Speed Optimization

- Algorithm 2:

  For each MS/MS spectrum

    For each protein in the database

      In-silico digest the protein into peptides

      For each peptide

        if (precursor mass error < allowed error tolerance)

          Evaluate the peptide-spectrum match

  Assign the highest-scoring peptide to the spectrum

error Tol = 0.1 Da

precursor 500 – 1500 Da

- Precursor mass error = | theoretical precursor mass – observed precursor mass |
- The mass filtration reduces the number of PSM evaluation.

Speed Optimization

- Algorithm 3:

  Sort the spectra according to precursor mass.

  For each protein in the database

    In-silico digest the protein into peptides

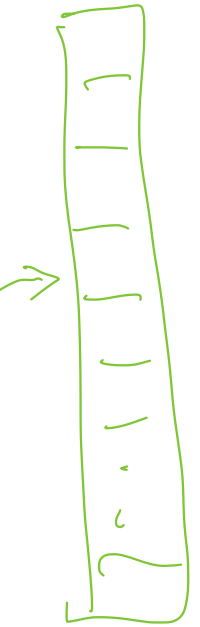
      For each peptide

        For each spectrum with matching precursor mass

          Evaluate the peptide-spectrum match

          Keep the highest scoring peptide for the spectrum

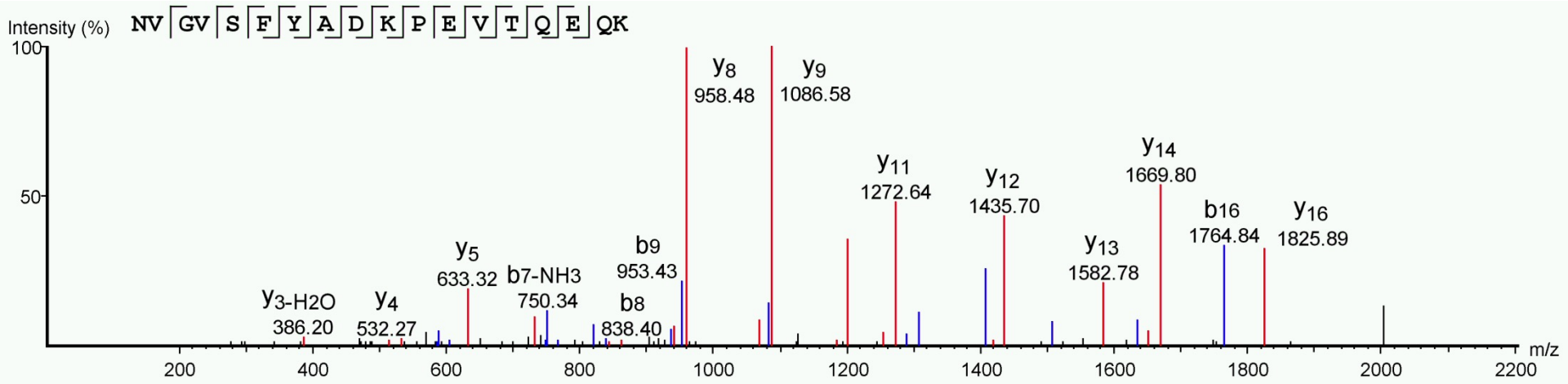
n spectra
m peptides.
n · log n
m peptides.
m · log n
Sorted spectra.
peptide

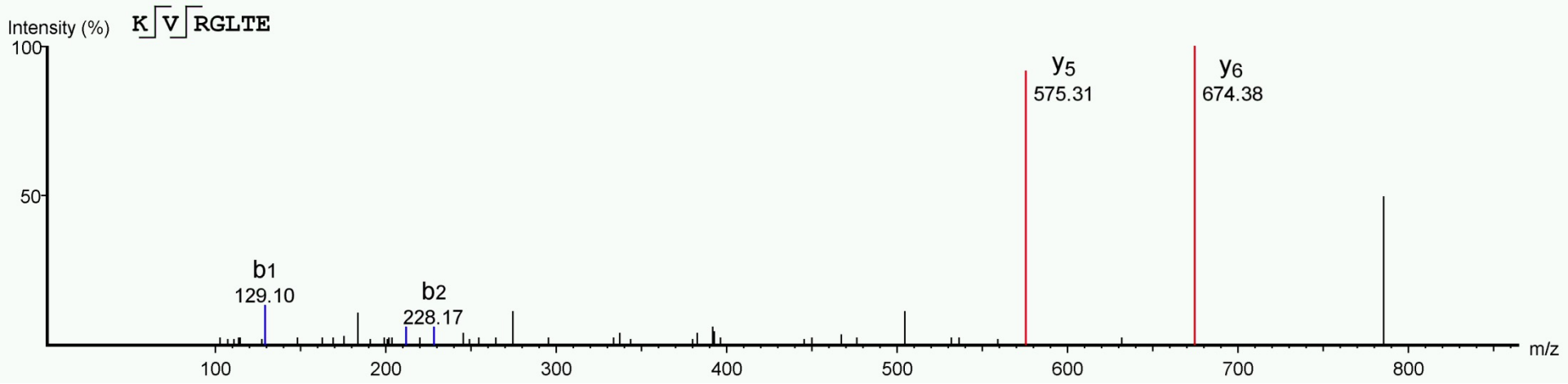


Result Validation

- Some spectra are of lower quality.
  - Peptides do not fragment well
  - Peptides fragment too much
  - Peptides do not get charged → Don't fly
  - Peptides are of low concentration
  - Etc.
- Some spectra's true peptides are not in database.
- Therefore, search results of some spectra are junk.
- Computer scientists may think these are not their problems. After all, they've reported the "optimal" peptide for each spectrum.

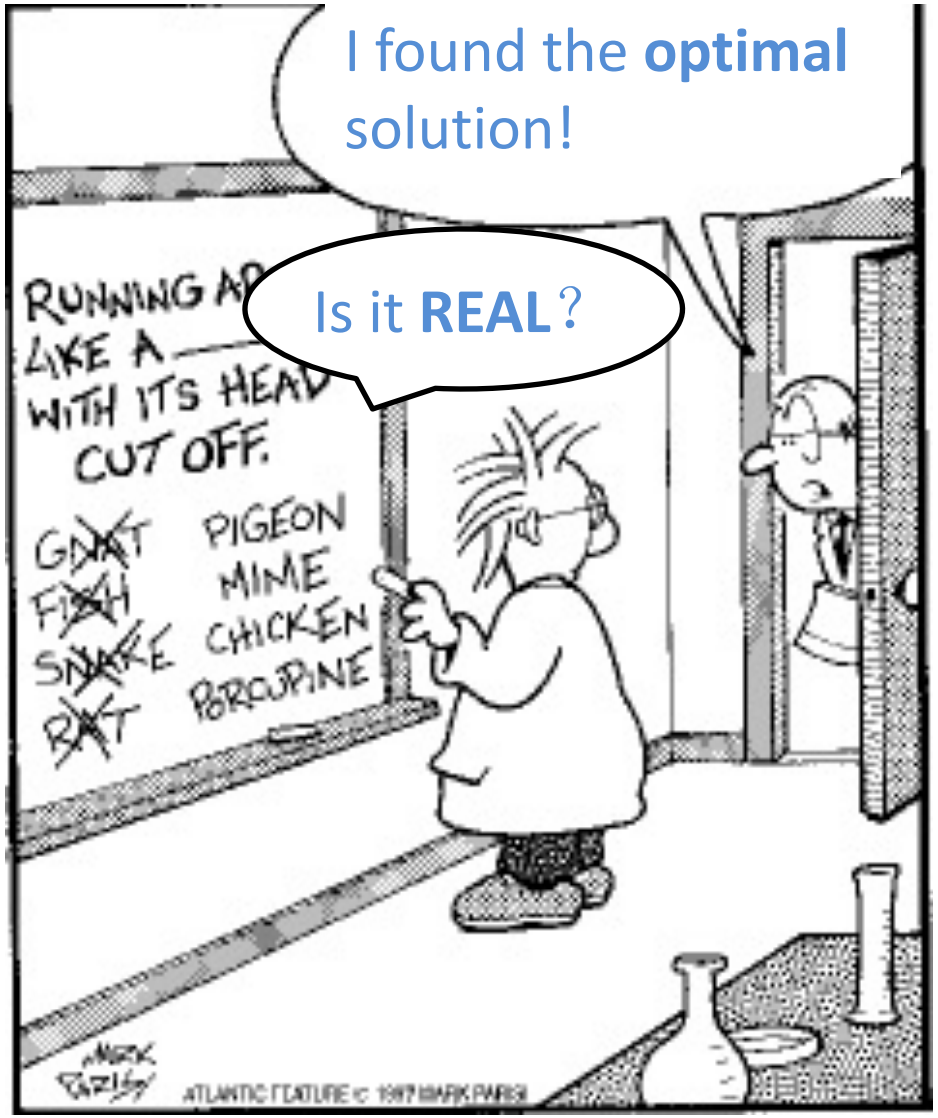
Peptide Spectrum Matches (PSM)



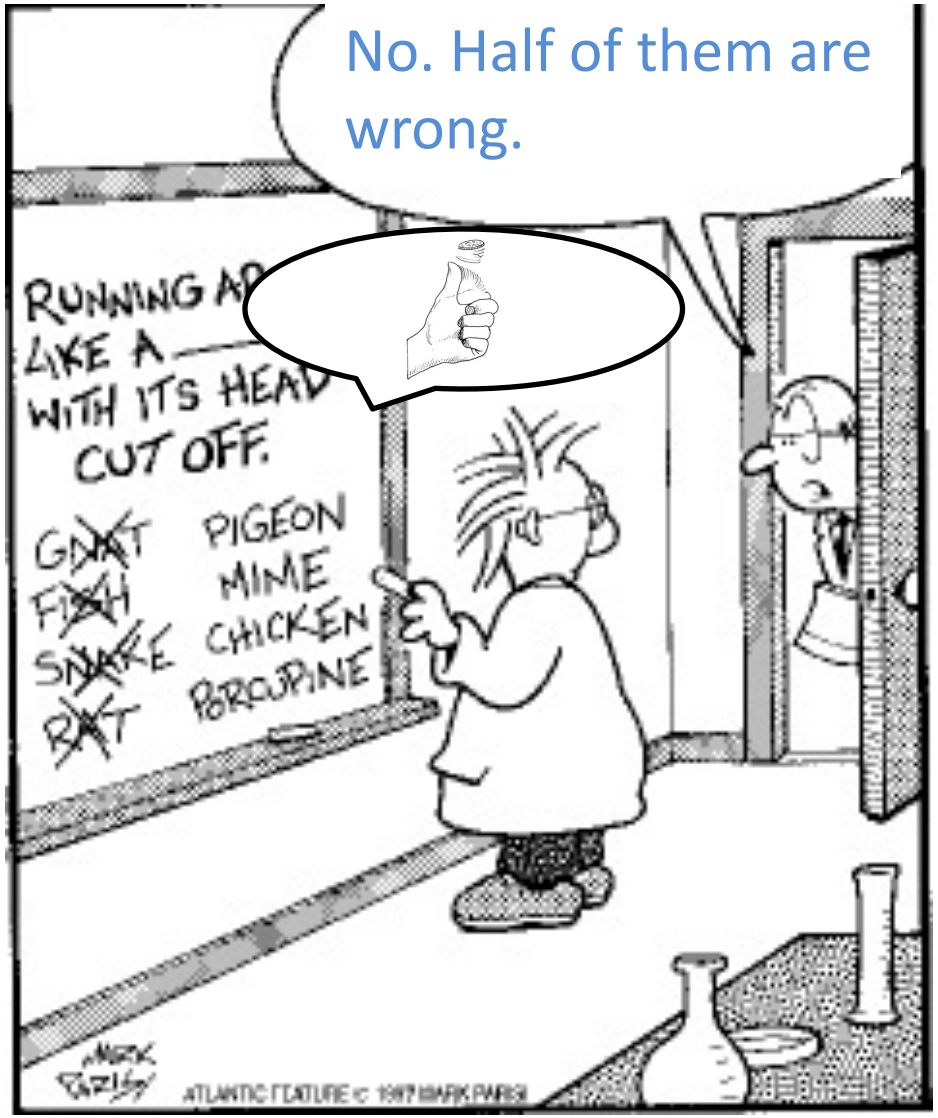


These two PSMs are all the best match from database for two different spectra.
Their confidences are clearly different.

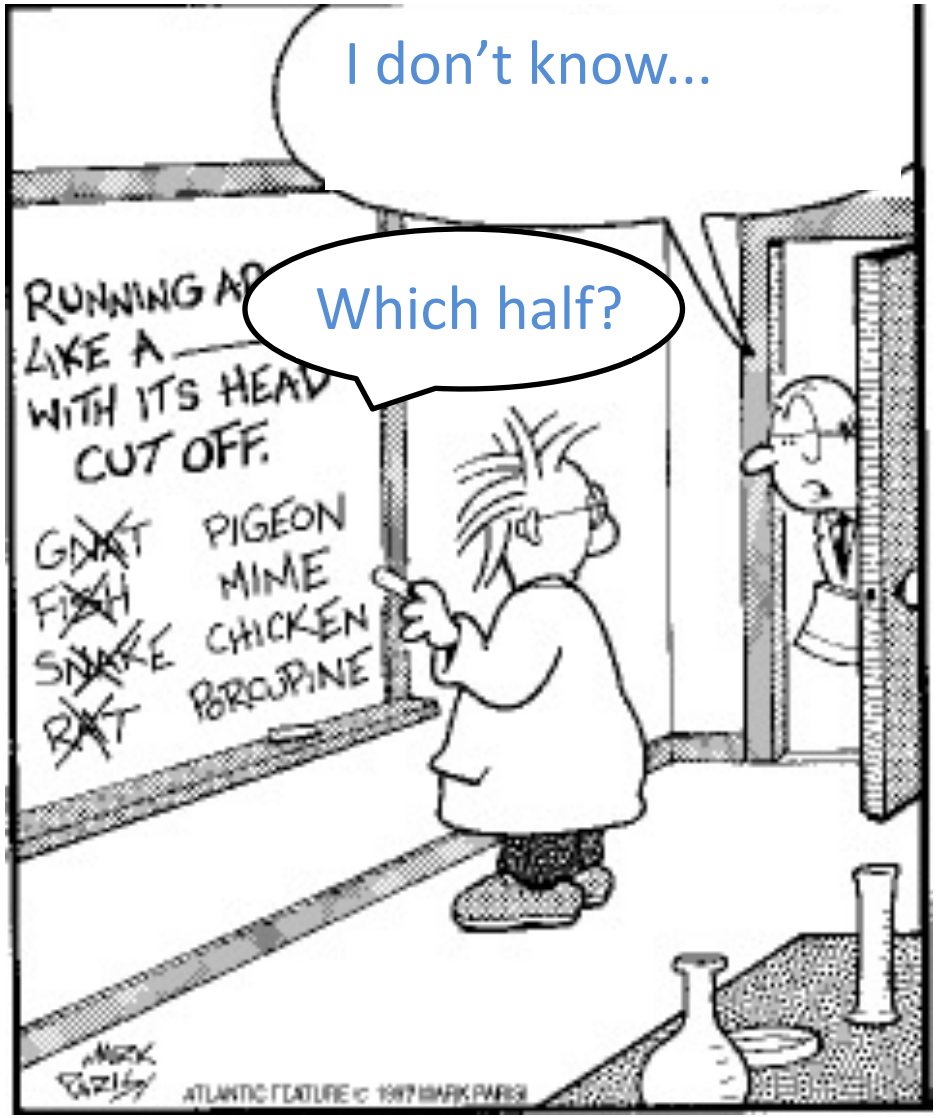
Biologists vs. Computer Scientists



Biologists vs. Computer Scientists



Biologists vs. Computer Scientists



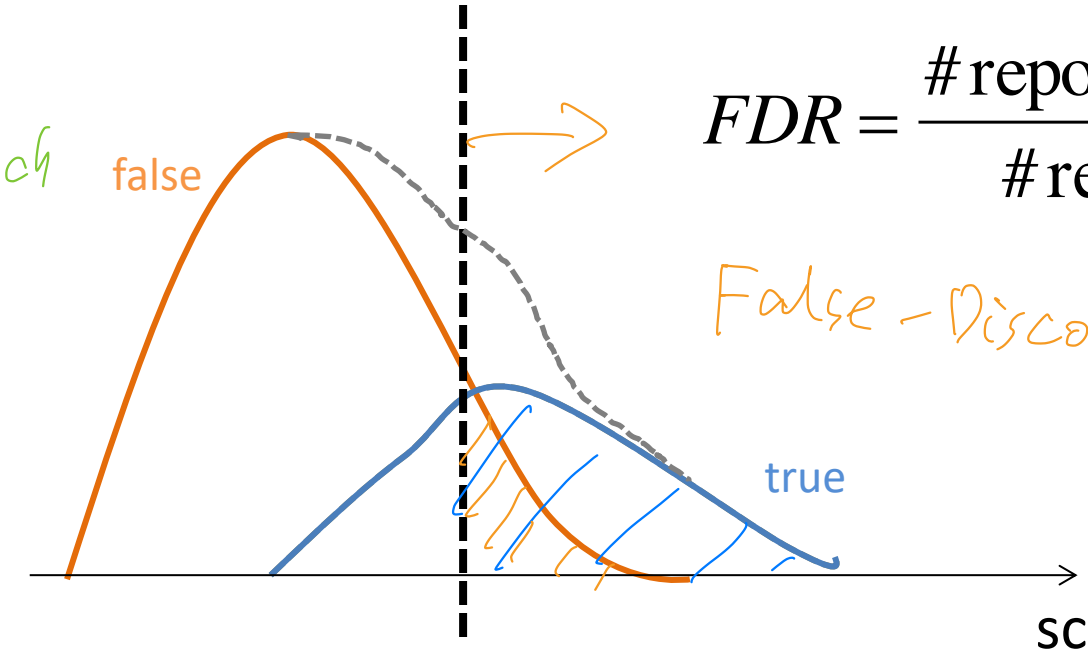
Solution to Noisy Input

- Only report results if you're confident.
- Discard the less confident ones.
- This increases accuracy to make the results useful at the price of discarding some data.

Only Report Highly Confident Results

PSM:

peptide-spectrum match

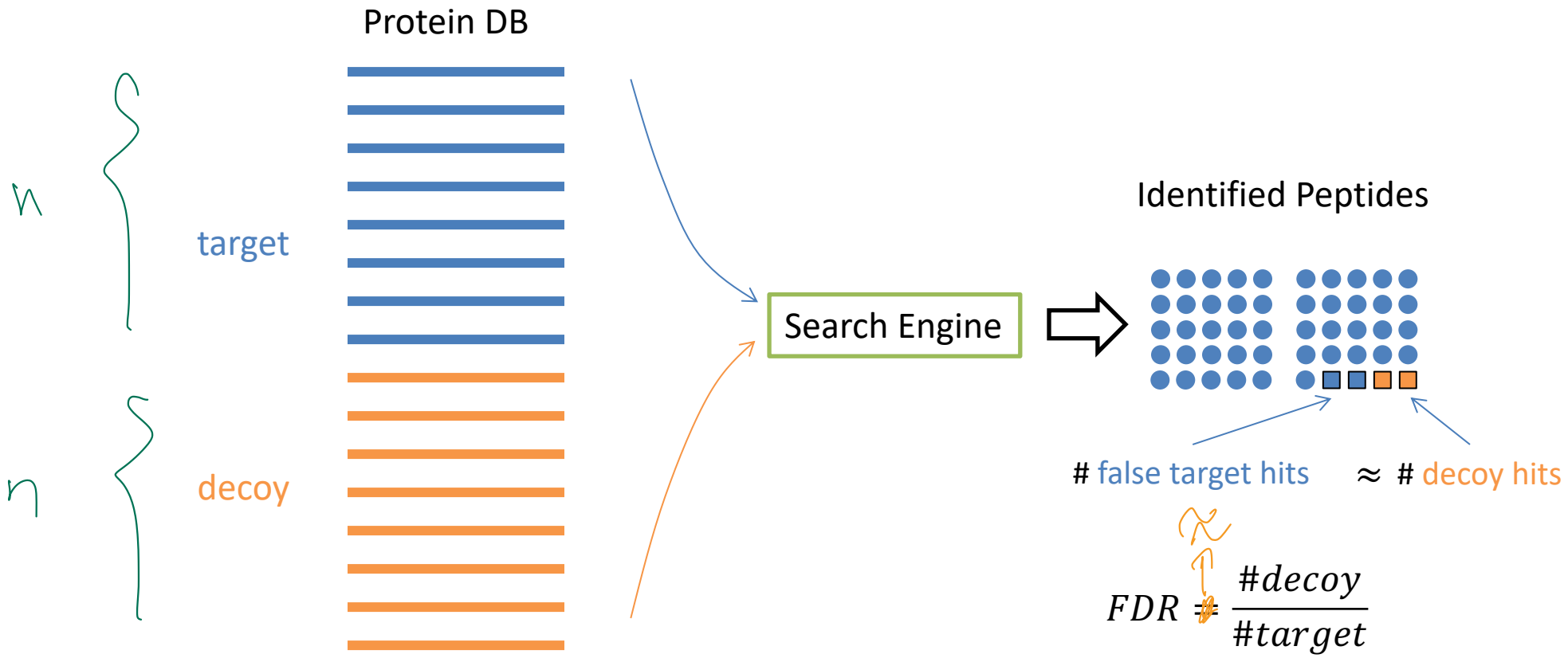


$$FDR = \frac{\#\text{reported false hits}}{\#\text{reported hits}}$$

False-Discovery Rate

- By choosing different score threshold, one can calculate the FDR for all target PSMs above the threshold. Or conversely, one can choose a proper threshold to meet a FDR requirement.
- As of today, a typical FDR requirement is 1%.
- Unfortunately, we only know the aggregated distribution (grey curve)

FDR Estimation with Target-Decoy



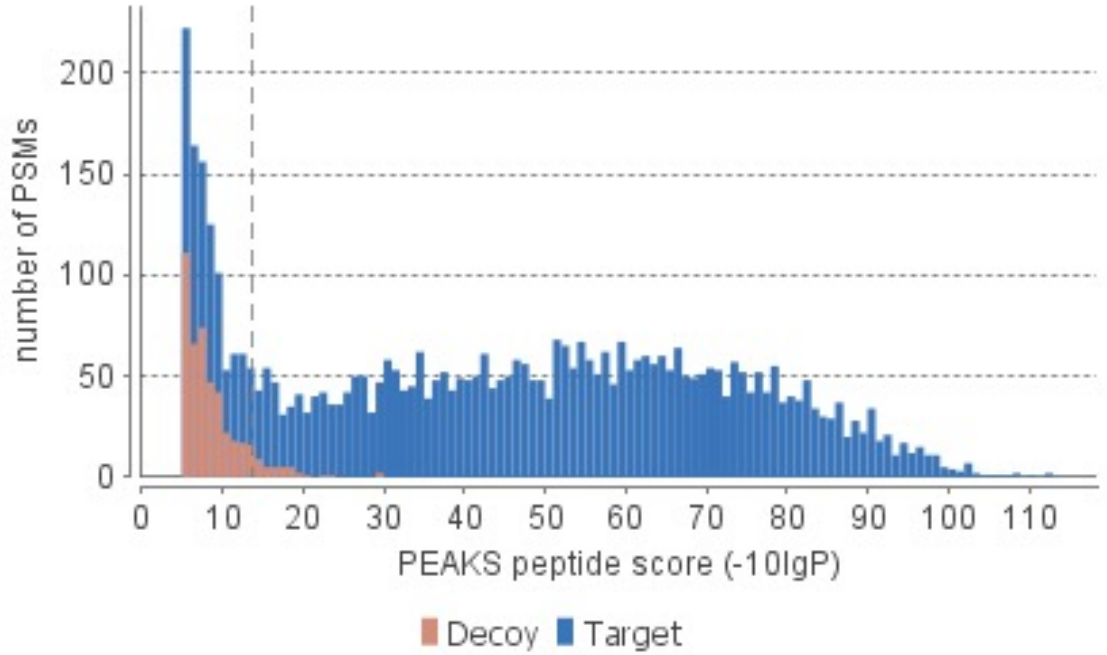
Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry

Joshua E Elias[1] & Steven P Gygi[1,2]

Nature Methods 4, 207 - 214 (2007)

FDR Estimation

Distribution of PSM scores



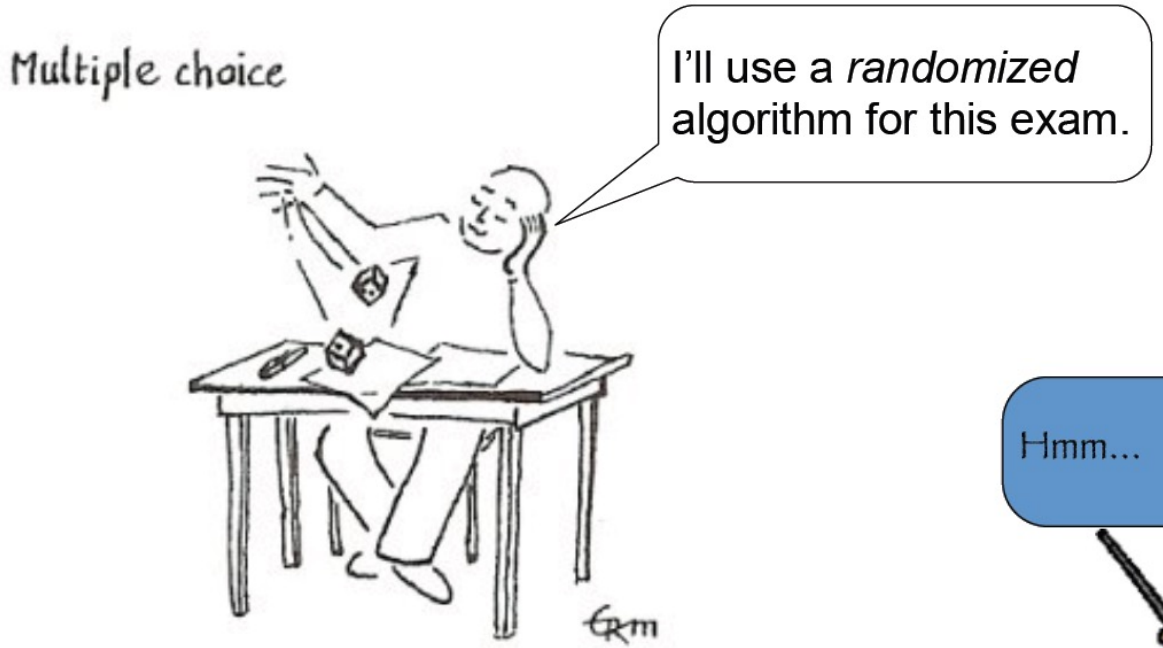
Question

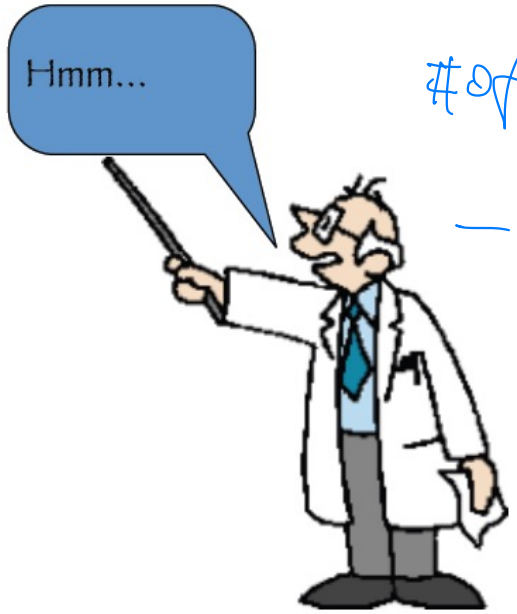
- Consider in assignment 2, we mix N true and N random peptides together.
- By using your score, from the N top-scoring peptides, there are about 0.6N true peptides and 0.4N random peptides.
- How many of the “true” peptides are selected because your scoring function is truly amazing, and how many are pure luck?

A Test

- Now you've learned about target-decoy method.
- Can you help me out in the following problem?
- Suppose I'm asked to make and mark a final exam as a substitute teacher. But I know absolutely nothing about the subject.
- Fortunately, I have previous year's exam and all questions are multiple choices. But I do not know the correct answers.

The Challenge





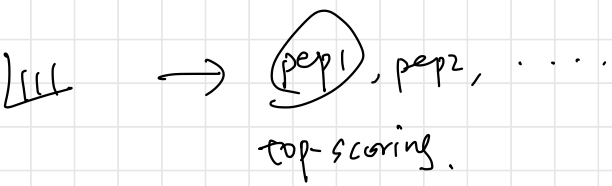
Review:

① database search

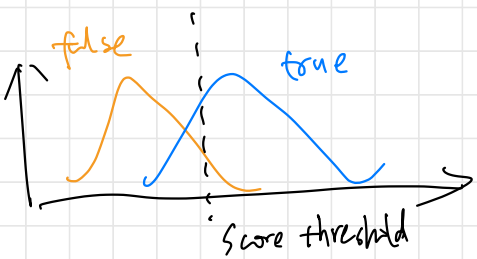
precursor/peptide mass

> protein 1
MTAKE....
> protein 2
....

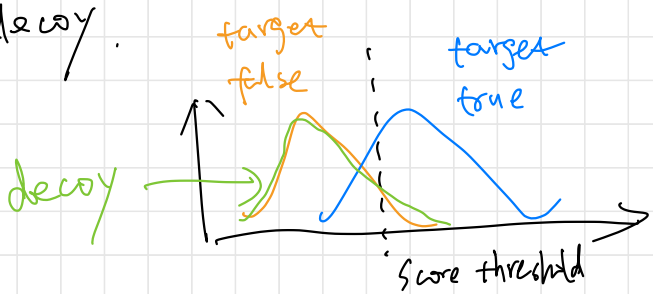
② PSM (peptide-spectrum match)



③ FDR (false-discovery rate).



④ target-decoy.



Roadmap:

- More details
  - better score
  - troubles. (PTM, speed, FDR).
- De novo sequencing.
- Quantification & Multiple Myeloma
- Spectrum prediction with deep learning.
- AlphaFold2
- classic bioinformatics

Better Scoring Function

- The empirical score is only good for start up.
- Soon competition will get fierce and you'll need a better scoring function.

Likelihood Ratio

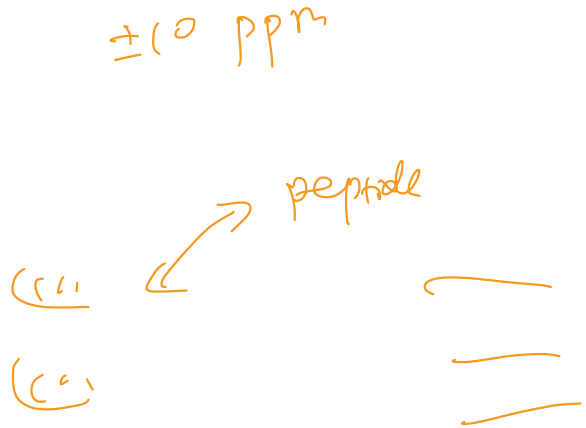
- Let m be the m/z of a y-ion, and indeed, we see a peak with m/z = m in the spectrum.
- Two assumptions:
  - The peptide is the real peptide so peak is caused by the y-ion.
    - Pr(observe a peak at m|m is a y-ion m/z of the real peptide)
  - The peptide is a random peptide so the match is purely by chance.
    - Pr(observe a peak at m|m is a random mass)

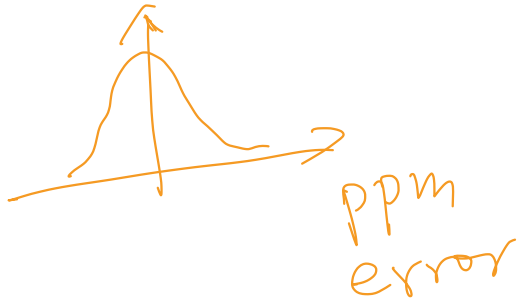
Log Likelihood Ratio

- Learn two probabilities from large training data
  - p : Prob(a peak is observed at a y-ion m/z).
  - q : Prob(a peak is observed at a random m/z).
  - Usually $p > q$.
- Given a peptide sequence, calculate m/z of all possible y-ions. For each y-ion,
  - If a peak observed, $\log \frac{p}{q}$ is added to score.
  - If no peak is observed, $\log \frac{1-p}{1-q}$, is added to score.
- Thus, matching ion is rewarded and missing ion is penalized.
- Other fragment ion types can be considered similarly, and added to the score.

Ideas of Even Better Scores

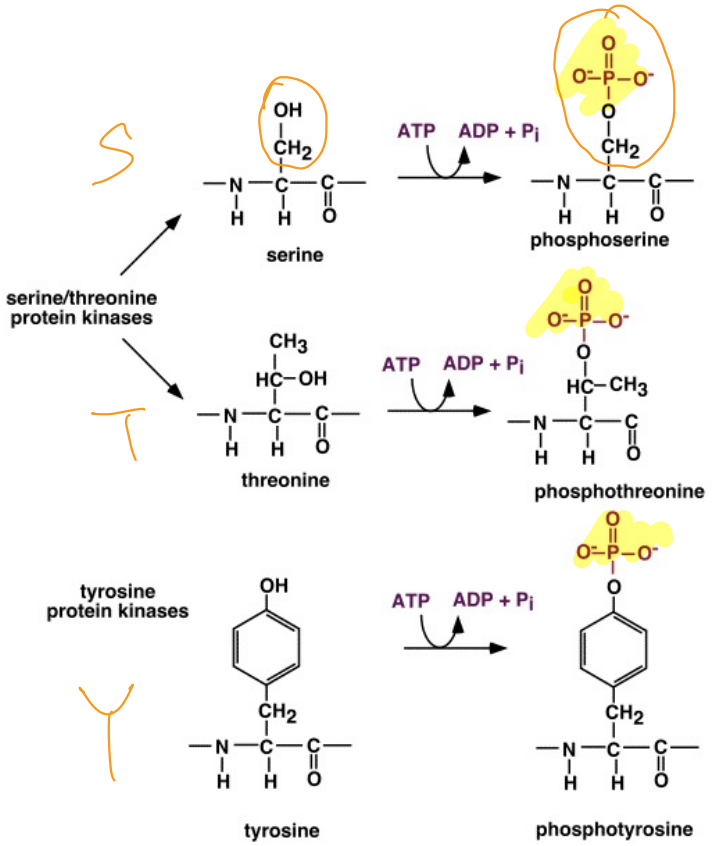
- Machine learning that combines many factors
  - Log likelihood ratio score
  - Empirical score (log of relative intensity)
  - Precursor error tolerance
  - Number of matching peaks
  - Number of unmatched peaks
  - Number of unmatched y-ions
  - Include b-ions.
  - Include charge 2 fragment ions.
  - Etc.





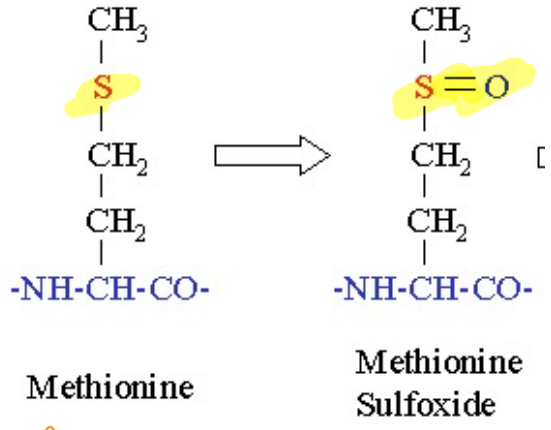
OTHER PRACTICAL CONCERNS

Post-Translational Modifications (PTM)



Phosphorylation (Δm = +80)

- PTM important to protein functions.
- Hundreds of different types of PTMs
- PTM normally change the mass of an amino acid.
- Some PTMs can be on and off.
- The figure shows two common types of PTMs.



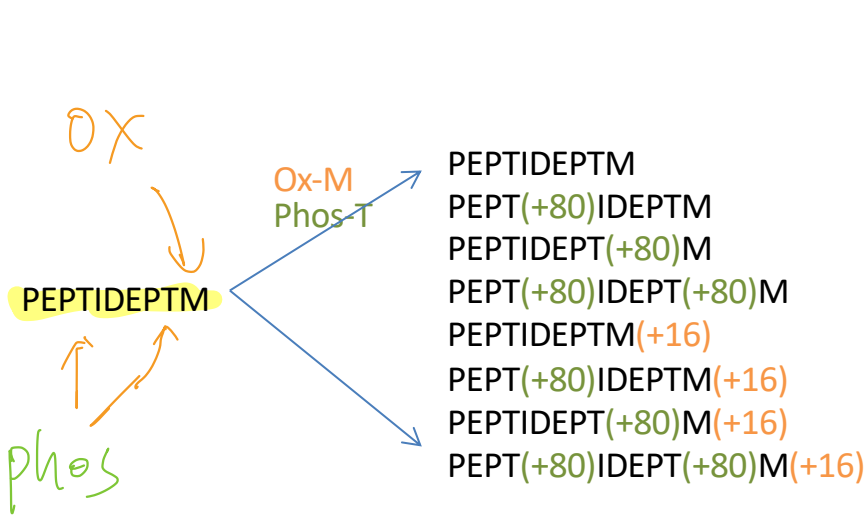
Oxidation (Δm = +16)

Post-Translational Modifications

- There are many hundreds of different types of PTMs included in the unimod PTM database.
- 30% of human proteins are phosphorylated, 50% are glycosylated.
- PTMs are important to the functions of proteins.
  - For example: Reversible phosphorylation of proteins is an important regulatory mechanism. Many enzymes are switched "on" or "off" by phosphorylation and dephosphorylation. The structural change caused by the PTM changes the function of the protein.

Variable PTMs

- If user selects some PTMs as “variable”, all possible modification forms of a database peptide need to be tried to match the spectra. This results in exponential growth of search space. E.g.

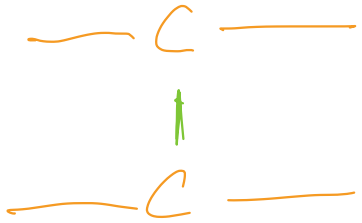


- Consequently, one can only search with a few variable PTMs.

Fixed PTMs

- Certain modifications are deliberately added during the sample preparation and is (almost) 100%. These are called fixed PTMs.
- The most common one is that cysteines are usually modified chemically. And the most common modification changes the mass from 103.00919 to 160.03065. Roughly 57.02 Da were added.
- Fixed modification changes the amino acid residue mass table, but does not affect the database search speed.

- For curiosity only, cysteines are modified to avoid the formation of “disulphide bonds”.



Missed and Nonspecific Cleavages

- The proteolyses may not be 100% efficient.
  - Assuming Trypsin digests the following protein with 100% efficiency
  - SSAYSR/GVFR/R/DTHK/SEIAHR/F
- Missed cleavages: a digestion site is not cut.
  - E.g. peptide GVFRR
- Non-specific cleavages: a non-digestion site got cut.
  - E.g. peptide SEIAH
- Allowing them will both affect the algorithm's time complexity.
  - Which one has a bigger impact?

$O(n^2)$

$O(N^2)$

N : length

n : cutting sites

Summary

- MS/MS data includes survey scans and MS/MS scans.
- Database search to assign peptides to MS/MS scans.
- Scoring functions.
- Target-decoy for FDR estimation.
- Practical issues:
  - Fixed and variable PTMs
  - Nonspecific cleavages