Bioinformatics C: Proteomic identification

DAVID L. TABB, PH.D.

Overview

- Tandem mass spectrometry measures the fragment ions produced by energizing peptide ions.
- Database search algorithms compare peptide sequences drawn from FASTA databases to tandem mass spectra.
- Target / decoy or distribution fitting models allow us to limit the rate of false IDs among peptide-spectrum matches.
- Protein assembly seeks a minimal set of proteins to explain the peptide evidence we trust.

What is proteomics?

Proteomics focuses on the identification, localization, and functional analysis of the protein make-up of the cell. The proteins present in a cell, together with their function, sub-cellular location, and perhaps even structure, change dramatically with the organism, and the conditions faced by their host cells including: age, checkpoint in the cell cycle, and external or internal signaling events.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Switching from parallel measurement to serial

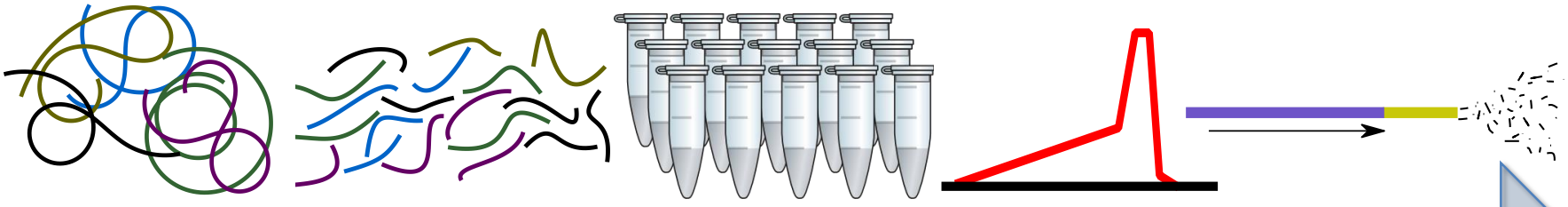
- High-throughput sequencing operates on millions of templates in parallel.
- Microarrays are rapid because the probes can all hybridize to cDNA independently.
- Mass spectrometry is a ***serial*** process; each spectrum requires the full attention of the mass analyzer. *Scan rate drives sensitivity*!

UNIVERSITEIT STELLENBOSCH UNIVERSITY

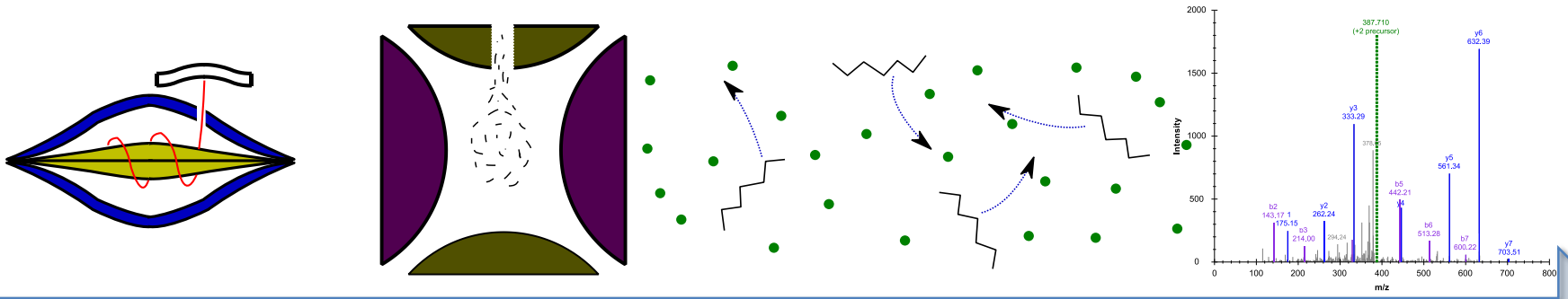
Three features of all mass spectrometers

- Ion Source:
  Produce ions from biological materials.
- Mass Analyzer:
  Separate or select ions by mass-to-charge (m/z) ratio.
- Detector:
  Report intensity of ions in mass spectrum.

Discovery Proteomics



Protein Mixture → Peptide Mixture → Peptide Fractionation → Liquid Chromatography → Electrospray Ionization



High-Resolution Mass Spectrometry → Isolate Ions of Peptide → Collide Ions to Dissociate → Collect Fragments in Tandem MS

Which proteins are present? How have they been modified?
Which proteins differ most between cohorts?

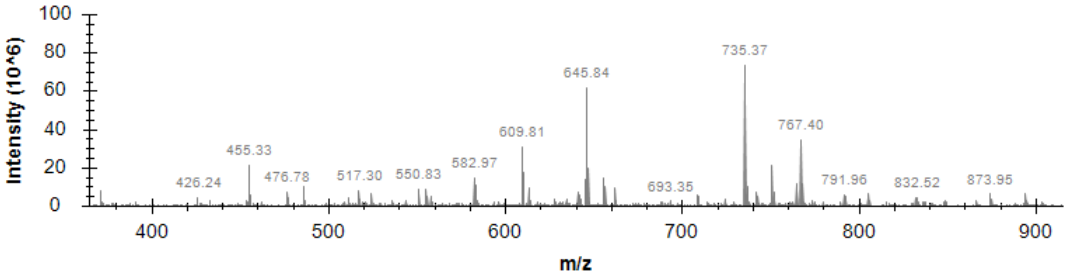
M. Bantscheff et al. *Anal. Bioanal. Chem.* (2007) 389: 1017

Disruption steps in proteomics

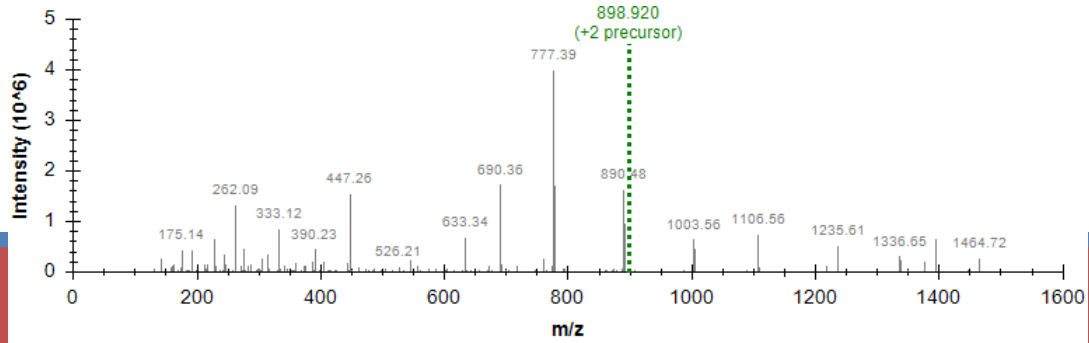
- Denaturation and reduction: disruption of protein structure, leaving backbone intact
- Enzymatic cleavage: cutting a protein into peptides, typically by the enzyme trypsin
- CID: fragmenting a peptide ion through Collision-Induced Dissociation.

Shotgun proteomics: Two types of scans

- Mass spectrum (MS): Peaks represent intact peptide ions that are eluting from LC-ESI.

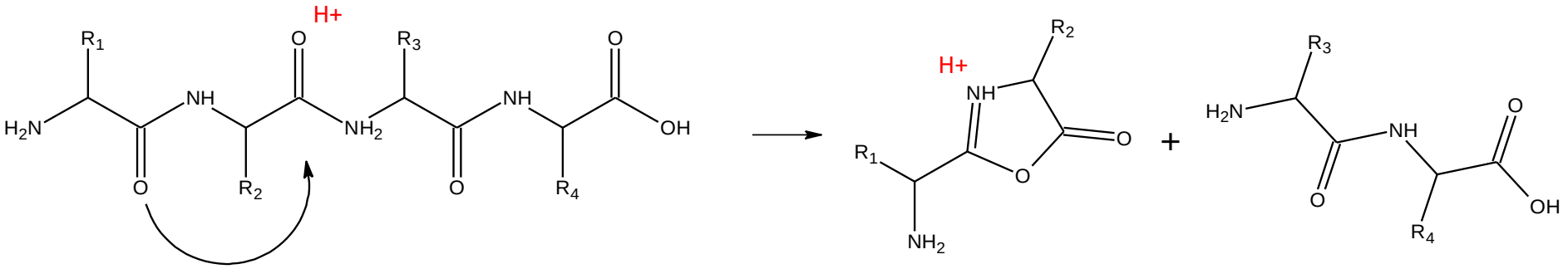


- *Tandem mass spectrum* (MS/MS): Peaks represent the fragment ions that result from CID of many copies of a peptide ion.



UNIVERSITEIT STELLENBOSCH UNIVERSITY

Tandem mass spectrometry



1. Collision-Induced Dissociation (CID) adds energy to peptides through gas collisions.
2. Vibrations mobilize protons, drawing carbonyl electrons and leaving carbon partially positive.
3. Electrons on preceding carbonyl attack!

DF Hunt, *PNAS* (1986) 83: 6233-6237.

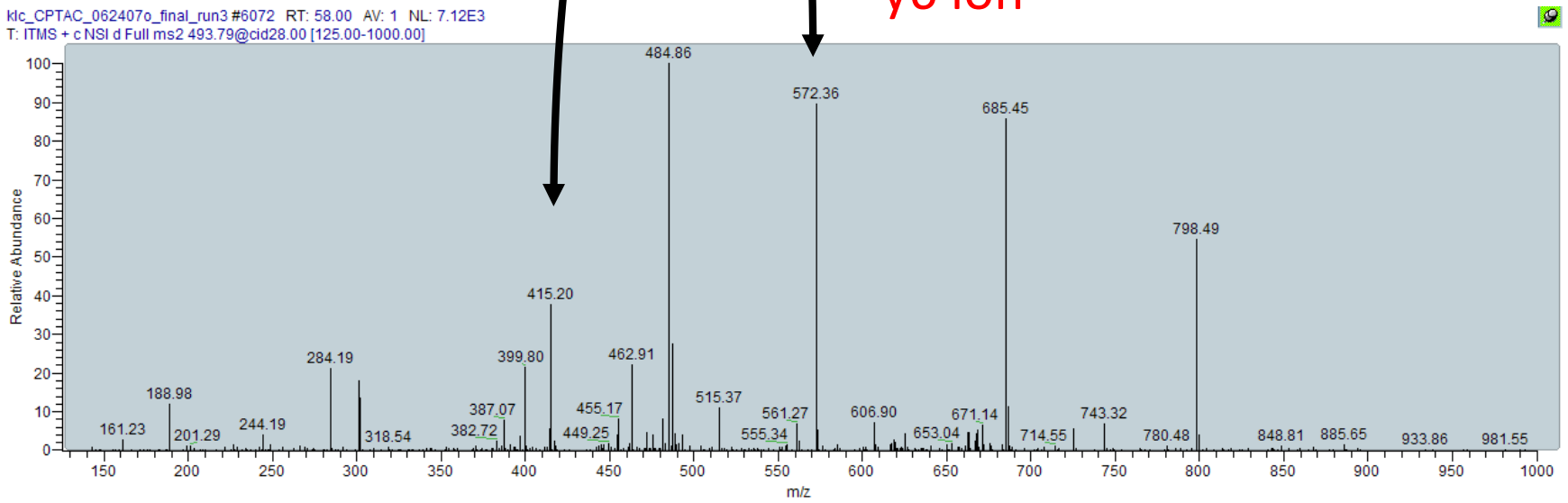
UNIVERSITEIT STELLENBOSCH UNIVERSITY

Fragment ions result from breakage of peptide bonds

TSII|GTIGPK

N-terminal b4 ion

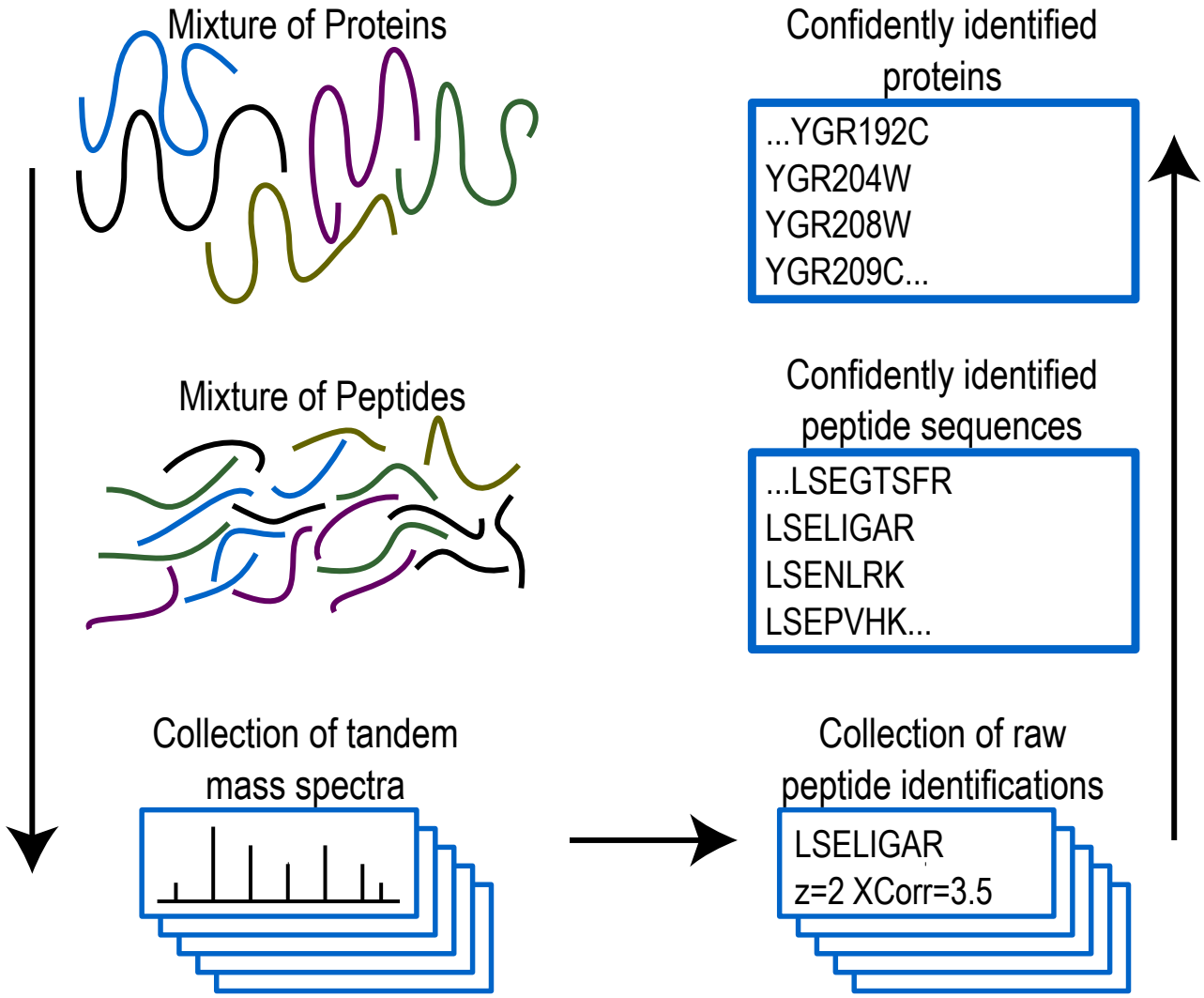
C-terminal y6 ion



Tandem mass spectrum (MS/MS) from an ion trap; peaks are fragments.

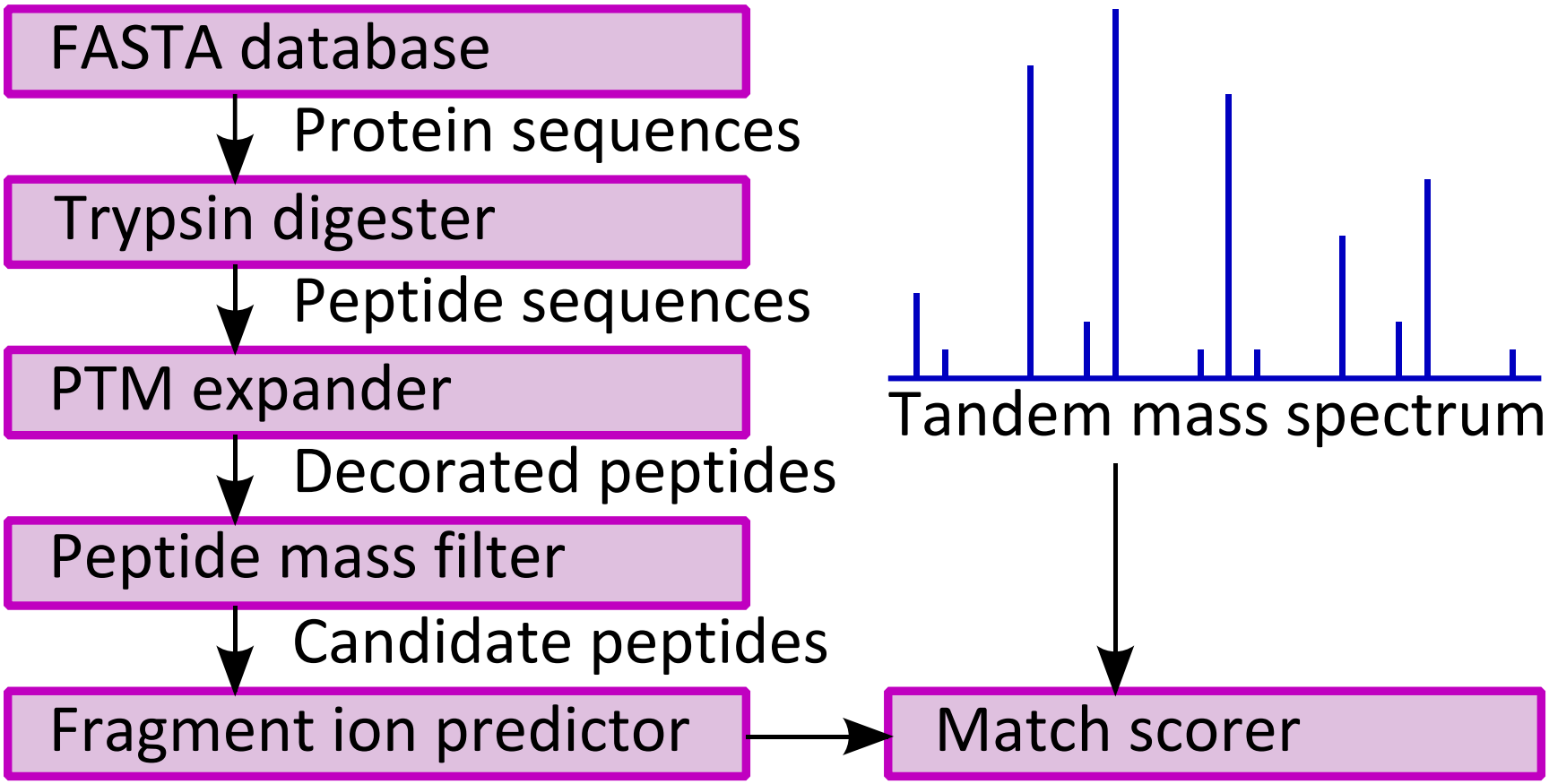
Database search algorithms

Disassembly and reassembly



After AI Nesvizhskii, *Mol Cell Proteomics* (2005) 4: 1419-40.

Database search overview

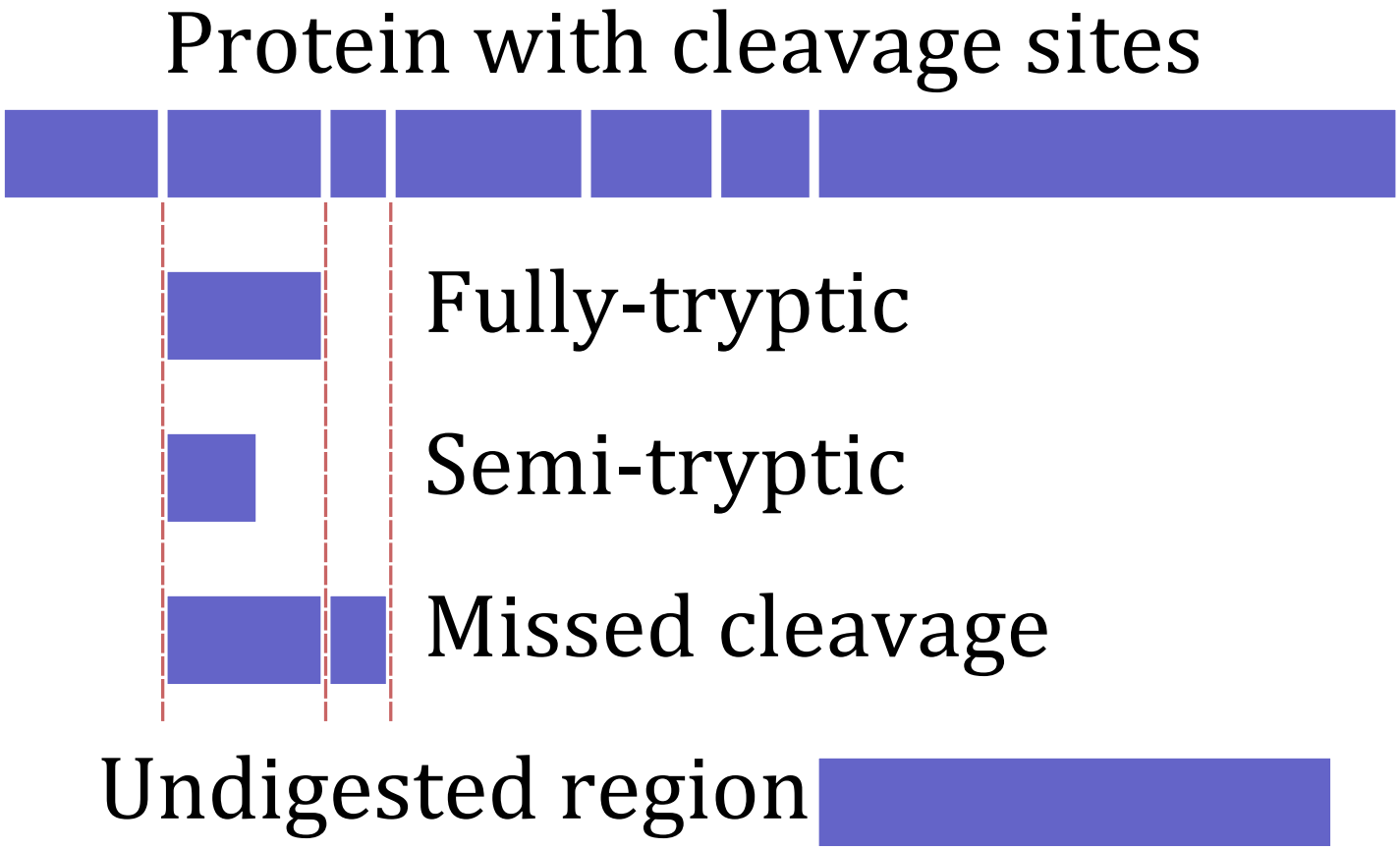


Eng et al (1994) *J. Amer. Soc. Mass Spectrom*. 5: 976-989.

Yates et al (1995) *Anal. Chem*. 67: 1426-1436.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Emulating proteases *in silico*



N Edwards and R Lippert. *Lecture Notes In Computer Science* (2002) 2452: 68-81.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Dynamic PTMs grow search space

Because multiple PTMs may be in each peptide, adding PTMs to a search creates an *exponential* cost.

Here, three sites lead to eight PTM variants.

D I G S E S T E K

D I G S*E S T E K

D I G S E S*T E K

D I G S E S T*E K

D I G S*E S*T E K

D I G S*E S T*E K

D I G S E S*T*E K

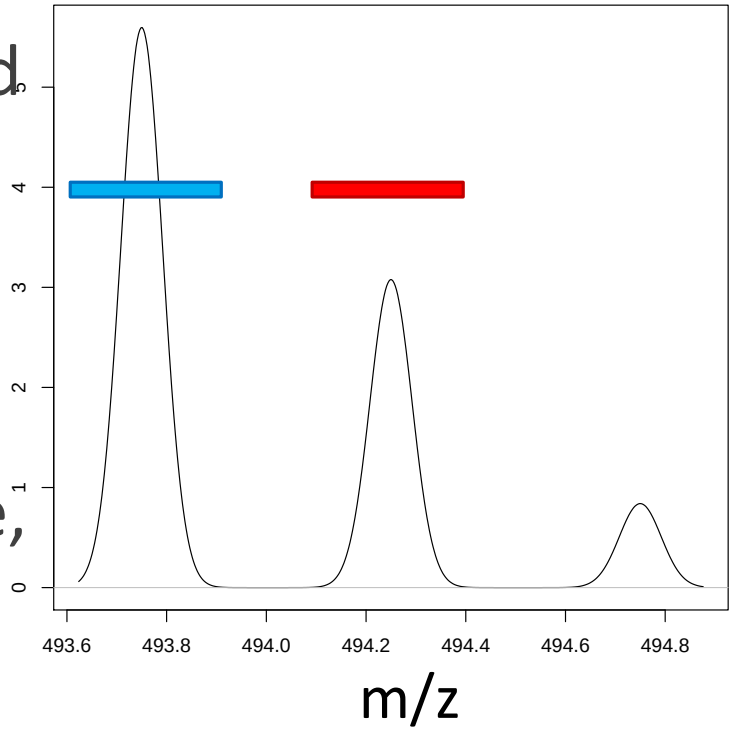
D I G S*E S*T*E K

CASA1_BOVIN

Yates et al *Anal. Chem*. (1995) 67: 1426-1436.

Peptide mass filter

- Peptides can only be compared with MS/MS if their computed masses agree with measured mass within a “precursor mass tolerance” (blue bar).
- If “isotope correction” is in use, comparisons may be made as if red peak were misreported as monoisotope.



Mass analyser (for MS1)	Typical setting
Quadrupole ion trap	1.25 m/z (comingled isotopes)
Time Of Flight	100 ppm (without correction)
FT / Orbitrap	10 ppm

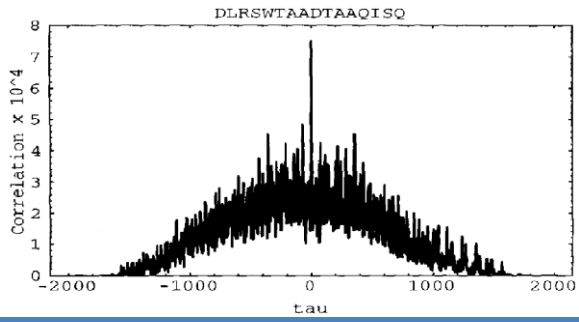
Zubarev and Mann. *Mol. Cell. Proteomics* (2007) 6: 377-381.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Cross correlation: Sequest scores

- Normalize observed spectrum.
- Generate model spectrum for each candidate.
- Convert observed and model spectrum to frequency domain by FFT.
- Cross-correlate, reporting ratio between zero-offset alignment and nearby alignments.

$$\text{xcorr} = x_0 \cdot y' \quad \text{where} \quad y' = y_0 - \left(\sum_{\tau=-75, \tau \neq 0}^{\tau=+75} y_\tau \right) \Big/ 150$$



J Eng et al. *J. Proteome Res.* (2008) 7: 4598-4602.

J Eng et al. *J Amer. Soc. Mass. Spectrom.* (1994) 5: 976-989.

Andromeda (MaxQuant) scoring approach

- The *binomial distribution* computes the probability of matching this many or more peaks in MS/MS by chance alone.
- Code asks if the score improves with *neutral losses* from peptide modification.
- Code maximizes score for different MS/MS densities.

n = total number of theoretical ions
k = number of matching ions in spectrum

Approx. probability of getting at least *k* matches by chance

$$s(q,\text{loss}) = -10\log_{10}\sum_{j=k}^{n}\left[\binom{n}{j}\left(\frac{q}{100}\right)^{j}\left(1-\frac{q}{100}\right)^{n-j}\right]$$

Optimize inclusion of losses

$$s(q) = \max_{\text{loss = true/false}} s(q,\text{loss})$$

Optimize *q* (peaks per 100 Da)

$$s = \max_{p\geq 2} s(q)$$

Cox et al. *J. Proteome Res*. (2011) 10:1794-1805.

Filtering PSMs and assembling proteins

DB Search scores play two roles:

BOOSTING SEQUENCE RANKS

- Only top matches for each MS/MS get considered further.
- Sensitivity requires true sequence to score highest on even marginal spectra.

FILTERING PEPTIDE-SPECTRUM MATCHES (PSMS)

- Scores also serve to separate accepted from rejected PSMs.
- Well-calibrated scores can be compared among different spectra.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Controlling PSM* error

* *PSM=Peptide-spectrum match*

- Distribution modeling
  - How do scores distribute for false PSMs?
  - How do scores distribute for true PSMs?
  - *What is the probability this PSM is true?*
- Target-Decoy (a.k.a. reversed search)
  - Database includes protein sequences in both target (natural) and decoy (bogus) forms.
  - Matches to known false sequences model matches to unknown false hits.

Tabb. *J. Proteome Res*. (2007) 7:45-46.

Discriminant Function Analysis combines sub-scores from Sequest

Table 1. Discriminant Functions Derived from Training Dataset Spectra of $[M + 2H]^{2+}$ and $[M + 3H]^{3+}$ Precursor Ions[a]

	$[M + 2H]^{2+}$		$[M + 3H]^{3+}$	
variable	coefficient	correlation	coefficient	correlation
Xcorr′	8.362	0.798	9.933	0.698
ΔC_n	7.386	0.746	11.149	0.806
ln SpRank	−0.194	−0.510	−0.201	−0.491
d_M	−0.314	−0.306	−0.277	−0.251
constant	−0.959		−1.460	

[a] The coefficients weighting each variable are indicated, as well as the correlation between each variable and the discriminant function given by the loading matrix, indicative of the contribution of each variable to discrimination. Contributions can range from none (correlation of 0) to complete (correlation of ±1).

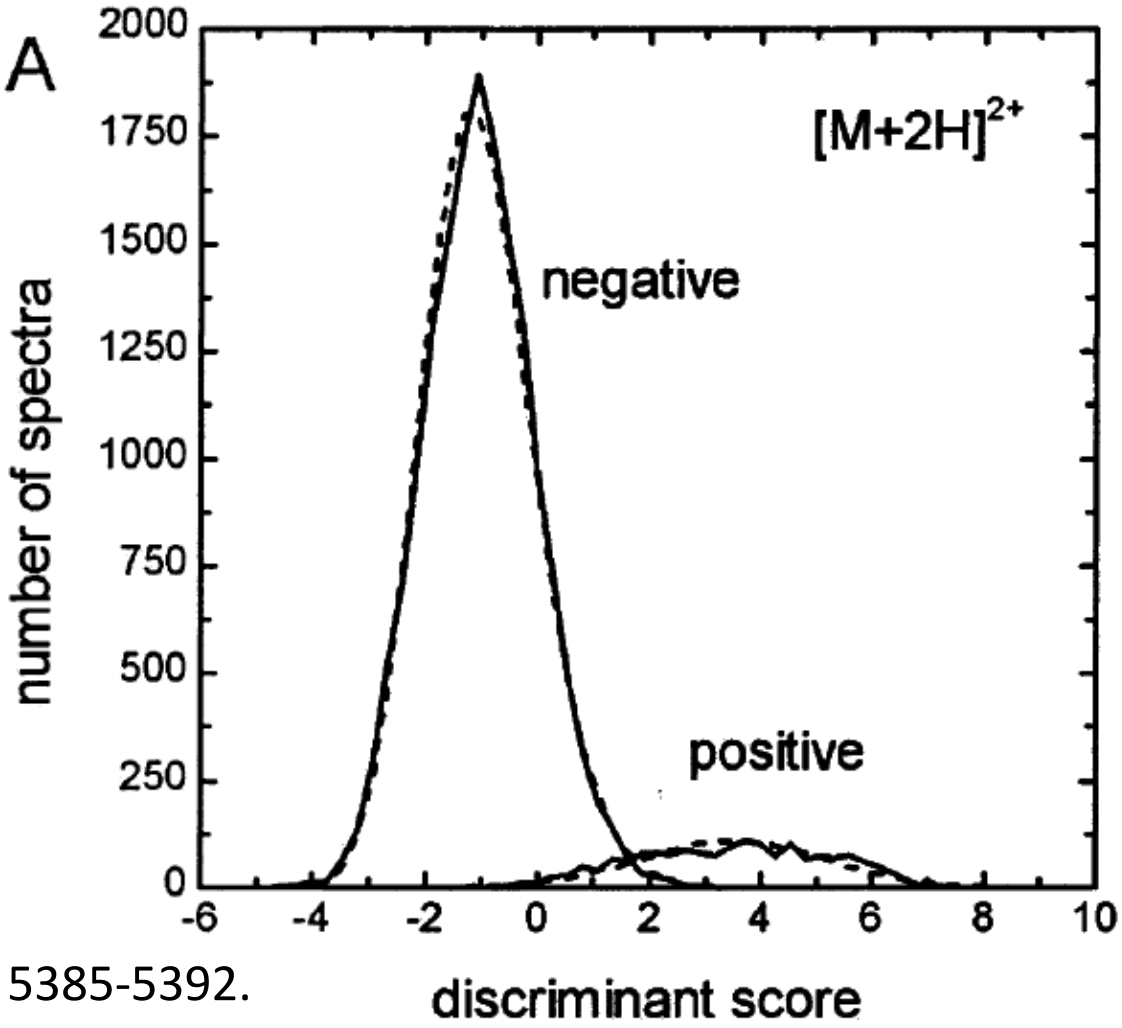
Absolute Score →

Relative Score →

A Keller. *Anal. Chem.* (2002) 74: 5383-5392.

PeptideProphet: Mixture Model separates true and false distributions

- *Discriminant score* combines multiple features.
- Expectation maximization adjusts parameters for two distributions to match empirical score distribution.
- Peptide probability reflects both heights for that score.



Keller et al. *Anal. Chem*. (2002) 74: 5385-5392.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Simpler FDR error control: Target/decoy analysis estimates FDR

- Sequence database has equal numbers of *target* and *decoy* sequences.
- False IDs split evenly among target or decoy sequences.
- Apply a threshold, and:
  - False estimate = 2 x [decoy hit count].
  - False Discovery Rate (FDR) = False estimate divided by number of passing IDs.

$$FDR = \frac{2 * NumDecoy}{(NumDecoy + NumTarget)}$$

Elias and Gygi. *Nature Methods* (2007) 4: 207-214

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Peptides are identified; proteins are *inferred*

- "Shotgun" proteomics produces MS/MS that represent individual peptides.
- We infer a set of protein sequences that best explains the peptides we detected.
- The MS/MS has no memory of the protein in which the peptide originated.
- "Top-Down" proteomics identifies intact proteins *without the digestion step*.

L.M. Smith et al. *Nature Methods* (2013) 10: 186-187.

UNIVERSITEIT STELLENBOSCH UNIVERSITY

Each peptide may appear in multiple database proteins.

- In multi-species databases, a peptide may be shared across species due to *orthology*.
- Gene duplication makes *paralogs* share peptides within a single species.
- Many sequence databases report proteins derived from multiple transcripts (*isoforms*) separately, with high peptide overlap.

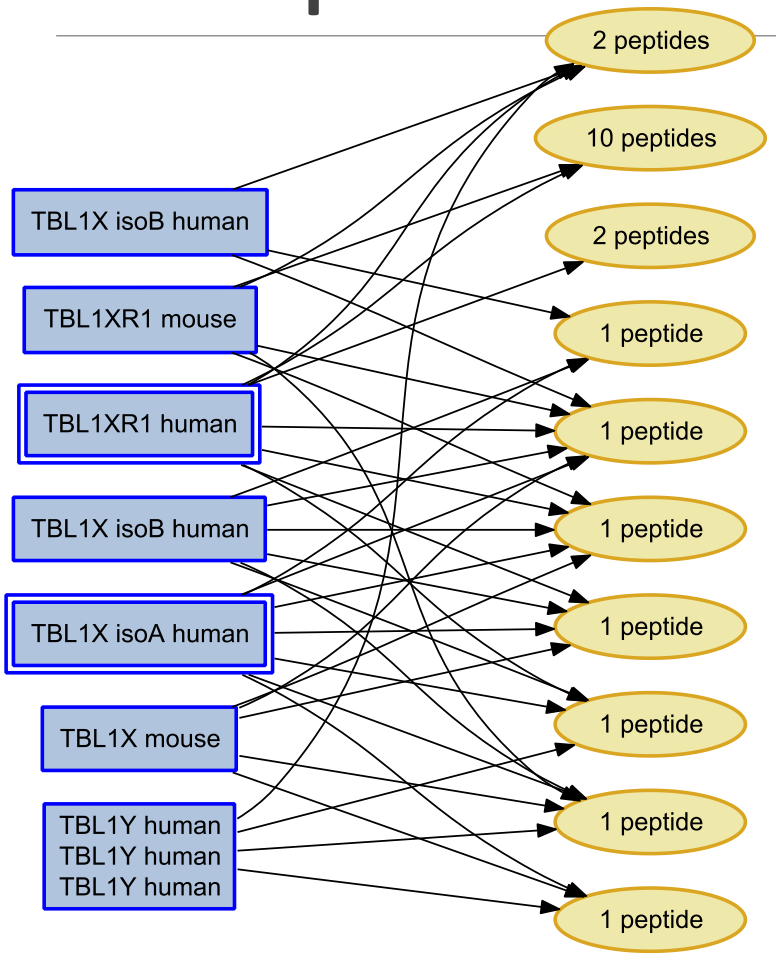
Koonin. *Genome Biology* (2001) 2: comment 1005.1-1005.2.

Parsimony

- *noun*: “economy of explanation in conformity with Occam's razor”
  Merriam Webster OnLine

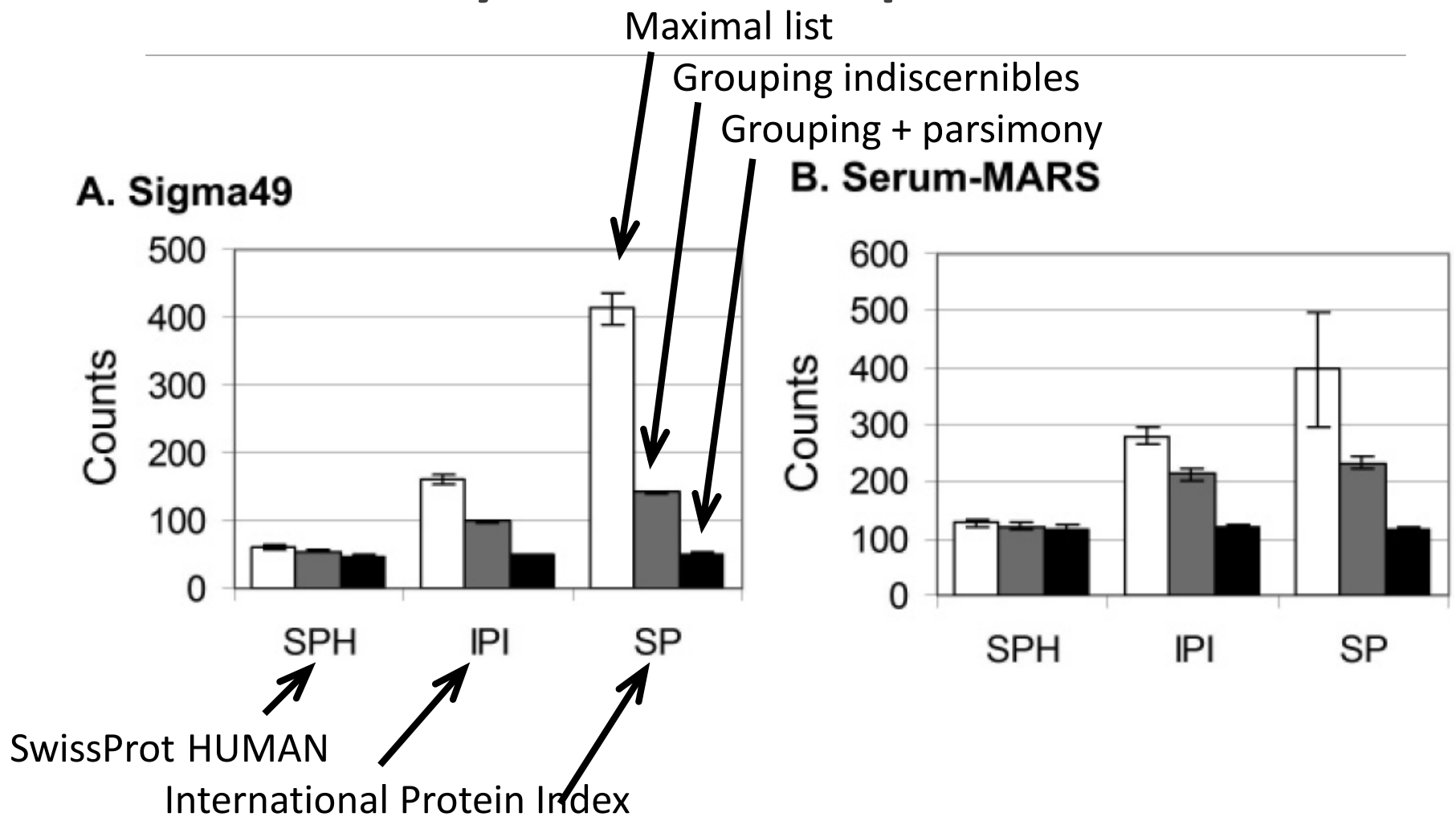
- *“Plurality ought never be posed without necessity.”*
  William of Occam

Two proteins or seven?



- Sample mixes mouse and human proteins.
- Isoforms, paralogs, and orthologs complicate protein-peptide map.
- Protein evidence may be indiscernible or subsumable

Parsimony reduces protein lists



B Zhang. *J. Proteome Res.* (2007) 6: 3549-3557.

Takeaway messages

- Proteomic identification depends upon database of known protein sequences to define the set of possible peptides.
- Controlling error in identified peptides is critical to reliable proteomics interpretation.
- Detecting indiscernible proteins and applying parsimony is a conservative guide.