Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

An R package to compute diffusion-based scores on biological networks: diffuStats

Sergio Picart-Armada [1,2]*, Wesley K. Thompson[3,4], Alfonso Buil [3] and Alexandre Perera-Lluna [1,2]

[1] B2SLab, Departament d'Enginyeria de Sistemes, Automàtica i Informàtica Industrial, Universitat Politècnica de Catalunya, CIBER-BBN, Barcelona, 08028, Spain, [2] Institut de Recerca Pediàtrica Hospital Sant Joan de Déu, Esplugues de Llobregat, Barcelona, 08950, Spain, [3] Mental Health Center Sct. Hans, 4000 Roskilde, Denmark and [4] Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, California, USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: Label propagation and diffusion over biological networks are a common mathematical formalism in computational biology for giving context to molecular entities and prioritising novel candidates in the area of study. There are several choices in conceiving the diffusion process -involving the graph kernel, the score definitions and the presence of a posterior statistical normalisation- which have an impact on the results. This manuscript describes diffuStats, an R package that provides a collection of graph kernels and diffusion scores, as well as a parallel permutation analysis for the normalised scores, that eases the computation of the scores and their benchmarking for an optimal choice.
Availability and Implementation: The R package diffuStats is publicly available in Bioconductor, `https://bioconductor.org`, under the GPL-3 license
Contact: sergi.picart@upc.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Network analysis can help finding therapeutic targets and understanding biology in networks obtained from protein-protein interactions, gene regulation and metabolic reactions. In this context, label propagation and diffusion algorithms (Zoidi, 2015) address a general problem of molecular entity ranking according to a seed node list. Examples include finding significantly mutated subnetworks in cancer (Vandin, 2010), predicting gene function (Mostafavi, 2008), prioritising genome-wide association hits (Lee, 2011) and classifying proteins (Tsuda, 2005).

In general, the mentioned methods involve diffusion processes with ad-hoc parameter and network settings, making comparisons fundamentally difficult. The RANKS R package (Valentini, 2016) is an effort to collect a range of diffusion kernels and scores, but the effects of label codification and a recently proposed statistical normalisation (Bersanelli, 2016) have not been explored. To that end, we introduce the diffuStats R package gathering diffusion kernels, input codifications and statistical normalisations to benchmark single-network diffusion settings.

2 Methods

The diffuStats R package offers scoring schemes for diffusing a label vector on a network, determined by (a) the graph kernel, (b) the translation of labels into a numeric vector y to be smoothed, and (c) the statistical normalisation. In general, diffusion scores f are based on modifications of the quantity $f = K \cdot y$, where y are the input labels, f the diffusion scores and K is a graph kernel.

Regarding (a), most of the cited applications use the regularised Laplacian kernel, but our package also offers the diffusion kernel, the p-step random walk kernel, the inverse cosine kernel (Smola and Kondor, 2003) and the commute time kernel (Yen, 2007). In practice, they differ on the reach and the behaviour of the spreading inside the network - further detail for its choice can be found in the documentation. The decision can be data-driven, based on prior studies on the subject or on desirable properties of the kernel. For (b) and (c), the implemented scores are variations of the propagation of a binary vector whose ones are the positive labels y^+ and whose zeroes are the negative y^- and unlabelled y^u entities (Table 1). The statistical normalisation (c) compares the diffusion scores with the distribution of scores that arise from a permuted input, in order to spot

Table 1. Implemented diffusion scores. f_{raw}, f_{ml} and f_{gm} differ on the weights of the positive, negative and unlabelled nodes. f_{ber_s} quantifies the change of the f_{raw} scores relative to the input scores. f_{ber_p}, f_{mc} and f_z are statistically normalised by permuting the labelled examples, but not the unlabelled*. f_{mc} derives from an empirical p-value, whereas f_{ber_p} combines f_{mc} and f_{raw}. f_z is a parametric alternative to f_{mc} requiring no stochastic permutations. Quantitative inputs are allowed in all the scores except f_{ml} and f_{gm}.

Score	y^+	y^-	y^u	Normalised	Stochastic	Quantitative	Reference
raw	1	0	0	No	No	Yes	Vandin, 2010
ml	1	-1	0	No	No	No	Tsuda, 2005
gm	1	-1	k	No	No	No	Mostafavi, 2008
ber$_s$	1	0	0	No	No	Yes	Bersanelli, 2016
ber$_p$	1	0	0*	Yes	Yes	Yes	Bersanelli, 2016
mc	1	0	0*	Yes	Yes	Yes	Bersanelli, 2016
z	1	0	0*	Yes	No	Yes	Harchaoui, 2013

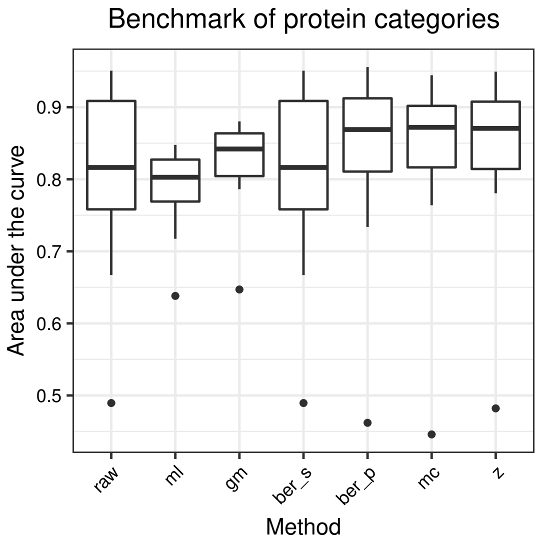

Fig. 1. Performance comparison for diffusion scores in predicting 12 functions on half of the proteins using the area under the Receiver Operating Characteristic curve.

nodes whose score is systematically high or low regardless of the input. However, not all normalised scores require stochastic simulations.

The diffuStats R package contains proper documentation and unit testing to facilitate its development. Its algorithms are written in R except the permutations, which use C++; further details on the implementation can be found in the supplementary materials. Manipulating networks with more than 10,000 nodes might require extra RAM memory and computational power due to the kernel matrix size.

3 Results

The example data is a yeast interactome with 12 annotated protein functions (Von Mering, 2002). The functionalities of our package are demonstrated by (i) obtaining a prioritised list of annotations given a set of labels, and (ii) benchmarking all the available diffusion scores in a dataset containing validation data. For both analyses, half of the proteins in the interactome will be treated as unlabelled and will receive a score from the propagation of the other half using the default regularised Laplacian kernel. Regarding (i), the function “diffuse” allows to compute the desired diffusion scores with a starting set of scores (labels) and a network:

```
scores_diff <- diffuse(graph = yeast,
  scores = scores, method = "raw")
```

When assessing the performance of different diffusion scores in a given dataset (ii), the desired metrics involving the diffusion scores and the validation can be computed on a grid of parameters:

```
performance <- perf(graph = yeast, scores = scores,
  validation = validation, grid_param = grid_param)
```

The results are returned as a table that can be directly plotted (Fig. 1). In this case study, statistically normalised scores f_{mc}, f_z and f_{ber_p} seem preferable than their unnormalised counterparts comparing the area under the ROC curves. For instance, f_z outperforms f_{raw}, f_{ml}, f_{gm} and f_{ber_s} (FDR < 25%, Wilcoxon test), thus highlighting the usefulness of a prior screening in score performance and its potential impact.

In summary, the R package diffuStats gathers diffusion kernels and scores with statistical normalisation that are object of active research in bioinformatics, like functional prediction or module identification. In addition, it facilitates benchmarking diffusion scoring methods to find the optimal configuration for the application of interest.

Funding:

This work was supported by the Spanish Ministry of Economy and Competitiveness (MINECO) [TEC2014-60337-R to A.P.] and the National Institutes of Health (NIH) [R01GM104400 to W.T.]. AP. and S.P. thank CIBERDEM and CIBER-BBN for funding, both initiatives of the Spanish ISCIII. SP. thanks the AGAUR FI-scholarship programme.

Conflict of Interest: none declared

References

Bersanelli, M. *et al.* (2016). Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Sci. Rep.*, **6**.

Harchaoui, Z. *et al.* (2013). Kernel-based methods for hypothesis testing: A unified view. *IEEE Signal Process Mag*, **30**(4), 87–97.

Lee, I. *et al.* (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res.*, **21**(7), 1109–1121.

Mostafavi, S. *et al.* (2008). Genemania: a real-time multiple association network integration algorithm for predicting gene function. *Genome Biol.*, **9**(1), S4.

Smola, A. J. and Kondor, R. (2003). Kernels and Regularization on Graphs. *Mach. Learn.*, **2777**, 1–15.

Tsuda, K. *et al.* (2005). Fast protein classification with multiple networks. *Bioinformatics*, **21**(SUPPL. 2), 59–65.

Valentini, G. *et al.* (2016). RANKS: A flexible tool for node label ranking and classification in biological networks. *Bioinformatics*, **32**(18), 2872–2874.

Vandin, F. *et al.* (2010). Algorithms for detecting significantly mutated pathways in cancer. *Lect. Notes Comput. Sci.*, **6044 LNBI**(3), 506–521.

Von Mering, C. *et al.* (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**(6887), 399–403.

Yen, L. *et al.* (2007). Graph nodes clustering based on the commute-time kernel. *Advances in Knowledge Discovery and Data Mining*, pages 1037–1045.

Zoidi, O. *et al.* (2015). Graph-Based Label Propagation in Digital Media: A Review. *ACM Comput. Surv.*, **47**(3), 48.