**ProteoStats - a library for estimating false discovery rates in proteomics pipelines**

<table>
<thead>
<tr>
<th>Journal:</th>
<th><em>Bioinformatics</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>BIOINF-2013-0992.R1</td>
</tr>
<tr>
<td>Category:</td>
<td>Applications Note</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>n/a</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Yadav, Amit; CSIR-Institute of Genomics and Integrative Biology, Bioinformatics Kadimi, Puneet; CSIR-Institute of Genomics and Integrative Biology, Bioinformatics Kumar, Dhirendra; CSIR-Institute of Genomics and Integrative Biology, Bioinformatics Dash, Debasis; CSIR-Institute of Genomics and Integrative Biology, Bioinformatics</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Mass spectrometry, Proteomics, Bioinformatics, Data analysis, Protein sequence analysis, Visualization</td>
</tr>
</tbody>
</table>
**Gene Expression**

**ProteoStats - a library for estimating false discovery rates in proteomics pipelines**

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

CSIR-Institute of Genomics and Integrative Biology, South Campus, Mathura Road, Sukhdev Vihar, New Delhi-110020, India

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

**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 lead 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; aky.compbio@gmail.com; amit.yadav@igib.in

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

High throughout shotgun proteomics generates millions of spectra which 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.

Due 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, (iii) q-value calculation. Leucine/Isoleucine cannot be distinguished by collision induced dissociation (CID), 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 which 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 above, we have developed an open-source, cross-platform scripting library, ProteoStats, which employs 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 peptide spectral matches. ProteoStats supports OMSSA(Geer et al., 2004), MassWiz(Yadav et al., 2007).
TD searches can be conducted as:

- 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.

- 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).

Since 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 (PIT) 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, (vi) calculating q-values and writing output to a file. For visual analysis, ROC curves 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) while the FlexiFDR method (Yadav et al., 2012) was 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: This work was funded by CSIR Network Project - GENESIS (BSC0121).

REFERENCES


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


