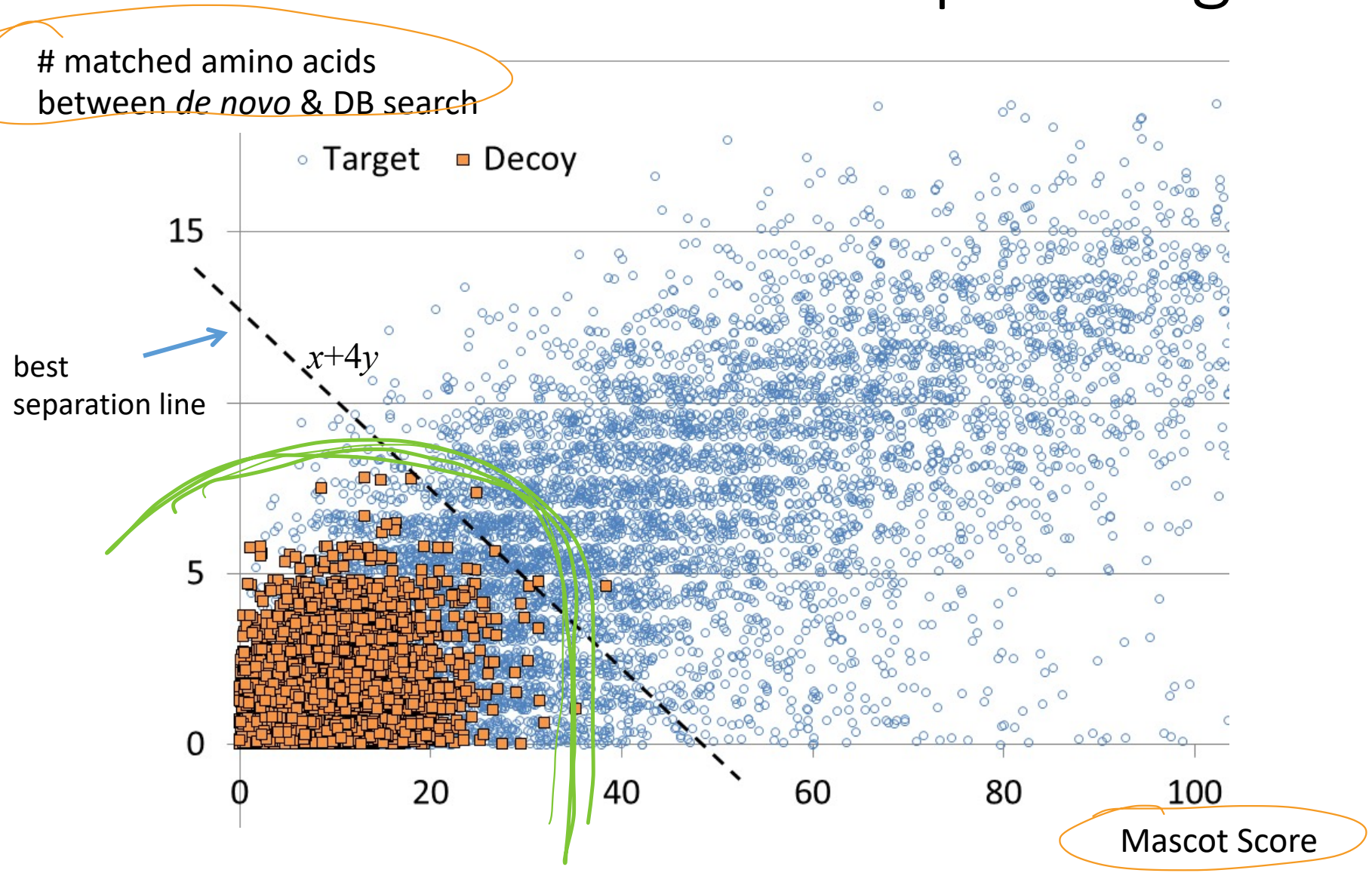
Database Search Details

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

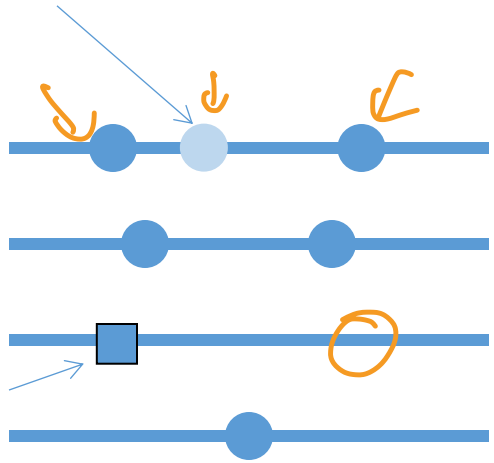


Use Protein Information

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

peptide - spec match

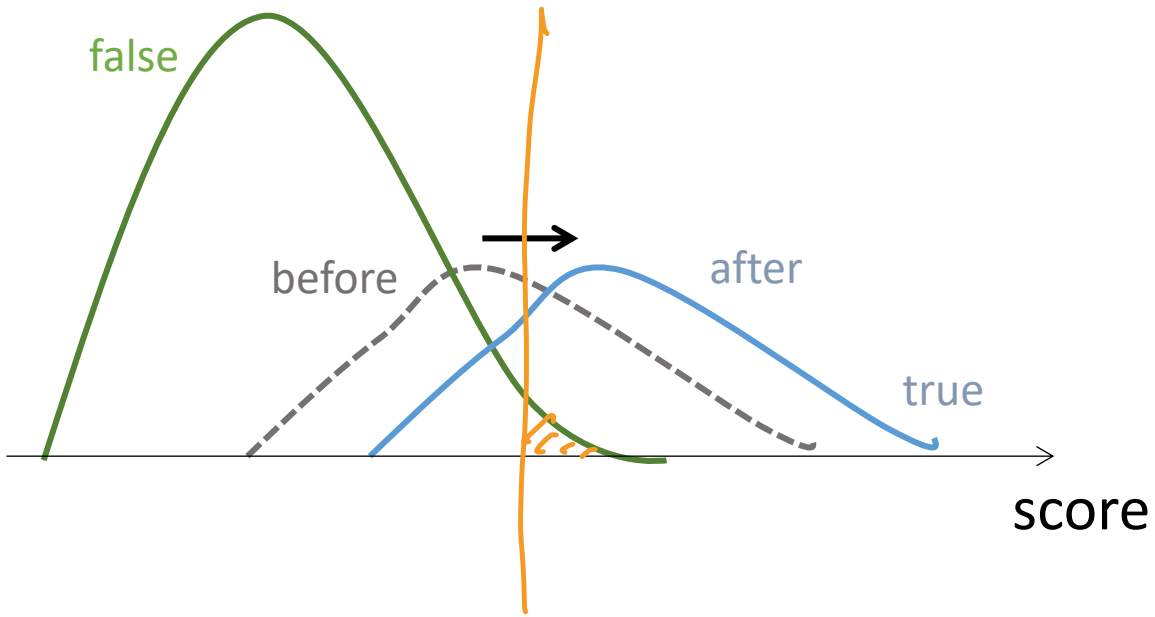
A weak hit is "saved" due to the bonus.



false hit

Better Scoring Function

- More features make the score function better separate true and false matches.

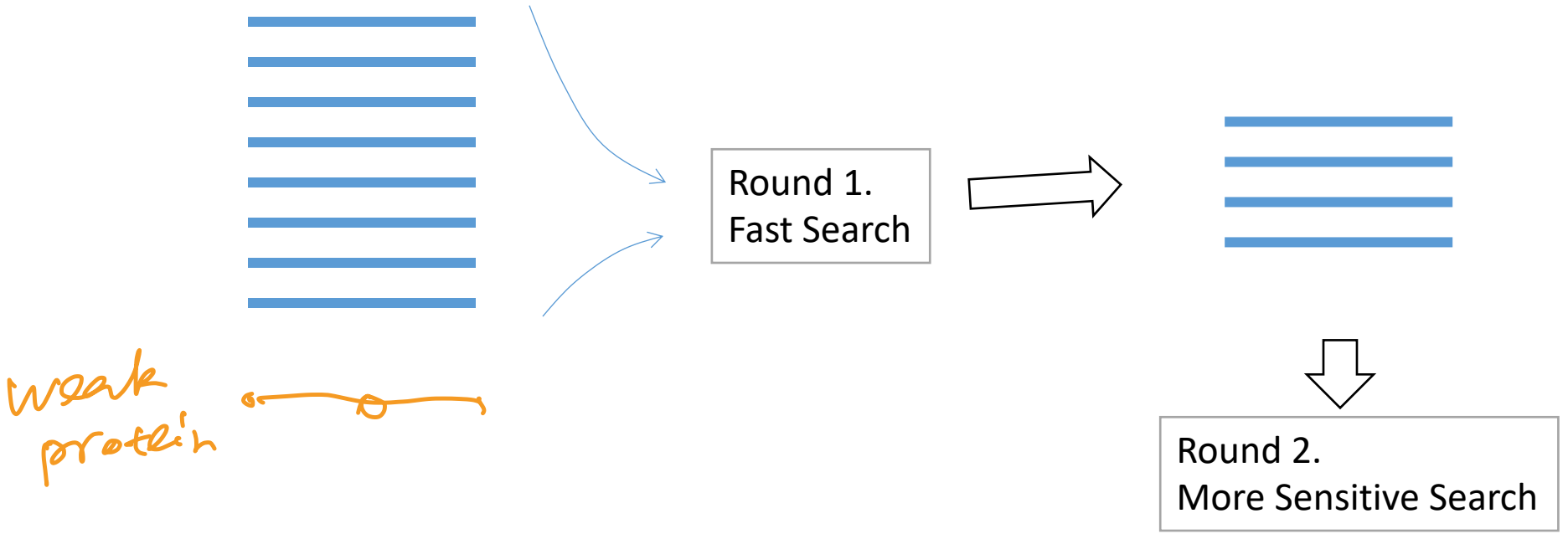


Speed Concern

- 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,

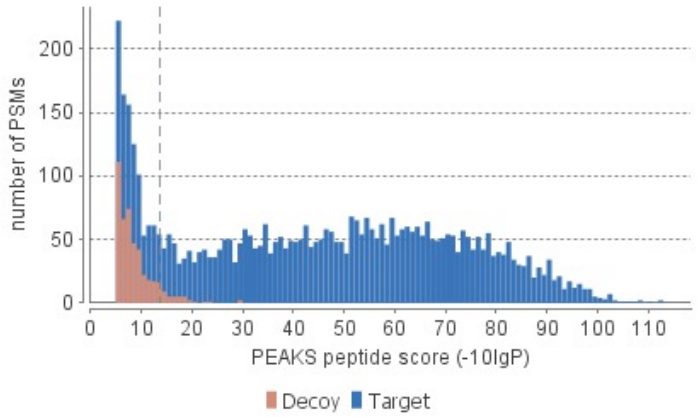


Craig and Beavis 2004. *Bioinformatics* 20, 1466–67.

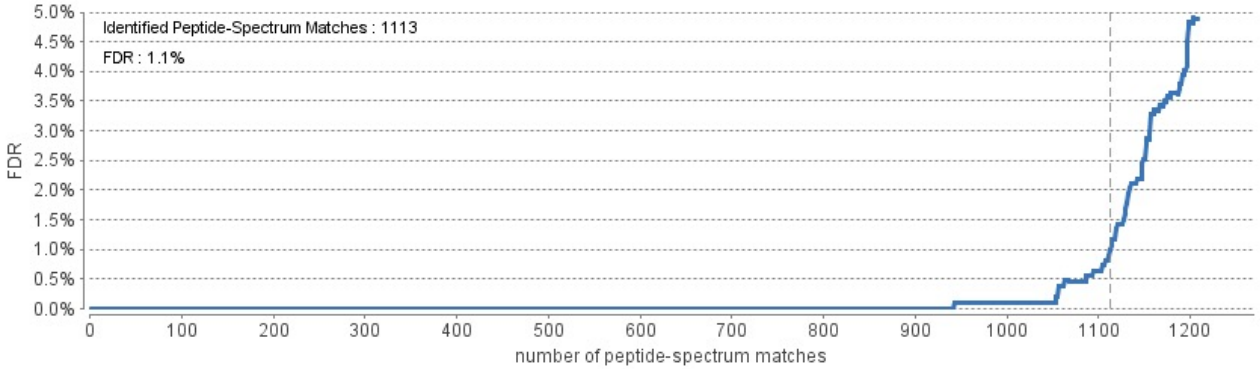
Pitfalls in FDR Estimation

FDR Estimation

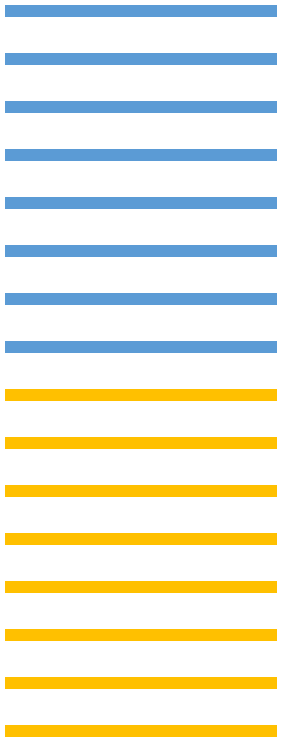
Distribution of PSM scores



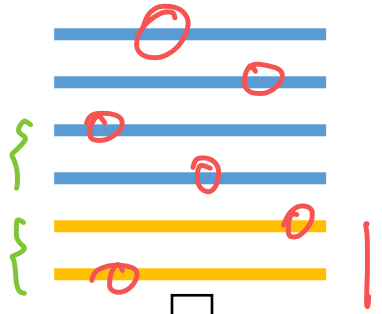
Corresponding FDR curve



Pitfall 1 – Multiple Round Search

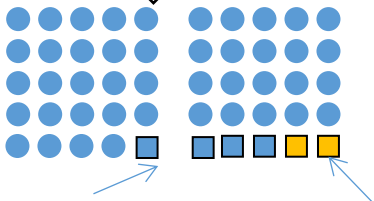


Round 1.
Fast Search



more targets
than decoys

Round 2.
More Sensitive Search



false target hits > # decoy hits

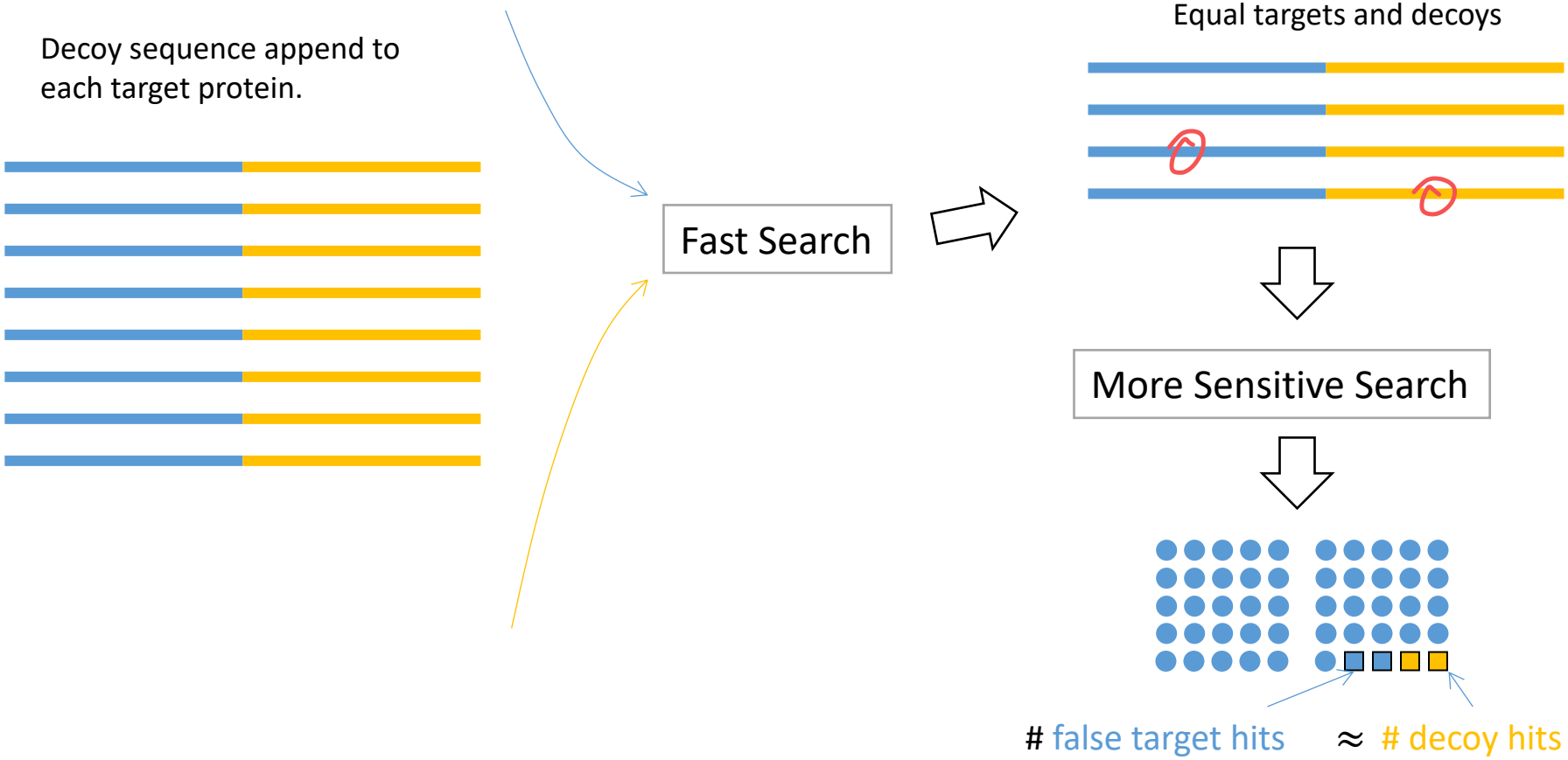
FDR underestimation.

Craig and Beavis 2004. *Bioinformatics* 20, 1466–67.

Evertt et al. 2010. *J Proteome Res*. 9, 700-707.

Bern and Kil 2011, *J Proteome Res*. 10, 2123-27.

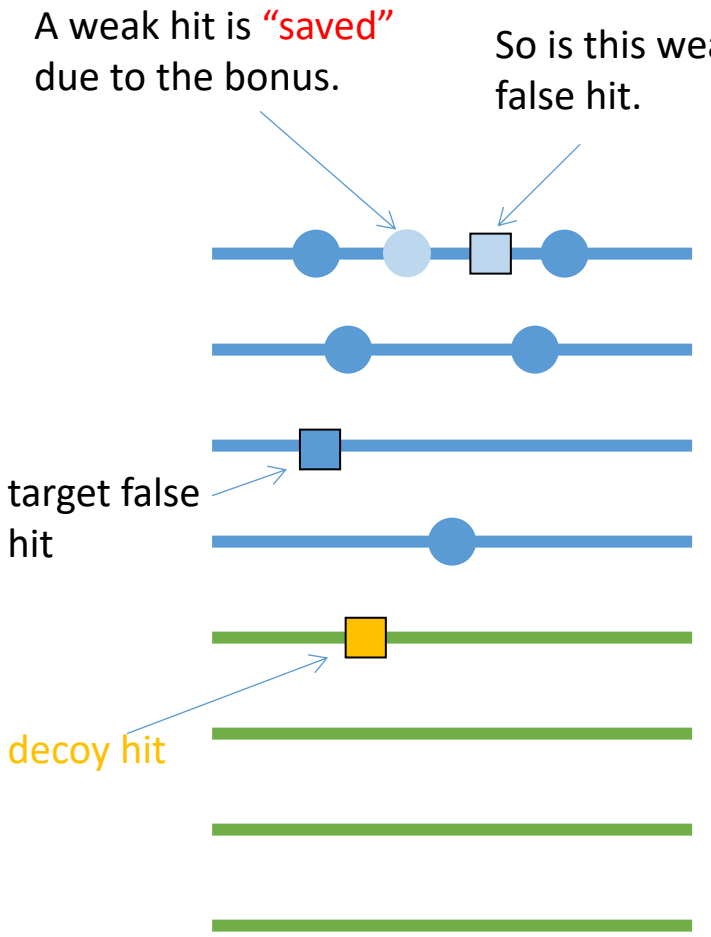
Solution: Decoy Fusion



PEAKS DB paper. MCP 2012.

Pitfall 2 – Mix Protein and Peptide ID

○ true
□ false



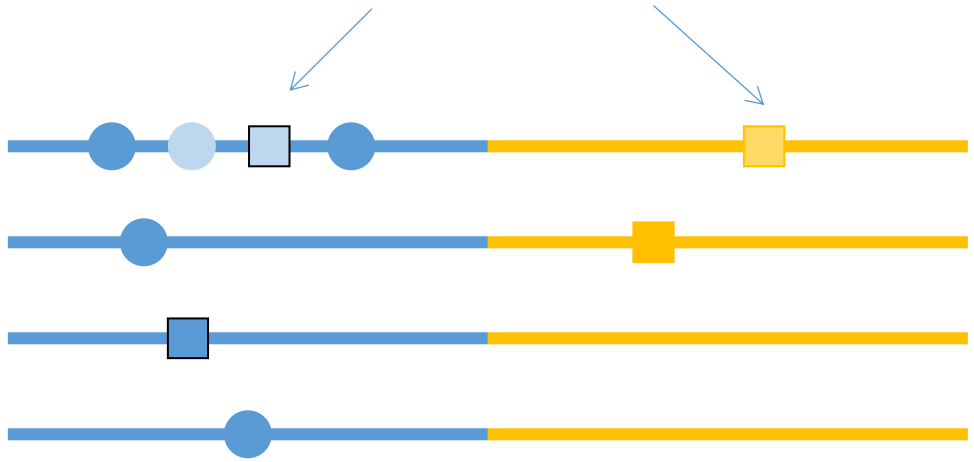
Idea: Peptides on a multi-hit protein get a bonus on their scores to increase sensitivity.

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.

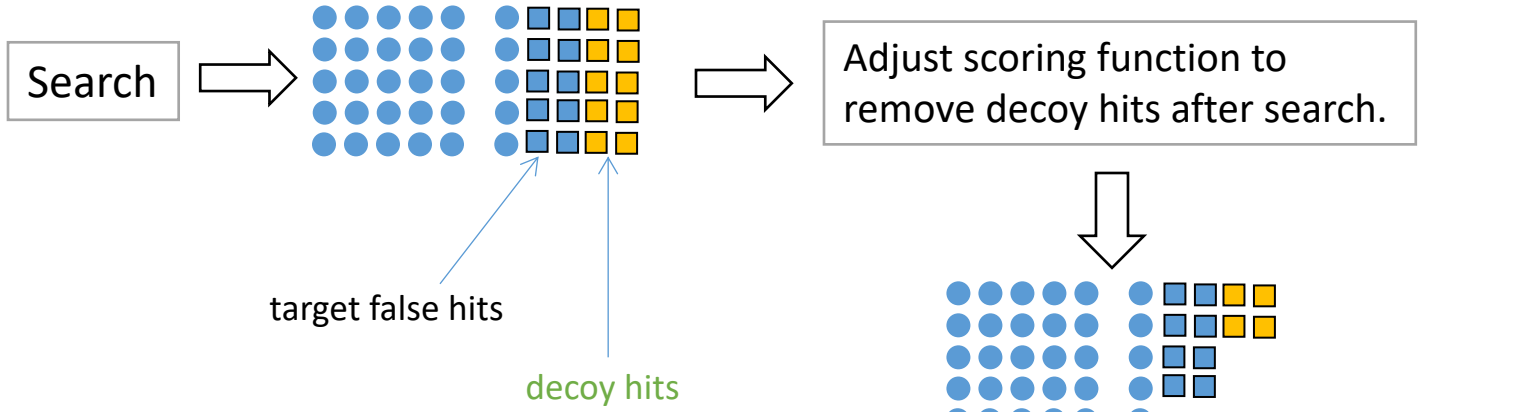


Got the sensitivity, but still estimate the FDR correctly.

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.



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.

or

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