BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 21 2013, pages 2799–2800
doi:10.1093/bioinformatics/btt490

Gene expression Advance Access publication August 20, 2013

ProteoStats—a library for estimating false discovery rates in proteomics pipelines

Amit Kumar Yadav*, Puneet Kumar Kadimi, Dhirendra Kumar and Debasis Dash*

G.N.R. Knowledge Center for Genome Informatics, CSIR-Institute of Genomics and Integrative Biology, New Delhi-110020, India

Associate Editor: Janet Kelso

ABSTRACT

Summary: Statistical validation of peptide assignments from a large-scale shotgun proteomics experiment is a critical step, and various methods for evaluating significance based on decoy database search are in practice. False discovery rate (FDR) estimation of peptide assignments assesses global significance and corrects for multiple comparisons. Various approaches have been proposed for FDR estimation but unavailability of standard tools or libraries leads to development of many in-house scripts followed by manual steps that are error-prone and low-throughput. The ProteoStats library provides an open-source framework for developers with many FDR estimation and visualization features for several popular search algorithms. It also provides accurate *q*-values, which can be easily integrated in any proteomics pipeline to provide automated, accurate, high-throughput statistical validation and minimize manual errors.

Availability: https://sourceforge.net/projects/mssuite/files/ProteoStats/.

Contact: ddash@igib.res.in or aky.compbio@gmail.com or amit.yadav@igib.in

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 11, 2013; revised on August 2, 2013; accepted on August 14, 2013

1 INTRODUCTION

High-throughout shotgun proteomics generates millions of spectra that are assigned to peptides by database searching in an automated manner. A scoring function is used to calculate the level of similarity found between the experimental spectrum and the theoretical spectrum from a peptide. This quantitative value, known as the score, may be a probability, an expectation value or a simple score.

Owing to random matching of peaks in a peptide spectral match (PSM), there is a fuzzy boundary between true and false matches leading to sensitivity versus specificity trade-off at any chosen threshold score. Conventional metrics like *P*-value/*E*-value are intended for significance assessment of a single hit (or PSM) but are not suitable for global significance assessment in large-scale datasets. A multiple testing correction procedure known as *false discovery rate* (FDR) is applied to the results for controlling the false positives globally. Target-decoy (TD) search-based FDR estimation is used as a standard method for global significance assessment (Elias and Gygi, 2007; Kall *et al.*, 2008).

Several challenges can arise in (i) dealing with multiple non-standard file formats, (ii) presence of correct peptides in decoy results, which should be removed prior to FDR calculation and (iii) *q*-value calculation. Leucine/Isoleucine cannot be distinguished by collision-induced dissociation, resulting in identical scores for both target and decoy peptides. Such peptides need to be removed from decoy hits. This is a non-trivial exercise if carried out manually but critical for accurate estimation of FDR because even one such peptide may penalize hundreds of matches in target database. The calculation can be tedious for large datasets and numerous files.

FDR estimation has recently been a major field of research in statistical proteomics, and several methods and their refined variants have been proposed. Unfortunately, not as many tools are available as there are FDR estimation methods. Most methods involve custom in-house scripts that may not implement all nuances of the algorithm and thus lack provenance. Custom pipelines only include the most popular methods like concatenated (Elias and Gygi, 2007) or separate (Kall *et al.*, 2008) methods, even though the refined methods improve the results. FDRAnalysis (Wedge *et al.*, 2011) provides FDR for three algorithms but only provides FDR for concatenated search. The barrier to their widespread adoption is the lack of software. This may also lead to different implementations of same method, code duplication and a considerable waste of research effort.

2 DESCRIPTION OF PROTEOSTATS

To address the issues mentioned earlier, we have developed an open-source cross-platform scripting library, ProteoStats, which uses several FDR estimation procedures. The library is written in Perl and can be easily interfaced with other pipelines. The library also provides a simple programming interface for visualizing these results and quality control of PSMs. ProteoStats supports OMSSA (Geer *et al.*, 2004), MassWiz (Yadav *et al.*, 2011b), Mascot (Perkins *et al.*, 1999), X!Tandem (Craig and Beavis, 2004), Sequest (Eng *et al.*, 1994), MyriMatch (Tabb *et al.*, 2007) and Comet (Eng *et al.*, 2013), besides the generic PepXML and text file formats (Fig. 1).

TD searches can be conducted as follows:

- Separate searches: The spectra are searched separately against the target and decoy databases independent of each other. Each spectrum has two best hits—one from the target and another from decoy.

*To whom correspondence should be addressed.

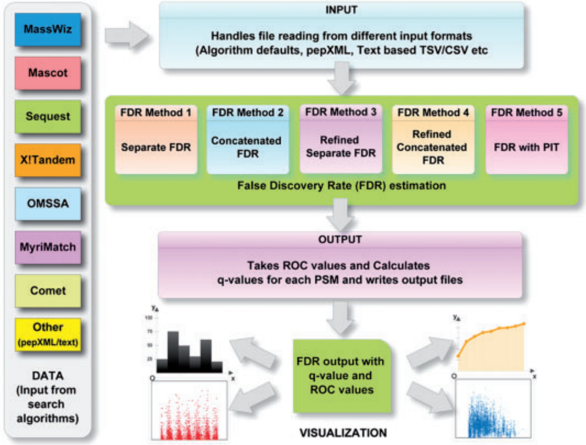
MassWiz
Mascot
Sequest
X!Tandem
OMSSA
MyriMatch
Comet
Other (pepXML/text)
DATA (Input from search algorithms)
INPUT
Handles file reading from different input formats (Algorithm defaults, pepXML, Text based TSV/CSV etc
FDR Method 1
Separate FDR
FDR Method 2
Concatenated FDR
FDR Method 3
Refined Separate FDR
FDR Method 4
Refined Concatenated FDR
FDR Method 5
FDR with PIT
False Discovery Rate (FDR) estimation
OUTPUT
Takes ROC values and Calculates q-values for each PSM and writes output files
FDR output with q-value and ROC values
VISUALIZATION

Fig. 1. Overview of ProteoStats library design

- Concatenated searches: The spectra are searched in a single concatenated database formed by combining the target and decoy proteins. The target and decoy candidate peptides compete against each other. Each spectrum will be assigned either a target or a decoy peptide as the best hit.

The two approaches are different only in the context of TD competition, which can be held *a posteriori* (Fitzgibbon *et al.*, 2008). Because it is easy to perform the TD competition even after a separate search, ProteoStats library requires results from separate search irrespective of the FDR method a user wants to select.

The ProteoStats library supports many formulations on FDR:

- Concatenated FDR (Elias and Gygi, 2007)
- Separate FDR (Kall *et al.*, 2008)
- FDR with percentage of incorrect target correction (Kall *et al.*, 2008)
- Refined FDR on separate method (Navarro and Vazquez, 2009)
- Refined method on concatenated FDR (Cerqueira *et al.*, 2010)

The ProteoStats library estimates FDR using the desired method in a series of steps that involve (i) native result file reading and conversion to text format, (ii) removing decoy peptides identical to target peptides, (iii) preparing target and decoy score arrays, (iv) estimating false positives based on FDR method, (v) iteratively calculating FDR for every decoy score as threshold and (vi) calculating *q*-values and writing output to a file. For visual analysis, receiver-operating characteristic curves (ROC) and various types of scatter plots and histograms can be plotted from FDR files (see Supplementary Data). The output is formatted in an easy to manipulate CSV/text file for further processing using any scripting language making subsequent analysis hassle-free. For charts, Excel spreadsheets can be created. These files can be easily imported into R environment for further data visualization/post-processing.

The test datasets, sample input and expected outputs with ROC curves are provided along with a comparative table for various FDR methods across search engines (Supplementary Data). ProteoStats library has been extensively used in MassWiz, GenoSuite (Kumar *et al.*, 2013) and integrated multi-algorithmic analysis on plasma (Yadav *et al.*, 2011a). The FlexiFDR method (Yadav *et al.*, 2012) was also developed on top of the core framework of this library.

3 CONCLUSION

Proteostats is a highly versatile, platform independent, open-source, extensible and easy-to-use framework for FDR estimation and statistical control of results from shotgun proteomics database search. Apart from providing an assorted list of different FDR estimation procedures, it also provides single PSM metric-like *q*-values, and visualization features like ROC, histogram, Venn diagram and scatterplot for data quality assessment and comparative analysis.

ACKNOWLEDGEMENTS

The authors thank Dr. G.P. Singh and Dr. S. Ramachandran for their critical comments on the manuscript.

Funding: Council of Scientific and Industrial Research (CSIR, India) Network Project-GENESIS (BSC0121).

Conflict of Interest: none declared.

REFERENCES

Cerqueira,F.R. *et al.* (2010) MUDE: a new approach for optimizing sensitivity in the target-decoy search strategy for large-scale peptide/protein identification. *J. Proteome Res.*, **9**, 2265–2277.

Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

Elias,J.E. and Gygi,S.P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods*, **4**, 207–214.

Eng,J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.*, **5**, 976–989.

Eng,J.K. *et al.* (2013) Comet: an open-source MS/MS sequence database search tool. *Proteomics*, **13**, 22–24.

Fitzgibbon,M. *et al.* (2008) Modes of inference for evaluating the confidence of peptide identifications. *J. Proteome Res.*, **7**, 35–39.

Geer,L.Y. *et al.* (2004) Open mass spectrometry search algorithm. *J. Proteome Res.*, **3**, 958–964.

Kall,L. *et al.* (2008) Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *J. Proteome Res.*, **7**, 29–34.

Kumar,D. *et al.* (2013) Proteogenomic analysis of *Bradyrhizobium japonicum* USDA110 using Genosuite, an automated multi-algorithmic pipeline. *Mol. Cell Proteomics*, [Epub ahead of print, July 23, 2013].

Navarro,P. and Vazquez,J. (2009) A refined method to calculate false discovery rates for peptide identification using decoy databases. *J. Proteome Res.*, **8**, 1792–1796.

Perkins,D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, **20**, 3551–3567.

Tabb,D.L. *et al.* (2007) MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.*, **6**, 654–661.

Wedge,D.C. *et al.* (2011) FDRAnalysis: a tool for the integrated analysis of tandem mass spectrometry identification results from multiple search engines. *J. Proteome Res.*, **10**, 2088–2094.

Yadav,A.K. *et al.* (2011a) A systematic analysis of eluted fraction of plasma post immunoaffinity depletion: implications in biomarker discovery. *PLoS One*, **6**, e24442.

Yadav,A.K. *et al.* (2011b) MassWiz: a novel scoring algorithm with target-decoy based analysis pipeline for tandem mass spectrometry. *J. Proteome Res.*, **10**, 2154–2160.

Yadav,A.K. *et al.* (2012) Learning from decoys to improve the sensitivity and specificity of proteomics database search results. *PLoS One*, **7**, e50651.