Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Systems Biology

OXFORD

Systems Biology

PAFway: pairwise associations between functional annotations in biological networks and pathways

Mahiar Mahjoub [1, 2, 3] and Daphne Ezer [2,4,5]*

[1] Department of Mathematics, University of Cambridge, Cambridge, UK,
[2] The Alan Turing Institute, London, UK,
[3] Royal Prince Alfred Hospital, Central Clinical School, University of Sydney, Sydney, Australia,
[4] Department of Statistics, University of Warwick, Coventry, UK and
[5] Department of Biology, University of York, York, UK.

*To whom correspondence should be addressed: daphne.ezer@york.ac.uk.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Large gene networks can be dense and difficult to interpret in a biologically meaningful way.
Results: Here, we introduce PAFway, which estimates pairwise associations between functional annotations in biological networks and pathways. It answers the biological question: do genes that have a specific function tend to regulate genes that have a different specific function? The results can be visualised as a heatmap or a network of biological functions. We apply this package to reveal associations between functional annotations in an *Arabidopsis thaliana* gene network.
Availability: PAFway is available in CRAN.
Contact: daphne.ezer@york.ac.uk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biological networks can be very large, dense and difficult to visualise and interpret. PAFway is a tool to interpret large, dense biological networks in the context of functional annotations, such as gene ontology (GO). Some methods that analyse GO enrichment within networks, such as BiNGO [Maere *et al.*, 2005], operate by partitioning the network into clusters and then finding functional enrichment within each cluster.

Another family of methods, called topological network enrichment methods, utilise the structure of the network to find gene ontology terms that are enriched in a network or subnetwork [Mitrea *et al.*, 2013]. The output of these algorithms is generally a ranked list of annotations, ordered by how much they are enriched in the network.

In contrast, PAFway finds *pairwise associations of functional annotations* in biological networks and pathways, which is calculated efficiently using the Fast Fourier Transform (FFT). The results can be illustrated either in the form of a heat map or as a network where the nodes in the graph are functional annotations. We apply this method to AraNet [Lee *et al.*, 2015], a gene network for *Arabidopsis thaliana*.

2 Methods

The PAFway function takes as input a directed network, with or without edge weights, and a list all the functional annotations associated with each node. We refer to each *edge type* as an ordered pair of functional annotations, representing the scenario where a gene with the first functional annotation regulates a gene with the second functional annotation. The output of PAFway is the probability of observing at least the observed number (or sum of edge weights) of each edge type, under a null model in which the functional annotations are randomly distributed in the network (after correcting for multiple hypothesis testing).

2.1 P-value of edge counts

Let us say that the frequency of the first functional annotation in the network is p_a and the second is p_b. The probability of observing an edge between p_a and p_b is $p_{a,b} = p_a p_b$ if they are randomly distributed in the network. The probability of observing n edges between the first and second functional annotation in a network with N edges is determined by a binomial distribution:

$$n \sim B(N, p_{a,b}) \tag{1}$$

This means that it is possible to determine the probability of observing at least n edges of a certain type by using the binomial test.

2.2 P-value of sum of edge weights

When a gene network contains edge weights, we calculate the sum of the edge weights of each edge type, and we would like to know whether this value is higher than would be expected by chance. For two functional annotations a and b, let us define $z_{a,b}$ as the sum of the edge weights of edge type (a, b) in the network. Let us say that $c_{a,b}$ is the count of the number of edges of that type. $P(c_{a,b} = i)$ is the probability of observing exactly i edges of type (a, b) and $P(x \geq z_{a,b}|c_{a,b} = i)$ is the probability of observing a sum of edge weights greater than $z_{a.b}$ given that $c_{a,b} = i$. The probability of observing at least $z_{a,b}$ is:

$$P(x \geq z_{a,b}) = \sum_{i=1}^{N} P(c_{a,b} = i)P(x \geq z_{a,b}|c_{a,b} = i) \qquad (2)$$

where N is the number of edges in the network. Note that $P(x \geq z_{a,b})$ is the p-value. From the previous section, we see that $P(c_{a,b} = i)$ is the probability density function (pdf) of the binomial distribution $B(N, p_{a,b})$. $P(x \geq z_{a,b}|c_{a,b} = i)$ can be determined by a set of recursive functions described in SI 1.1. These functions are convolutions and so can be expressed in terms of Fourier transforms and calculated efficiently using the Fast Fourier Transform (FFT) (see SI 1.2).

3 Results

PAFway produces a network of functional annotations, which can be depicted either as a network (Fig. 1A-B) or a heatmap (Fig. 1C), This is shown for AraNet, a gene regulatory network for *Arabidopsis thaliana* (SI2.1). We are not aware of any other accessible tool for performing this precise task, but there are a number of alternative packages that perform other kinds of complementary analyses of GO terms.

First, we compare the results of PAFway to a pairwise association score similar to the one proposed by Chitale *et al.* [2011], Yerneni *et al.* [2018]. We find that our method produces results that are consistent with this score, but with the added benefit of providing a p-value (SI 2.2).

Next, we compare our results to those produced by NaviGO [Wei *et al.*, 2017], a tool that allows the user to calculate the similarity between pairs of GO terms, based on either semantic similarity [Resnik, 1999, Schlicker *et al.*, 2006, Lin, 1998] or how often they appear together in gene annotations [Chitale *et al.*, 2011], the scientific literature [Chitale *et al.*, 2011], and in physically interacting proteins [Yerneni *et al.*, 2018]. We find that the strength of the correlation between our p-values and these metrics varies quite substantially based on whether edge weight information is incorporated in the model (SI 2.3).

Finally, we cluster the AraNet network into communities, and visualise the GO terms within each community with both BiNGO [Maere *et al.*, 2005] and PAFway. We suggest that BiNGO can be used to help identify GO terms of interest whose relationships within the network could be further analysed with PAFway (SI 2.4).

In conclusion, PAFway provides information that is complementary to these alternative methods, providing an innovative way to improve our understanding of large biological networks.

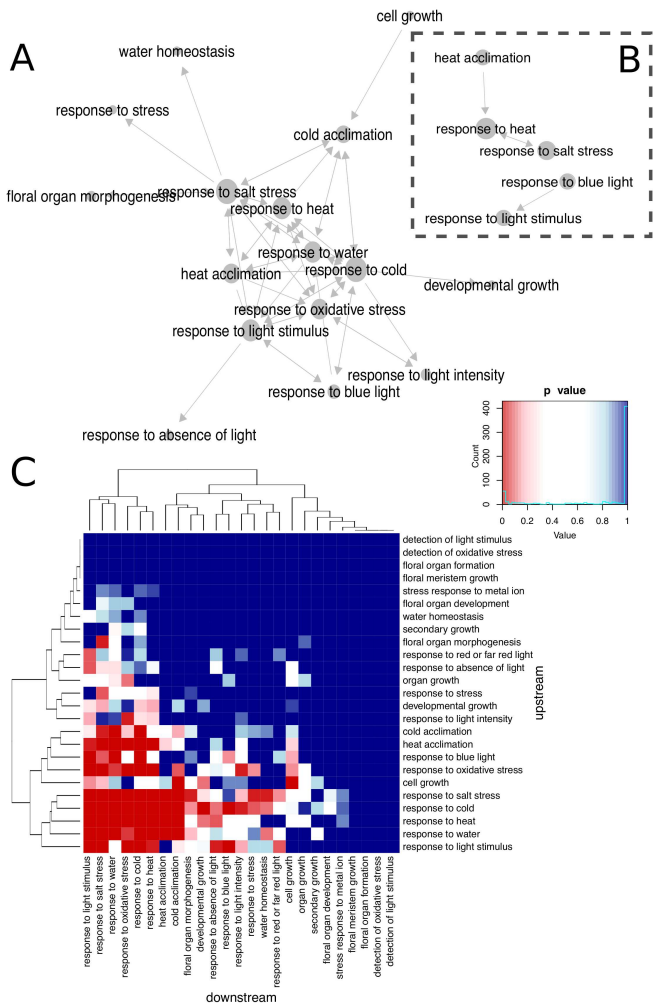

Fig. 1. These are the results of PAFway applied to the AraNet gene regulatory network of Arabidopsis thaliana. (A) This is the network, ignoring edge weights. (B) This is the network, including edge weights. (C) This is the network in (A) represented as a heatmap.

Funding

Turing Research Fellowship under EPSRC grant (TU/A/000017); EPSRC/BBSRC Innovation Fellowship (EP/S001360/1); UKRI Research Strategic Priority Fund (R-SPES-107).

References

Chitale, M. *et al.* (2011). Quantification of protein group coherence and pathway assignment using functional association. *BMC Bioinformatics*.

Lee, T. *et al.* (2015). AraNet v2: An improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. *Nucleic Acids Research*.

Lin, D. (1998). An Information-Theoretic Definition of Similarity. *ICML*.

Maere, S. *et al.* (2005). BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*.

Mitrea, C. *et al.* (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology*.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*.

Schlicker, A. *et al.* (2006). A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*.

Wei, Q. *et al.* (2017). NaviGO: Interactive tool for visualization and functional similarity and coherence analysis with gene ontology. *BMC Bioinformatics*.

Yerneni, S. *et al.* (2018). IAS: Interaction Specific GO Term Associations for Predicting Protein-Protein Interaction Networks. *IEEE/ACM transactions on computational biology and bioinformatics*.