Bioinformatics, 34(12), 2018, 2136–2138
doi: 10.1093/bioinformatics/bty063
Advance Access Publication Date: 7 February 2018
Applications Note

OXFORD

Gene expression

INfORM: Inference of NetwOrk Response Modules

Veer Singh Marwah[1,2], Pia Anneli Sofia Kinaret[1,2,3], Angela Serra[4], Giovanni Scala[1,2], Antti Lauerma[5], Vittorio Fortino[1,2] and Dario Greco[1,2,3,*]

[1]Faculty of Medicine and Life Sciences, University of Tampere, 33014 Tampere, Finland, [2]Institute of Biosciences and Medical Technologies (BioMediTech), 33520 Tampere, Finland, [3]Institute of Biotechnology, University of Helsinki, 00014 Helsinki, Finland, [4]NeuRoNeLab, DISA-MIS, University of Salerno, 84084 Fisciano, Italy and [5]Department of Dermatology and Allergology, University of Helsinki and Helsinki University Central Hospital, 00029 Helsinki, Finland

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on September 7, 2017; revised on January 8, 2018; editorial decision on February 4, 2018; accepted on February 6, 2018

Abstract

Summary: Detecting and interpreting responsive modules from gene expression data by using network-based approaches is a common but laborious task. It often requires the application of several computational methods implemented in different software packages, forcing biologists to compile complex analytical pipelines. Here we introduce INfORM (Inference of NetwOrk Response Modules), an R shiny application that enables non-expert users to detect, evaluate and select gene modules with high statistical and biological significance. INfORM is a comprehensive tool for the identification of biologically meaningful response modules from consensus gene networks inferred by using multiple algorithms. It is accessible through an intuitive graphical user interface allowing for a level of abstraction from the computational steps.

Availability and implementation: INfORM is freely available for academic use at https://github.com/Greco-Lab/INfORM.

Contact: dario.greco@staff.uta.fi

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

1 Introduction

Co-expression gene network analysis has become popular for analyzing DNA microarray and RNA sequencing data, for instance, to detect gene modules. A number of software solutions are available offering some important functionalities, but no tool to date implements a robust ensemble strategy for network inference (Supplementary Table S1). Here, we introduce INfORM (Inference of NetwOrk Response Modules), an R-Shiny (Chang *et al.*, 2017) application to assist performing all the computational steps needed to infer relevant responsive modules from transcriptomic data. Its graphical interface provides the user with a layer of abstraction and enables to analyze data without obligatory knowledge of gritty technical and statistical aspects. First, robust gene co-expression network is derived by combining multiple network inferences from an ensemble of methods (Marbach *et al.*, 2012). Second, gene modules are identified in the network by using community detection algorithms and evaluated based on multiple metrics of node importance. Third, the similarity between the identified modules is computed based on the overrepresented gene ontology (GO) terms in each module. Based on this information, the user can select a set of gene communities and merge them into a response module. The workflow of INfORM is depicted in Figure 1.

2 Methods and features

INfORM requires two mandatory input files: i) a gene expression data table; ii) a list of significantly differentially expressed genes. Gene symbols (Yates *et al.*, 2017) must be used as row names for

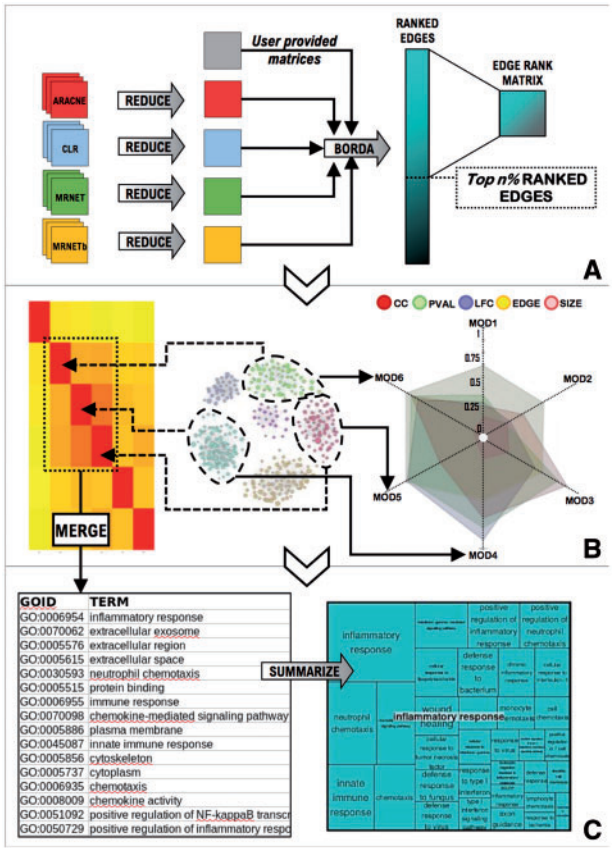

Fig. 1. (**A**) Matrices from same inference algorithm are summarized by median, mean or maximum of the correlation scores. Ranked edges from each summarized matrix are combined by the Borda method. Finally, the top ranked edges are selected until all the nodes are covered by at least one edge. (**B**) Modules are identified and annotated with node ranks, edge ranks. Modules can be selected and merged into a response module. (**C**) GO annotations of the response module are summarized and visualized by the means of tile plots

both the inputs. Optionally, the user can also upload directly one or more adjacency matrices inferred from any network inference algorithm. Then, the user is guided through the following steps: i) network inference; ii) module detection, iii) module evaluation and selection based on multiple gene-rank scores and overrepresented GO terms. The R functions for each step of the analysis are described in details in Supplementary Table S2. The user guide is also provided in Supplementary File S1.

2.1 Gene network inference

In INfORM, an ensemble of mutual information-based methods is used to infer gene co-expression networks. A compendium of networks is generated by using all possible distinct combinations of inference algorithms [ARACNE (Margolin *et al.*, 2006), MRNET (Meyer *et al.*, 2007), MRNETb (Meyer *et al.*, 2010) and CLR (Faith *et al.*, 2007)], mutual information and entropy estimators, and discretization methods available in the R package MINET (Meyer *et al.*, 2008), and integrated into the final network. These inferences can be integrated or even replaced by pre-computed adjacency matrices provided by the user. First, inferences by the same algorithm are reduced by taking the median, the mean or the maximum of the estimated edge weights. An evaluation of the effect of different summarization strategies is reported in Supplementary File S2. Second, the ranked connections from the summarized networks are aggregated by using the Borda method implemented in TopKLists R package (Schimek *et al.*, 2015) and the top ranked edges are progressively included in the final network until all the nodes are incorporated, i.e. all the nodes have degree ≥ 1 (Supplementary File S2). This procedure ensures the robustness of the network inference. Alternatively, the user can decide the top $n\%$ of the ranked edges to include in the final network.

2.2 Gene module detection and responsive module definition

INfORM provides widely used community clustering algorithms for identifying the relevant modules from the inferred gene network, the 'Walktrap' (Pons and Latapy, 2005), 'Spinglass' (Newman and Girvan, 2004; Reichardt and Bornholdt, 2006), 'Louvain' (Blondel *et al.*, 2008) and the 'Greedy' (Clauset *et al.*, 2004) algorithms. Benchmark analysis showed that all the algorithms generate highly similar modules (Supplementary File S2). The relevance of the identified modules is evaluated by scoring them based on characteristics of the member nodes and edges: i) centrality, ii) differential log2(fold-change), iii) differential *P*-value, iv) median rank of edge weights and v) number of nodes. These are graphically represented as a radar chart, making it easy to evaluate the modules. Moreover, enrichment analysis is performed to find the GO terms overrepresented in each module and compute the similarity between sets of GO terms from different modules. A heatmap representation of the GO-based module similarity is provided to aid the selection of functionally related modules. INfORM allows the user to merge statistically significant and biologically relevant modules into an optimized response module. The biological functions associated with the response module are visualized by the means of a tile plot (Supek *et al.*, 2011), in which the semantically similar GO terms are grouped together (Yu *et al.*, 2010).

3 Conclusion

Here we present INfORM, a novel tool for robust inference of gene co-expression networks from transcriptomics data. The graphical user interface helps to perform the analysis without any obligatory technical expertise and in-depth knowledge of the implementation. Results from a case study analysis of a publically available data are provided in Supplementary File S3.

Acknowledgements

We are grateful to Dr. Lasse Ruokolainen (University of Helsinki) for his comments on the methods. We thank MSc. Juho Väänänen and Dr. Iiris Hovatta (University of Helsinki) for benchmarking the INfORM software and providing valuable feedback.

Funding

This study was supported by the Academy of Finland (grant agreements 275151 and 292307), EU H2020 caLIBRAte project (grant agreement 686239), EU H2020 LIFEPATH (grant agreement 633666) and EU FP7 NANOSOLUTIONS project (grant agreement FP7-309329).

Conflict of Interest: none declared.

References

Blondel,V.D. *et al.* (2008) Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.*, **2008**, P10008.

Chang,W. *et al*. (2017) shiny: Web Application Framework for R. *R package version 1.0.5*.

Clauset,A. *et al*. (2004) Finding community structure in very large networks. *Phys. Rev. E*, **70**, 066111–066117.

Faith,J.J. *et al*. (2007) Large-scale mapping and validation of *Escherichia coli* transcriptional regulation from a compendium of expression profiles. *PLoS Biol*., **5**, e8.

Marbach,D. *et al*. (2012) Wisdom of crowds for robust gene network inference. *Nat. Methods*, **9**, 796–804.

Margolin,A.A. *et al*. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.

Meyer,P.E. *et al*. (2007) Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinform. Syst. Biol*., **2007**, 1.

Meyer,P.E. *et al.* (2008) minet: a R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, **9**, 461.

Meyer,P.E. *et al*. (2010) Information-theoretic inference of gene networks using backward elimination. In: *BIOCOMP International Conference Bioinformatics Computational Biology CSREA Press*, 2010, pp. 700–705.

Newman,M.E.J. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phy. Rev. E*, **69**.

Pons,P. and Latapy,M. (2005) Computing communities in large networks using random walks. In: *Computer and Information Sciences - ISCIS 2005, Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, pp. 284–293.

Reichardt,J. and Bornholdt,S. (2006) Statistical mechanics of community detection. *Phys. Rev. E*, **74**.

Schimek,M.G. *et al*. (2015) TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Stat. Appl. Genet. Mol. Biol*., **14**, 311–316.

Supek,F. *et al*. (2011) REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One*, **6**, e21800.

Yates,B. *et al*. (2017) Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res*., **45**, D619–D625.

Yu,G. *et al*. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.