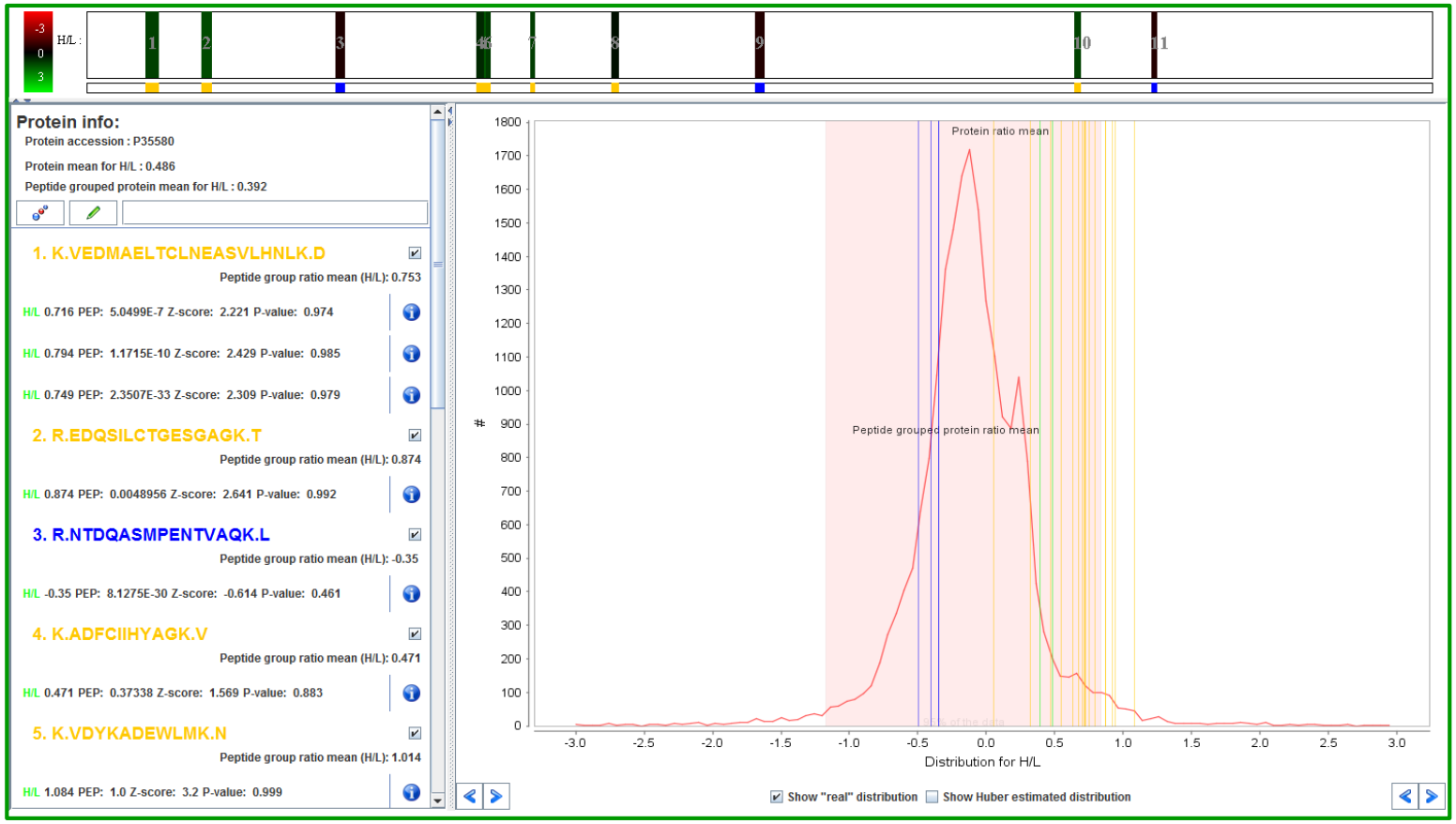
Protein inference is linked to quantification (iii)



A little less easy, up-regulated

Colaert, Proteomics, 2010

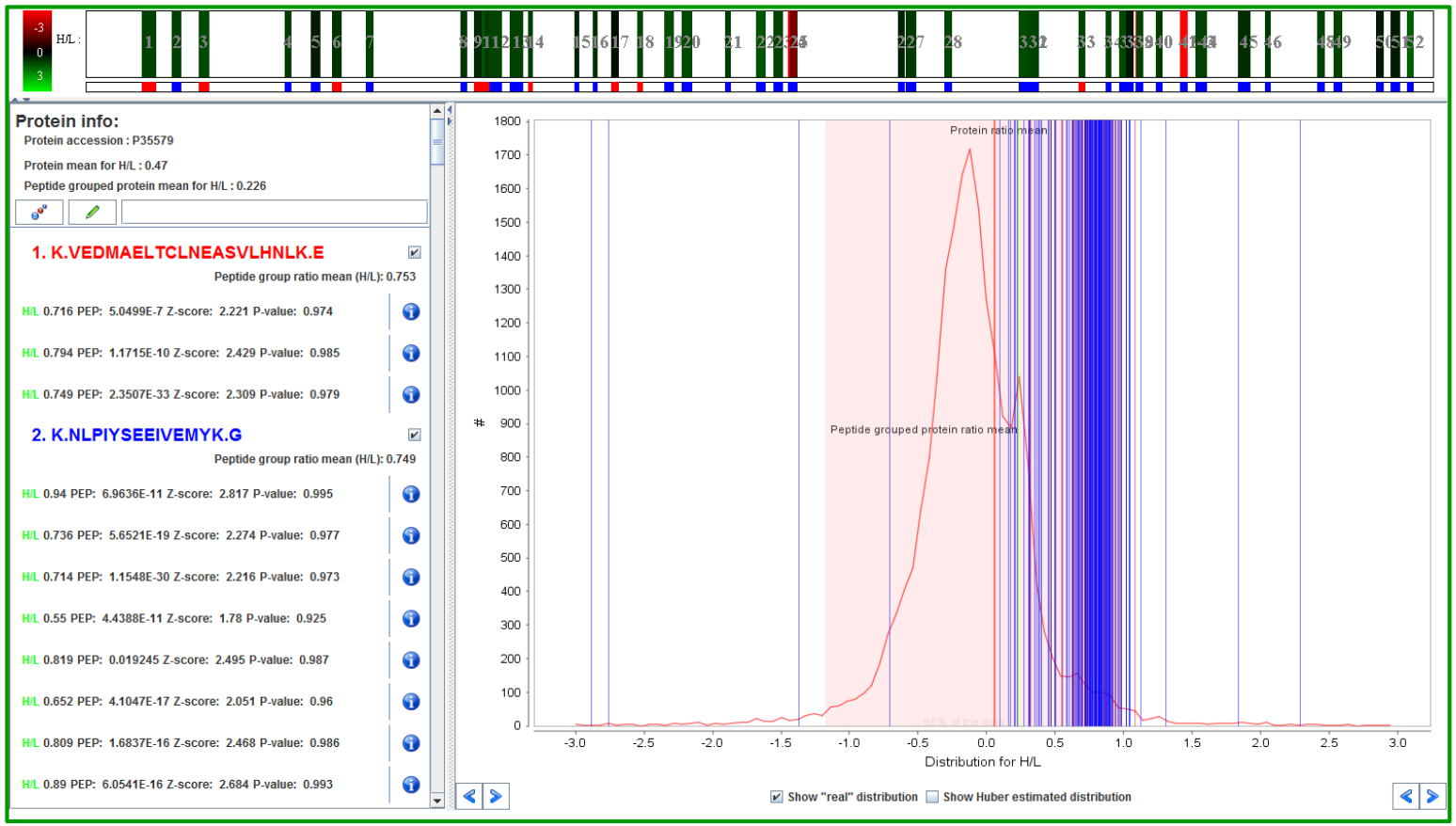
Protein inference is linked to quantification (iv)



A nice example of the mess of degenerate peptides

Colaert, Proteomics, 2010

Protein inference is linked to quantification (v)



A bit of chaos, but a defined core distribution

Colaert, Proteomics, 2010

Thank you!

Questions?