More Accurate Control of the False Discovery Rate in Mass Spectrometry Based Peptide Identification

Jing Zhang[1], Lei Xin[1], Bin Ma[2]
[1] Bioinformatics Solutions Inc, Waterloo, ON
[2] David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON

Background

The large volume of mass spectrometry data requires a reliable automated method for quality control of the peptide identified by software and submitted to public databases. The commonly used target-decoy method estimates the false discovery rate (FDR) of the software's results. However, in this abstract we illustrate that the target-decoy method makes some unrealistic assumption about the analytical software, and is critically over-confident. We further propose a decoy-fusion method to solve this problem.

Method

The existing target-decoy method appends a same-length decoy database with the target, and search in this joint database. The FDR is calculated as the ratio between the numbers of decoy and target identifications (Figure 1). This requires the false identifications to be evenly distributed in the target and decoy.

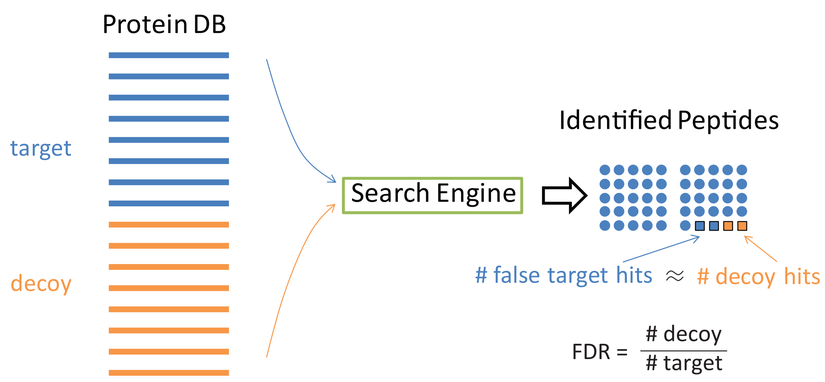

Figure 1. Conventional target-decoy method for FDR estimation.

However, today's peptide identification software often utilizes the protein information to help identify peptides in order to achieve better search sensitivity. This causes that a false peptide from the target sequences has a higher chance to be identified than a false peptide from the decoy sequences, and leads to the over-confidence problem. Many optimization techniques in the search engine can cause this effect and Figure 2 highlights one of such possibilities.

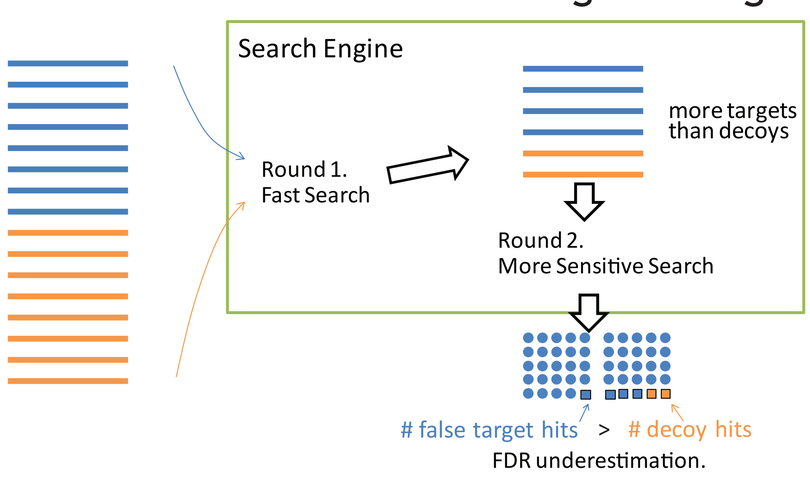

Figure 2. A multi-round search strategy commonly used in today's search engine will select more target proteins than decoy proteins in its first round. Thus the second round will produce fewer false identifications from the decoy than from the target, causing FDR underestimation.

We propose a decoy-fusion method to solve this problem. A same-length decoy sequence is appended to each target protein to become a new fused protein. The fused database contains all the fused proteins. Peptides identified from the first and second halves of each sequence are regarded as target and decoy identifications, respectively. Because each fused protein contains the same length of target and decoy, the protein information will affect the target and decoy identifications equally. This recreates the balance and avoids the over-confidence (Figure 3).

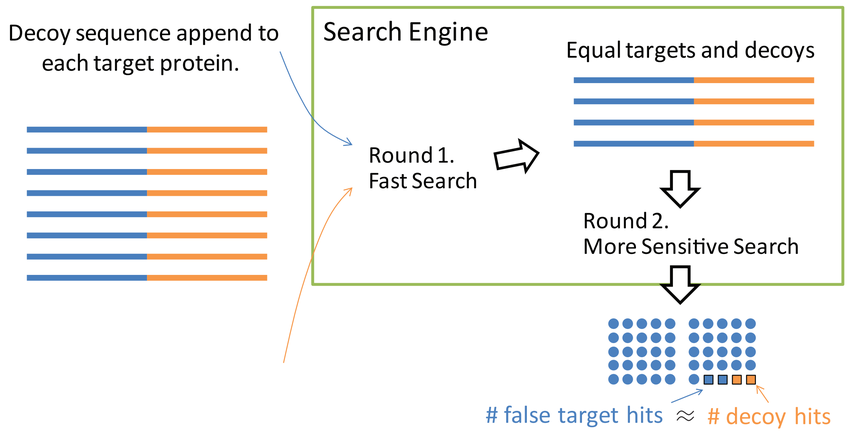

Figure 3. By appending the decoy to the end of target for each protein, the false identifications will fall equally in the target and decoy, even if a multi-round search approach is used.

Conclusion

The current target-decoy method for result quality control is over-confident. A new decoy-fusion method was proposed to solve this problem.