A Robust and Effective Strategy for Combining Results of Multiple Peptide Identification Engines

Bioinformatics Solutions Inc.

University of Waterloo

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

Background

Many software packages have been developed for identifying peptides from mass spectrometry data. Their abilities are often complementary to one another. It is therefore useful to combine multiple search engines' results to improve the overall peptide identification performance.

Empirical statistical methods have been developed for unifying the scores of different engines and combining the results together. Implementations of these methods include the Trans-Proteome Pipeline [1] and the Scaffold software [2]. While these methods have contributed greatly to proteomics research, the complexity of the statistical model makes it difficult or impossible to add a new search engine by an end-user.

We propose a simple model for combining multiple engines' results and demonstrate its effectiveness.

Method

Suppose *n* search engines are used. The method first generates a combined database by concatenating the target and decoy databases. Then, each engine is used to search in this combined database separately. By the standard target-decoy approach, the FDR of each engine's results above a certain score threshold is estimated. In the consensus report, we keep those peptide-spectrum matches (PSMs) that satisfy any of the following two conditions:

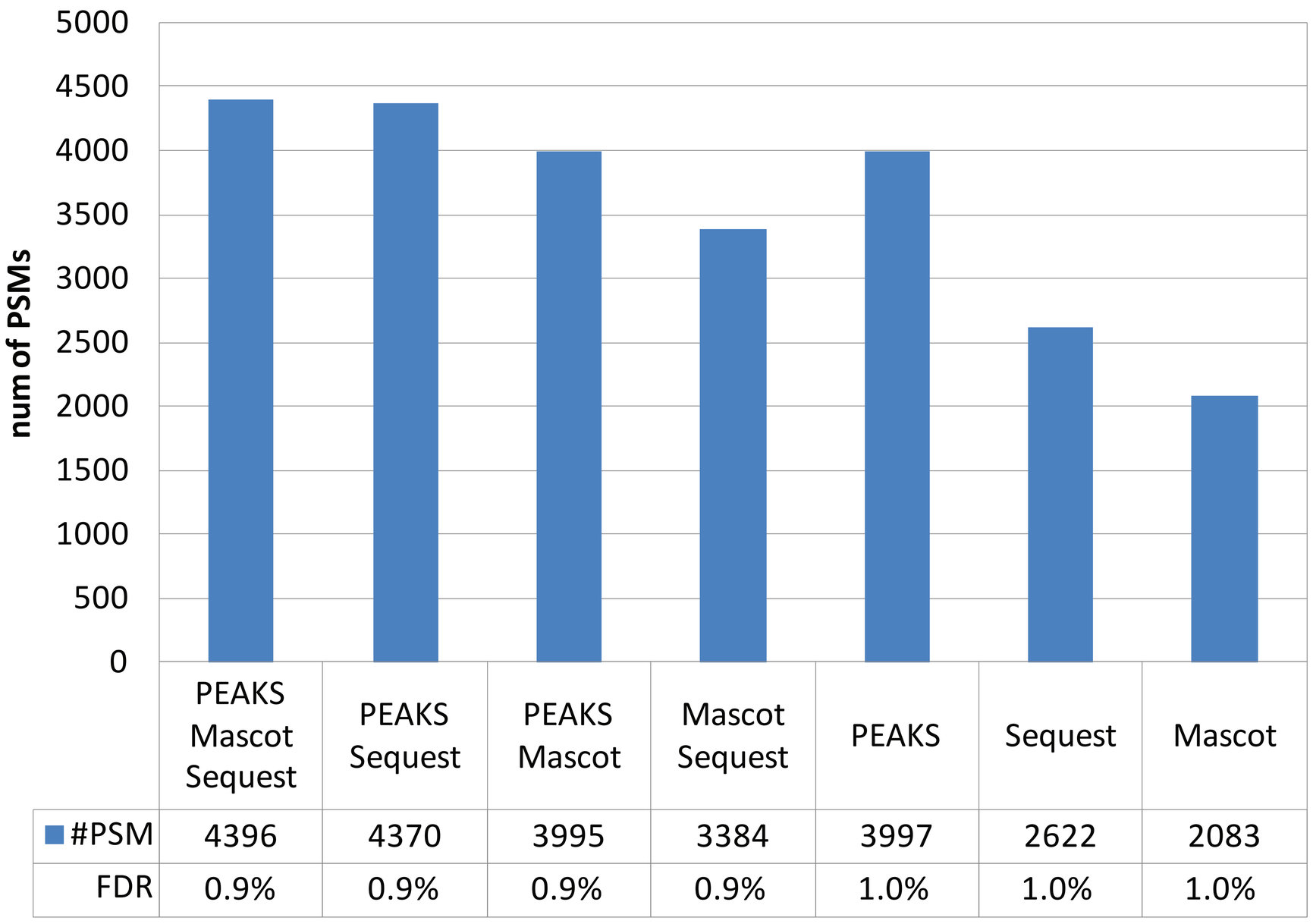
(1) It is identified by any engine with FDR ≤ 0.01/*n*, or
(2) It is identified by at least two engines, each with FDR ≤ 0.05.

The FDR of the consensus report can also be estimated by the target-decoy method.

Results

Mascot, Sequest, and PEAKS were used to analyze an Orbitrap dataset of 7743 CID MS/MS spectra of lysed C-Elegans. The performances (reported PSMs and FDR) of different combinations of the three engines are shown in Figure 1.

Figure 1. The combination of the three engines identified more than twice as many PSMs as using Mascot alone. If only two search engines are allowed, then Sequest+PEAKS provide the optimal combination.



	PEAKS Mascot Sequest	PEAKS Sequest	PEAKS Mascot	Mascot Sequest	PEAKS	Sequest	Mascot
#PSM	4396	4370	3995	3384	3997	2622	2083
FDR	0.9%	0.9%	0.9%	0.9%	1.0%	1.0%	1.0%

Conclusion

The new model for combining multiple peptide identification engines' results works effectively. The new model has the advantage of not requiring any additional training when a new software tool becomes available.

References

[1] Deutsch, E. W. et al., A guided tour of the Trans-Proteomic Pipeline. Proteomics 2010, 10, 1–10.

[2] Searle, B. C., Turner, M., Nesvizhskii, A. I., Improving sensitivity by probabilistically combining results from (multiple MS/MS search methodologies. J. Proteome Res. 2008, 7, 245–253.