H3ABioNet

H3ABioNet

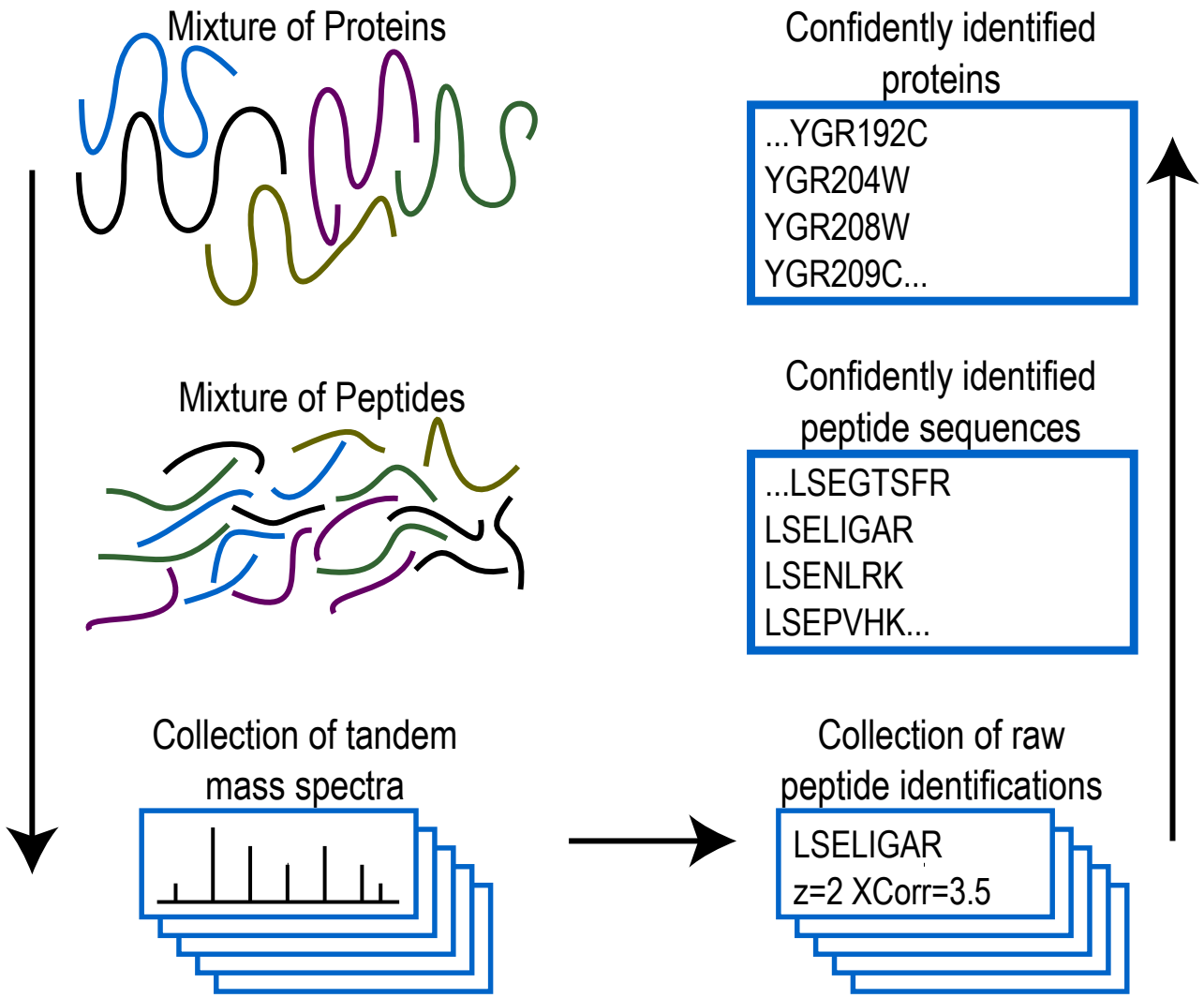
Pan African Bioinformatics Network for H3Africa

Introduction to Bioinformatics Online Course: IBT

Gene Expression

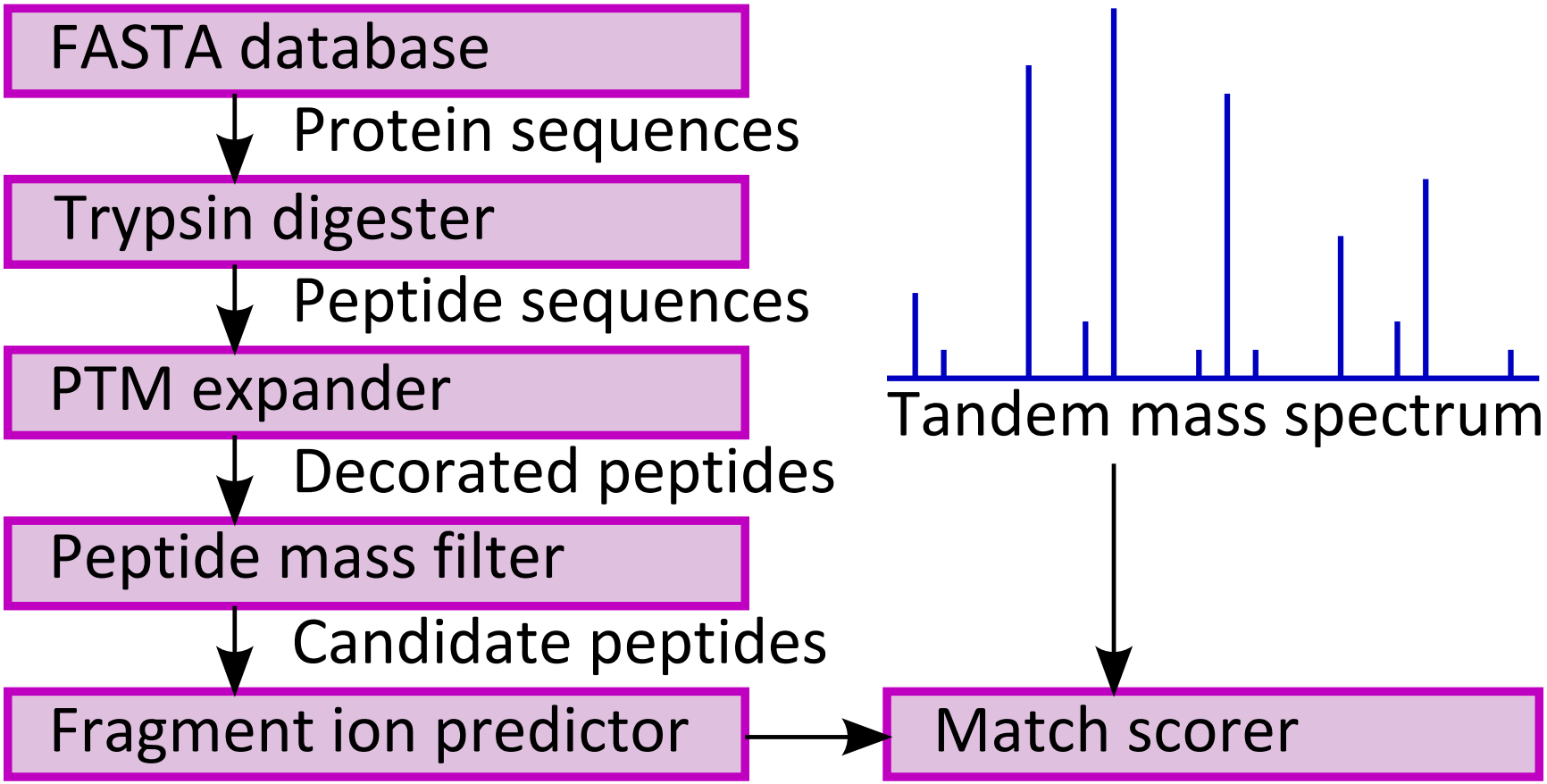
Proteome Informatics

Disassembly and reassembly



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

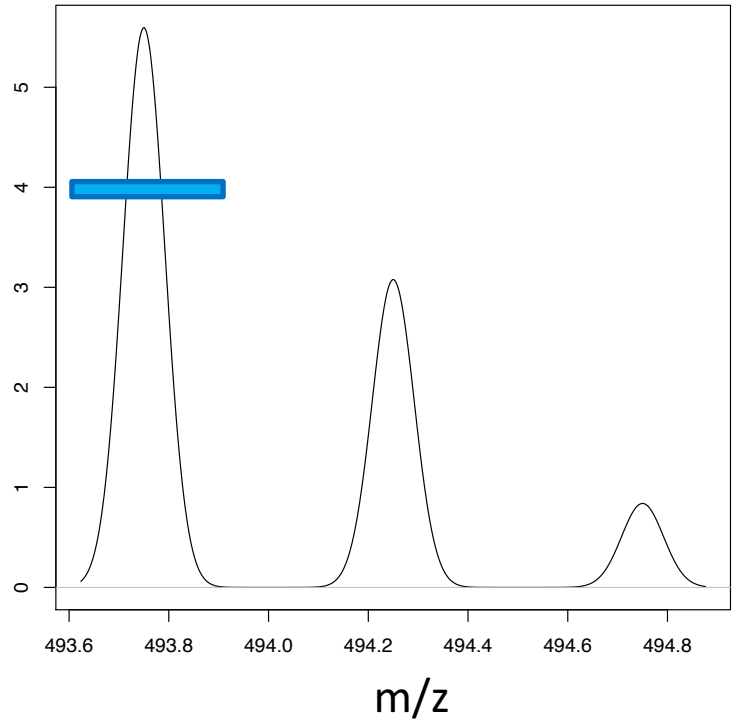
Database search overview



Eng et al (1994) *J. Amer. Soc. Mass Spectrom*. 5: 976-989.
Yates et al (1995) *Anal. Chem*. 67: 1426-1436.

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).
- Different mass analyzers measure mass to different accuracies.

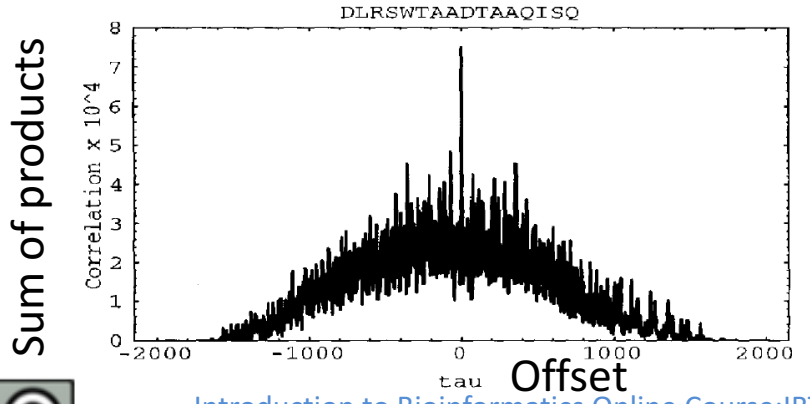


Mass analyser (for MS1)	Typical setting
Quadrupole ion trap	1.25 m/z (comingled isotopes)
Time Of Flight	100 ppm (without correction)
FT / Orbitrap	10 ppm

Sequest cross correlation

- 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.

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



Random match probabilities

- Imagine spectrum as jar of 100 black and 900 white marbles (peaks and voids).
- Sample 20 marbles for a predicted peaklist, drawing 15 black and 5 white.
- Probability that this happened by random chance from hypergeometric distribution:

$$p = \frac{\binom{100}{15}\binom{900}{5}}{\binom{1000}{20}} = 3.63146\text{E-}12$$

T Fridman
J. Bioinfo. Computat. Bio.
(2005) 3: 455-476.

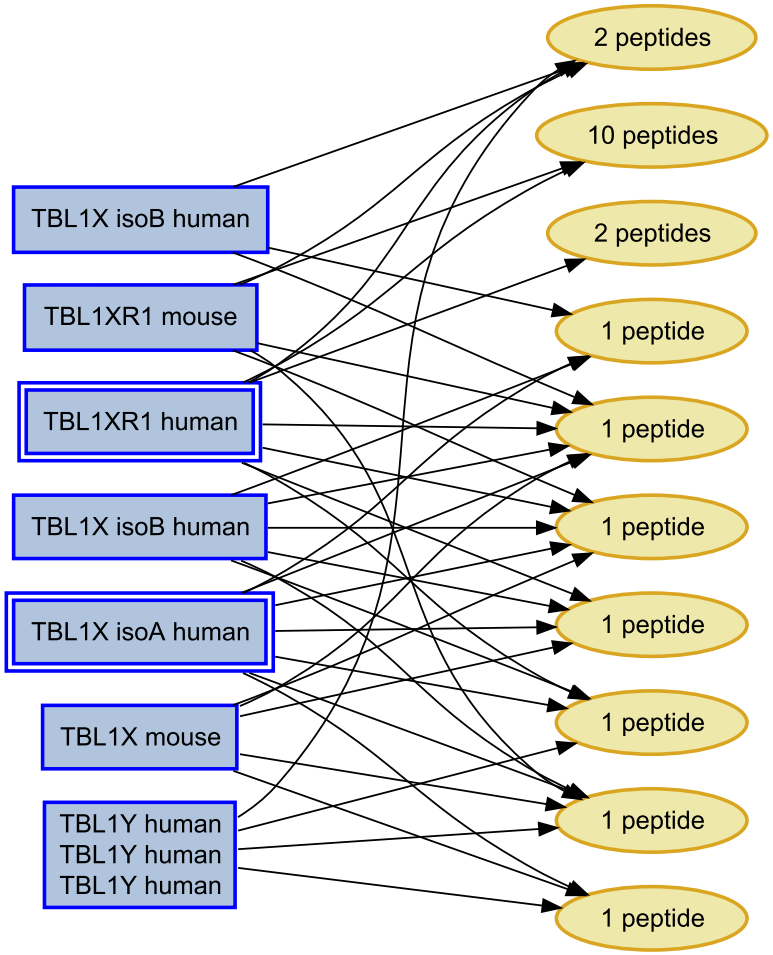
Controlling PSM* error

- 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 falses.

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

* *PSM=Peptide-spectrum match*
A particular MS/MS to which one peptide sequence has been matched with a particular score

Claim no more proteins than you must to explain observed peptides.



Peptides may be found in multiple proteins due to:

- Orthologous genes from different species
- Paralogous gene copies within a species
- Different isoforms from the same gene

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

Takeaway messages

- An experiment may produce a million spectra; matching them to DB peptides is intensive.
- Estimating the error rate of identification is essential to publishing successfully.
- Proteins are inferred on basis of observed peptides; these numbers can easily drift.
- Identifying 100,000 peptides from a sample is feasible with today's instruments.