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

Elizabeth A. McClellan1; Perry D. Moerland2,3,4, Peter J. van der Spek1, and Andrew P. Stubbs1

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

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 materials are available at Bioinformatics online.

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 (FDR) threshold. The inclusion of the FDR threshold parameter influences the discovery of the optimal module by negatively scoring nodes considered not significant. While multiple testing corrections and arbitrary significance cutoffs may be useful for detecting individual regulated genes or proteins, utilizing such procedures in network analysis can possibly increase the false negative rate. This is true especially when only one sub-network, albeit “optimal”, is detected, or when regulated genes or proteins interact with unregulated ones that are crucial to the connectivity of the sub-network. 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; 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
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 PPIs 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.

ACKNOWLEDGEMENT

We would like to 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).

REFERENCES