Combine Multiple Database Search Engines with Uniform FDR

Bioinformatics Solutions Inc.

Lei Xin[1], Brian Munro[1], Bin Ma[2]
Bioinformatics Solutions Inc, Waterloo, ON[1]
University of Waterloo, Waterloo, ON[2]

Overview

- A new method for combining multiple search engines is proposed.
- The method does not depend on any prior knowledge of the search engines, thus no training is needed for this method to add a new search engine.
- The method naturally comes with controlled FDR.

Introduction

For peptide identification with mass spectrometry, multiple database search software packages are available. Each of these search engines uses its own scoring function, which gives them complementary abilities in assigning different spectra from the same MS/MS dataset. Therefore, combining multiple engines' results helps improve sensitivity. But the different scoring functions also impose a challenge in result validation. Previous methods that use empirical probabilistic models are not consistent with the generally accepted false discovery rate (FDR) measurement for validation. We propose a new method to combine the multiple engines according to their respective FDR. The new method does not require any parameter training and supports the adding of a new search engine without any prior knowledge of its scoring function.

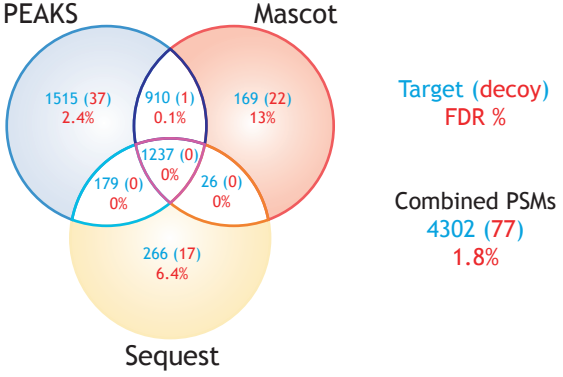

Figure 1. Venn diagram by setting 1% FDR for each individual search engine

Method

Multiple search engines are used to search against the same target-decoy database. The FDR curve for each search engine is determined by the target and decoy hits, without knowing the exact scoring function of each search engine. Suppose k engines are used and the desired FDR is x. An initial uniform FDR cut-off for each individual engine is set. Each engine's confident identifications passing the uniform FDR cut-off are combined. The target and decoy hits in the combined result are used to estimate the combined FDR. If the combined FDR is below (or above) the desired x, the initial uniform FDR is adjusted, increased (or reduced) accordingly. This process is iterated until the combined FDR is approximately equal to the target FDR.

Results

A public data set from the ABRF iPRG-2011 study was used to test the accuracy of this method. This LC-MS/MS dataset was generated from a Lys-C digest of a yeast lysate following SCX peptide fractionation.* Of 15 fractions collected across the SCX gradient, the dataset for this study constitutes only fraction 10. These fractions were then analyzed by a Thermo LTQ-Orbitrap XL with ETD fragmentation. For fraction 10, a total of 8031 MS/MS spectra were generated. In the ABRF iPRG-2011 study, a yeast database with 6666 protein sequences was also provided with the dataset. We generate our target-decoy database based on this small database. We use PEAKS [1], MASCOT [2] and SEQUEST [3] to search the spectra set against the target-decoy database. Figure 1 shows the Venn diagram of combining the three search engines at their own 1% FDR. Since overlap exists between different engines, the FDR for this combined result is 1.8%, which is higher than the respective FDR for each individual engine as shown in Figure 2. To get 1% FDR for the combined result, our method automatically determines the FDR for each individual engine to be 0.6% as shown in Figure 3. Thus we get 4168 PSMs with 40 decoy hits for the combined result as showed in Figure 4.

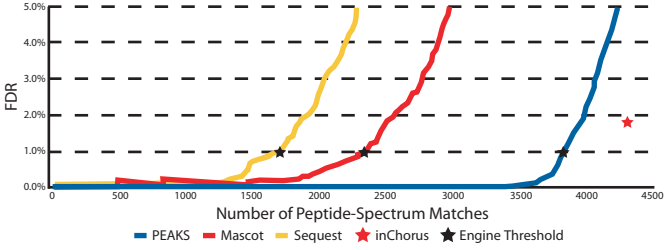

Figure 2. Combined FDR is 1.8% when each individual engine's FDR is set to be 1%

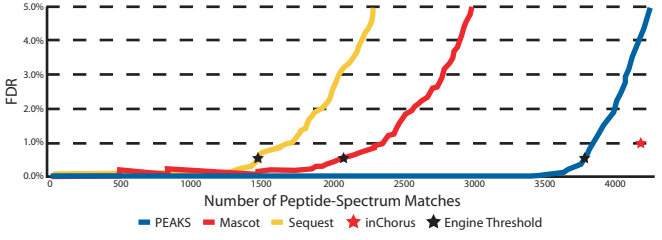

Figure 3. Individual search engine FDR is set to be 0.6% to achieve 1% combined FDR

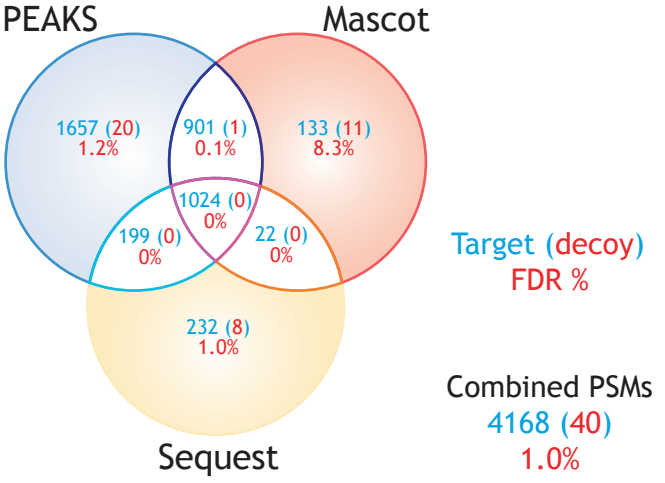

Figure 4. Venn diagram by setting 1% combined FDR

Conclusion

By combining different search engines with uniform FDR, our method improved the sensitivity of the search results while keeping the FDR of the combined result under control. This method has been implemented in the PEAKS 6 inChorus function.

References

[1]. J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G. Lajoie, and B. Ma. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. Molecular & Cellular Proteomics. Accepted. 2011.

[2]. D.N. Perkins, D.J.C. Pappin, D.M. Creasy, J.S. Cottrell. Probability-based protein identifcation by searching sequence databases using mass spectrometry data. Electrophoresis, 20(18):3551-3567, 1997.

[3]. J.K. Eng, A.L. McCormack, J.R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Am.Soc Mass Spectrom 5(11):976-989, 1994.

Data - The undigested Saccharomyces cerevisiae lysate, Reference Material (RM) 8323, was obtained from the National Institute of Standards and Technology (NIST) an described at https://www-s.nist.gov/srmors/view_report.cfm?srm=8323.