Influence of Target and Decoy Database Size on Peptide Identification Confidence

Franziska Zickmann, Bernhard Y. Renard
Research Group Bioinformatics (NG 4), Robert Koch-Institute, Berlin, Germany

Introduction: The reliable identification of peptides and proteins from tandem mass spectra is one of the main challenges in the field of proteomics. To ensure specificity of the results, the computation of false discovery rates (FDRs) based on target-decoy databases has become the standard [1]. The choice of target and decoy database is crucial for peptide identification, however, still little attention has been paid to the impact of database size and target to decoy size ratio on the accurateness and stability of FDRs. Hence, we analyse various combinations of target and decoy databases of different sizes and their effect on the FDR computation.

Methods: We used a previously described dataset consisting of several thousand high confidence spectra of helacells [2] and searched it with InSpecT and Sequest against a variety of databases created based on IPI human protein sequences. To simulate the effect of inappropriate small and large databases, we further sub-sampled databases of 10% and 50% of the original size and added protein sequences from two other species, chicken and *Arabidopsis*.
Reversed decoy databases have been created at sizes from 10% to 200% of the original protein sequences, resulting in a total of 16 combinations of database sizes. Each sampling has been performed 20 times to also obtain variance estimates of the FDR for the various database combinations.

Results: The influence of target to decoy database ratio differs between InSpect and Sequest. For InSpecT, we observed a general trend for decreased variances with increased decoy database sizes, whereas Sequest showed no clear relationship between the size of the decoy database and the variance. For both tools, we conclude that variance is more influenced by unbalanced ratios than by size alone, since smaller target samples of 10% and 50% size are less affected by changes in the decoy size.
In addition, we saw that for the target database of 10% of the original size the number of identified peptides is strongly reduced and the cut-off score is increased compared to the complete database. For the artificial expanded database no strong effect can be observed.

Conclusions: All in all, the simulation demonstrated that the reliability of FDRs can be influenced by the ratio of target and decoy database and their absolute size. Sequest and InSpecT showed different sensitivities to the various combinations of database sizes, indicating that robustness to the target-decoy choice is tool-specific and results cannot be generalized across search engines.
To ensure stability of results, the size of the decoy database should be equal to or

larger than the size of the target database. Otherwise a decoy hit is weighted higher than target hits, and inaccuracies are aggravated.

Also the absolute size of the target database should preferably be chosen larger than smaller. As expected, for the original database we generally observe the best stability. However, in case that the optimal database is unknown or not available, a larger database should be preferred over a smaller size, since we saw a negative effect on cut-off score and peptide number for sub-sampled databases, compared to moderate effects for the artificially expanded database size.

References

[1] Ralph A Bradshaw, Alma L Burlingame, Steven Carr, and Ruedi Aebersold. Reporting protein identification data: the next generation of guidelines. *Mol Cell Proteomics*, 5(5):787–788, May 2006.

[2] Bernhard Y Renard, Marc Kirchner, Flavio Monigatti, Alexander R Ivanov, Juri Rappsilber, Dominic Winter, Judith A J Steen, Fred A Hamprecht, and Hanno Steen. When less can yield more - computational preprocessing of ms/ms spectra for peptide identification. *Proteomics*, 9(21):4978–4984, Nov 2009.