BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 8 2010, pages 1129–1130
doi:10.1093/bioinformatics/btq089

Systems biology Advance Access publication February 25, 2010

BioNet: an R-Package for the functional analysis of biological networks

Daniela Beisser[1], Gunnar W. Klau[2], Thomas Dandekar[1], Tobias Müller[1,*] and Marcus T. Dittrich[1,*]

[1]Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany and [2]Life Sciences group, CWI, Science Park 123, 1098 XG Amsterdam, The Netherlands

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Increasing quantity and quality of data in transcriptomics and interactomics create the need for integrative approaches to network analysis. Here, we present a comprehensive R-package for the analysis of biological networks including an exact and a heuristic approach to identify functional modules.

Results: The BioNet package provides an extensive framework for integrated network analysis in R. This includes the statistics for the integration of transcriptomic and functional data with biological networks, the scoring of nodes as well as methods for network search and visualization.

Availability: The BioNet package and a tutorial are available from http://bionet.bioapps.biozentrum.uni-wuerzburg.de

Contact: marcus.dittrich@biozentrum.uni-wuerzburg.de; tobias.mueller@biozentrum.uni-wuerzburg.de

Received on December 14, 2009; revised on February 5, 2010; accepted on February 22, 2010

1 INTRODUCTION

Integrated analysis of microarray data in the context of biological networks such as protein–protein interaction (PPI) networks has become a major technique in systems biology. The primary objective is the identification of functional modules (significantly differentially expressed subnetworks) within large networks. This can be achieved by computing a score for each node in the network reflecting its functional relevance. Subsequently, a network search algorithm is required to find the highest scoring subgraph. In fact, this problem has been proven to be NP-hard (Ideker *et al.*, 2002). Various heuristic approaches have been proposed, most of them inspired by the seminal work from Ideker *et al.* (2002) that used a simulated annealing heuristic to identify high-scoring subgraphs in integrated networks. Recently, we have devised an algorithm (*heinz*, heaviest induced subgraph) that computes provably optimal and suboptimal solutions to the maximal-scoring subgraph (MSS) problem in reasonable running time using integer linear programming (ILP) (Dittrich *et al.*, 2008). In extension to this, we present an R package for (i) integrating multiple *P*-values obtained from different experiments, (ii) scoring the nodes of the network by a modular scoring function, (iii) calculating provably optimal and suboptimal solutions to the MSS problem, (iv) calculating high-scoring solutions with a novel heuristic, and (v) 2D and 3D visualization of network solutions, see also Figure 1.

2 DESCRIPTION

The BioNet package provides a comprehensive set of methods for the integrated analysis of gene expression data and biological networks. *P*-values are distributed uniformly under null hypotheses. Therefore, as a first step, multiple *P*-values derived from the analysis of different experiments (e.g. *t*-test or regression models) can be aggregated using a uniform order statistics (`aggrPvals`) (Dittrich *et al.*, 2008). The resulting distribution of combined *P*-values can be considered as a mixture of signal and noise, where the signal component is modelled to be Beta(a,1) distributed (Pounds and Morris, 2003). The model fit can be verified by the provided diagnostic plots (`plot.bum, hist.bum`). By fitting a beta-uniform mixture (BUM) model (`fitBumModel`), the maximum-likelihood estimates for the mixture model can be obtained. These parameters are subsequently used to score the nodes of the network (`scoreNodes, scoreFunction`). The adjusted node score is given by $(a-1)\big(\log(x)-\log(\tau(\text{FDR}))\big)$, where τ denotes the threshold for a given false discovery rate (FDR). The optimal and heuristic solutions of the MSS can be calculated by `runHeinz, runFastHeinz`. Bioconductor data structures and classes (Gentleman *et al.*, 2004) of the graph packages graph, RBGL as well as igraph are supported (Carey *et al.*, 2005; Csardi and Nepusz, 2006). Networks can be imported and exported in different formats, allowing a smooth data exchange with standard network analysis tools like Cytoscape (Shannon *et al.*, 2003).

3 APPLICATION

We apply our package to gene expression data from diffuse large B-cell lymphomas (DLBCL) and survival data (Rosenwald *et al.*, 2002) with a human PPI network based on human protein reference database (HPRD; Prasad *et al.*, 2009) as described in Dittrich *et al.* (2008). The data consist of 112 tumors with the germinal center B-like phenotype (GCB) and 82 tumors with the activated B-like phenotype (ABC) and includes information on patient survival. All data are available in the BioNet and DLBCL package. We use standard microarray analysis and Cox regression to obtain gene-wise *P*-values for differential expression and risk association,

*To whom correspondence should be addressed.

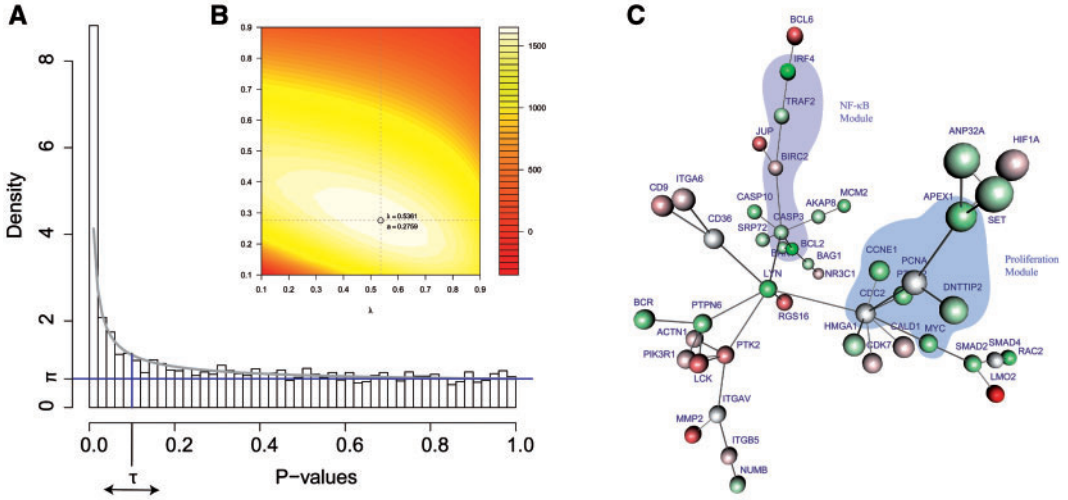

Fig. 1. Node scoring and network solution for the DLBCL dataset. (**A**) Fitted mixture model and empirical *P*-value distribution: π indicates the upper bound for the fraction of noise and τ the significance threshold according to a given FDR. (**B**) Log-likelihood surface for the mixture parameter λ (*x*-axis) and the shape parameter *a* (*y*-axis). The derived scores from the *P*-value distribution are used subsequently to calculate the MSS. (**C**) A 3D visualization of the identified optimal scoring module. Differential expression is depicted by node coloring (red: upregulated in ABC, green: upregulated in GCB). Disease-relevant modules (shaded) (Rosenwald *et al.*, 2002) are captured and extended by the network analysis.

respectively. Then we aggregate both *P*-values by the second-order statistics using BioNet.

```
> data(dataLym)
> pvals <- cbind(t=dataLym$t.pval, s=dataLym$s.pval)
> pval  <- aggrPvals(pvals, order=2, plot=FALSE)
```

We now fit a BUM model to the distribution of aggregated *P*-values and score the nodes using an FDR threshold of 0.001.

```
> fb <- fitBumModel(pval, plot=FALSE)
> scores <- scoreNodes(network, fb=fb, fdr=0.001)
> writeHeinz(network, file="lym_001",
+ node.scores=scores)
```

The exact search algorithm can be started from R by `runHeinz` if the CPLEX library is installed (Dittrich *et al.*, 2008). Alternatively, the fast heuristic search algorithm (`runFastHeinz`) often delivers a close approximation. Finally, the resulting modules can be visualized in 2D or 3D.

```
> module <- readHeinzGraph(node.file=
+ "lym_001_n.txt.0.hnz", network)
> plot3dModule(module)
```

BioNet captures an interaction module that has been described to play major biological roles in the GCB and ABC DLBCL subtypes (Fig. 1C). The combination of biological and clinical data with PPI networks generates a meaningful biological context in terms of functional association for differentially expressed, survival-relevant genes.

Funding: BMBF (Funcrypta); DFG (Da208/10-1); DFG (TR34/A5).

Conflict of Interest: none declared.

REFERENCES

Carey,V.J. *et al.* (2005) Network structures and algorithms in Bioconductor. *Bioinformatics*, **21**, 135–136.

Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal*, Complex Systems, http://igraph.sf.net, 1695.

Dittrich,M.T. *et al.* (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics*, **24**, i223–i231.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (Suppl. 1), S233–S240.

Pounds,S. and Morris,S.W. (2003) Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values. *Bioinformatics*, **19**, 1236–1242.

Prasad,T.S.K. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Rosenwald,A. *et al.* (2002) The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. *N. Engl. J. Med.*, **346**, 1937–1947.

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