BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 22 2013, pages 2946–2947
doi:10.1093/bioinformatics/btt513

Gene expression

Advance Access publication September 4, 2013

NetWeAvers: an *R* package for integrative biological network analysis with mass spectrometry data

Elizabeth A. McClellan[1,2,*], Perry D. Moerland[3,4,5], Peter J. van der Spek[1] and Andrew P. Stubbs[1]

[1]Department of Bioinformatics, Erasmus University Medical Center, Dr. Molewaterplein 50, 3015GE Rotterdam, The Netherlands, [2]Department of Mathematics and Computer Science, Metropolitan State University of Denver, Colorado, USA, [3]Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Bioinformatics Laboratory, Academic Medical Center, University of Amsterdam, Meibergdreef 9, Amsterdam, 1105AZ, The Netherlands, [4]Netherlands Proteomics Centre, H.R. Kruytgebouw, Padualaan 8, 3584CH Utrecht, The Netherlands and [5]Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525GA Nijmegen, The Netherlands

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Summary: The discovery of functionally related groups in a set of significantly abundant proteins from a mass spectrometry experiment is an important step in a proteomics analysis pipeline. Here we describe NetWeAvers (Network Weighted Averages) for analyzing groups of regulated proteins in a network context, e.g. as defined by clusters of protein–protein interactions. NetWeAvers is an *R* package that provides a novel method for analyzing proteomics data integrated with biological networks. The method includes an algorithm for finding dense clusters of proteins and a permutation algorithm to calculate cluster *P*-values. Optional steps include summarizing quantified peptide values to single protein values and testing for differential expression, such that the data input can simply be a list of identified and quantified peaks.

Availability and implementation: The NetWeAvers package is written in *R*, is open source and is freely available on CRAN and from netweavers.erasmusmc.nl under the GPL-v2 license.

Contact: e.mcclellan@erasmusmc.nl

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

Received on April 5, 2013; revised on July 31, 2013; accepted on August 28, 2013

1 INTRODUCTION

The statistical analysis of protein–protein interaction networks (PPINs) in conjunction with mass spectrometry (MS) data is an effective way to find functional groups of identified proteins in large networks. Several methods for network analysis are already implemented in *R*, but none are specific to label-free or labeled MS experiments. The package `ppiStats` provides tools for the analysis of PPINs, specifically for bait-prey technologies (Chiang *et al.*, 2013). `DEGraph` performs gene network differential expression (DE) testing on two conditions only (http://arxiv.org/abs/1009.5173). Few *R* packages are built specifically for MS data, and of those even fewer include downstream statistical analysis. None of them include the possibility to test on more than two conditions or perform network analysis. `MSnbase` and `MALDIquant` both process and quantify MS data without testing or network analysis (Gatto and Lilley, 2011; Gibb and Strimmer, 2012). The package `xcms` quantifies peaks and performs statistical analysis to find differences in two groups (*t*-tests) at the peak level (Smith *et al.*, 2006). The package `isobar` offers tools only for isobarically tagged MS proteomics data and includes a method for testing the difference in ratios between two groups (Breitwieser *et al.*, 2011).

`BioNet`, an *R* package that performs network analysis integratively with *P*-values from biological data, uses a maximal-scoring subgraph algorithm to find the optimal sub-network and, optionally, additional suboptimal solutions (Beisser *et al.*, 2010). In the algorithm, nodes are scored using a function of *P*-values, maximum likelihood estimates from a beta-uniform mixture model and a false discovery rate threshold. The inclusion of the false discovery rate threshold parameter influences the discovery of the optimal module by negatively scoring nodes considered not significant. Although multiple testing corrections and arbitrary significance cutoffs may be useful for detecting individual regulated genes or proteins, using such procedures in network analysis can possibly increase the false-negative rate. This is true especially when only one subnetwork, albeit 'optimal', is detected, or when regulated genes or proteins interact with unregulated ones that are crucial to the connectivity of the subnetwork. Considering this, we created an algorithm that finds and scores communities in a network without a subjective threshold and that does not require extra parameter specifications to find additional suboptimal subgraphs. Supplementary Table S1 presents a comparison of NetWeAvers and other network analysis tools; Supplementary Table S3 provides a rationale for removing *P*-value thresholds.

Here we present an *R* package that implements a network analysis method for finding dense clusters of DE proteins from MS data. It has three main components: peptide summarization, a test for DE and network analysis. The summarization and hypothesis testing steps allow for simple statistical analysis at the level of individual proteins: quantitative values for peptides are summarized to obtain protein quantities, and linear models are used to test for differences between groups to determine the

*To whom correspondence should be addressed.

statistical significance for each protein. The resulting *P*-values for individual proteins can be used in the network analysis step, which scores highly connected subgraphs, i.e. dense clusters, with these *P*-values. Because the need to specify many parameters can greatly impact the results, we chose a highly data-driven cluster-finding algorithm that requires only one parameter. Our protein and cluster scoring each require only one additional parameter.

2 DESCRIPTION

NetWeAvers provides a method for the integrated statistical analysis of MS data and biological networks. The input for NetWeAvers is a set of peaks from an MS experiment that has been identified, quantified and normalized. The data can be input as an *R*/Bioconductor ExpressionSet or a matrix to be converted into an ExpressionSet using `customSummarizer`. If the data are at the peptide level, then summarization to the protein level is required for use in NetWeAvers (`esetSummarizer`). This can be done before or after testing for DE (`DEtest`). The summarization step consists of aggregating all peptide quantities for a given protein using the mean or median so that each protein only has one value per sample.

The test for DE is implemented using the linear modeling framework of the `limma` package (Smyth, 2004). The output of the test includes *P*-values that may be used in the main algorithm of NetWeAvers (`runNetweavers`), which maps the proteins to a user-specified network in node–node format and performs the network analysis. The function `findDenseClusters` uses the Walktrap algorithm for finding highly connected subgraphs as implemented in the *R* package `igraph` (Csárdi and Nepusz, 2006) as a part of the network analysis algorithm.

The clusters are scored using a weighted mean or median of log-transformed *P*-values (`scoreClusters`). The weights are a function of the number of proteins with which a given protein interacts. A permutation test (`permTest`) may be carried out to determine the statistical significance of the clusters. See Supplementary File 1 for more details on the cluster scoring and the permutation test, as well as Supplementary Figure S1 for a schematic overview of the NetWeAvers procedure.

3 APPLICATION

We applied the *R* package to MS data from a phosphorylation study of human embryonic stem cells (Van Hoof *et al.*, 2009, see Supplementary File S1 for the experimental design). The *R* package vignette provided as Supplementary File S2 presents the code for summarizing the data, performing hypothesis testing and running the network analysis using the Reactome human PPIN, version 43 (Croft *et al.*, 2011). NetWeAvers identified clusters of proteins with roles in processes known to be involved in stem cell differentiation. See Supplementary File 1 for these results, results from NetWeAvers applied to a null dataset and an example using data that were summarized and tested in another *R* package.

4 CONCLUSIONS

NetWeAvers is a unique algorithm designed for quantitative MS data that incorporates key features of the proteins and networks (*P*-values and number of interactors, respectively) being analyzed. It uses only a few parameters and does not arbitrarily filter out non-significant proteins. We applied our method to a publicly available MS dataset and found statistically significant and biologically meaningful networks. The method may also be used with gene expression data. Many databases provide PPINs in node–node format, which makes it easy for users to connect NetWeAvers with their favorite databases. The format of the NetWeAvers output allows for simple connections to tools like Cytoscape (Shannon *et al.*, 2003) to visualize the resulting clusters.

ACKNOWLEDGEMENTS

The authors thank Javier Muñoz for discussions about the Van Hoof dataset and Steven V. Rødkær for suggesting changes to the algorithm.

Funding: This work was supported by The Netherlands Proteomics Centre, a program embedded in The Netherlands Genomics Initiative, and The Netherlands Bioinformatics Centre [NPC-GM WP3].

Conflict of Interest: none declared.

REFERENCES

Beisser,D. *et al.* (2010) BioNet: an *R*-package for the functional analysis of biological networks. *Bioinformatics*, **26**, 1129–1130.

Breitwieser,F. *et al.* (2011) General statistical modeling of data from protein relative expression isobaric tags. *J. Proteome Res.*, **10**, 2758–2766.

Chiang,T. *et al.* (2013) ppiStats: protein-protein interaction statistical package. *R* package version 1.25.0.

Croft,D. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691D697.

Csárdi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *Int. J. Complex Syst.*, 1695.

Gatto,L. and Lilley,K.S. (2011) MSnbase—an *R*/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, **28**, 288–289.

Gibb,S. and Strimmer,K. (2012) MALDIquant: a versatile *R* package for the analysis of mass spectrometry data. *Bioinformatics*, **28**, 2270–2271.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Smith,C.A. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching and identification. *Anal. Chem.*, **78**, 779–787.

Smyth,G.K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol.*, **3**, Article 3.

Van Hoof,D. *et al.* (2009) Phosphorylation dynamics during early differentiation of human embryonic stem cells. *Cell Stem Cell*, **5**, 214–226.