Database Search Details
Even Better Scoring Function

• Incorporate many other “features” for the scoring by a **machine learning** method.
• Features can apply to compute the matching/mismatching of certain fragment ion
  • Matched fragment ion intensities,
  • mass error
  • Correlation between intensity and surrounding amino acids
• Features can apply to the whole peptide score
  • Precursor ion mass error
  • **Agreement with de novo sequencing**
  • Protein information
Agreement with De Novo Sequencing

# matched amino acids
between de novo & DB search

Target  Decoy

best separation line

Mascot Score
Use Protein Information

**Idea:** Peptides on a multi-hit protein get a bonus on their scores to increase sensitivity.

A weak hit is “saved” due to the bonus.
Better Scoring Function

• More features make the score function better separate true and false matches.
Speed Concern

• General programming wisdom:
  • “Make it right, before make it fast.” (??)
  • “Premature optimization is the root of all evil” (Knuth)
Two-Round Search

- To further speed up the search,

Pitfalls in FDR Estimation
FDR Estimation

**Distribution of PSM scores**

[Graph showing the distribution of PSM scores with a peak at around 0 and a long tail extending to the right.]

**Corresponding FDR curve**

[Graph showing the FDR curve with an increase in FDR as the number of peptide-spectrum matches increases.]
Pitfall 1 – Multiple Round Search

Round 1. Fast Search

Round 2. More Sensitive Search


FDR underestimation.
Solution: Decoy Fusion

Decoy sequence append to each target protein.

Fast Search

Equal targets and decoys

More Sensitive Search

# false target hits \approx # decoy hits

PEAKS DB paper. MCP 2012.
Pitfall 2 – Mix Protein and Peptide ID

Idea: Peptides on a multi-hit protein get a bonus on their scores to increase sensitivity.

A weak hit is “saved” due to the bonus. So is this weak false hit.

Pitfall

More multi-hit proteins from target DB ⇒ more false hits are “saved” from target DB ⇒ FDR underestimation.
Solution: Decoy Fusion

Weak false hits are “saved” with approx. equal probabilities in target and decoy.

Got the sensitivity, but still estimate the FDR correctly.
Pitfall 3 – Machine Learning with Decoy

Idea: Re-train the coefficients of scoring function for every search after knowing the decoy hits.

Pitfall: Risk of over-fit. Machine learning experts only.

Search

target false hits

decoy hits

Adjust scoring function to remove decoy hits after search.

Fewer target false hits are removed ⇒ FDR underestimation
Solutions

1. Don’t use it.
   - Judges cannot be players.

2. Only use for **very** large dataset.

or

3. Train coefficients and reuse; don’t re-train for every search.
Wrap Up

• We’ve learned
  • Practical algorithmic concerns
  • Scoring function
  • Target-decoy result validation

• We’ve also learned
  • Scientists can make mistakes
  • In programming we call these mistakes bugs