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

Daniela Beisser¹, Gunnar W. Klau², Thomas Dandekar¹, Tobias Müller¹,∗
and Marcus T. Dittrich¹,∗

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

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, as well as (i) 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, 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 

\[ \frac{\alpha - 1}{\alpha}(\log(x) - \log(\tau) \cdot \text{FDR}) \]

where \( \tau \) 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 (GC) 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.
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 α (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.

```r
> 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.

```r
> 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.

```r
> 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


