BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 23 2014, pages 3387–3389
doi:10.1093/bioinformatics/btu536

Genome analysis Advance Access publication August 13, 2014

NetComm: a network analysis tool based on communicability

Ian M. Campbell[†], Regis A. James[†], Edward S. Chen and Chad A. Shaw[*]

Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77054, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Set-based network similarity metrics are increasingly used to productively analyze genome-wide data. Conventional approaches, such as mean shortest path and clique-based metrics, have been useful but are not well suited to all applications. Computational scientists in other disciplines have developed communicability as a complementary metric. Network communicability considers all paths of all lengths between two network members. Given the success of previous network analyses of protein–protein interactions, we applied the concepts of network communicability to this problem. Here we show that our communicability implementation has advantages over traditional approaches. Overall, analyses suggest network communicability has considerable utility in analysis of large-scale biological networks.

Availability and implementation: We provide our method as an R package for use in both human protein–protein interaction network analyses and analyses of arbitrary networks along with a tutorial at http://www.shawlab.org/NetComm/.

Contact: cashaw@bcm.edu

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

Received on June 10, 2014; revised on July 11, 2014; accepted on August 2, 2014

1 INTRODUCTION

Many types of biological data, for example protein–protein interactions, coexpression data and biomedical research literature naturally lend themselves to formation of interconnected networks. While finding meaningful connections between network members (i.e. nodes, proteins or genes) can be challenging, networks provide an opportunity to contextualize results at the system level. Standard analyses rely on relationships between individual pairs of nodes and can be vulnerable to stochasticity and technical errors. Identifying sets of relationships between individual nodes and groups of others is one strategy for a more robust metric of connectedness. The challenge remains to retain topological information while simplifying network interactions into univariate representations for statistical analysis and comparison with other data types.

A number of metrics have been successfully applied to the analysis of biological networks previously (Brohée *et al.*, 2008). Shortest path analysis assesses the fewest connections (i.e. edges) required to connect a query node to one or more targets. However, nodes which themselves have many connections (i.e. high degree) are on average closer to all nodes. Therefore, identifying meaningful relationships by shortest path often requires some form of degree correction. Another weakness of shortest path analysis is that two or more paths of the same length are equivalent to a single path of that length. Thus, information regarding connectedness is lost. Jaccard similarity between two nodes is the quotient of the intersection and the union of their k-th degree interactors (Jaccard, 1912). This metric is useful for testing connections to 'near by' nodes, but performs poorly for nodes that are far apart. The method of random walk mean first passage assesses how soon a random walk leaving a node takes to reach a given target (Noh and Rieger, 2003). This metric may also require degree correction.

Researchers in physics have proposed communicability as a metric of connectedness (Estrada and Hatano, 2008). We previously used communicability to identify proteins connected in the human PPI network to a set of epilepsy proteins to aid identification of novel epilepsy genes (Campbell *et al.*, 2013). Communicability considers paths of all lengths between two nodes scaled by the factorial of path length (Estrada and Hatano, 2008), retaining more information about connectedness compared with shortest path metrics. It remains useful across greater path lengths compared with Jaccard similarity and avoids algorithmic stochasticity. Communicability is still influenced by degree and is challenged by large networks such as the human protein–protein interaction network. Here, we propose a degree-normalized network connectedness metric inspired by network communicability and suitable for analysis of large complex networks. We provide an implementation for the human PPI network, as well as tools for arbitrary networks.

2 METHODS

Finite network communicability

Communicability $C_{i,j}$ between two nodes i and j counts paths of all lengths between them. Contributions are scaled so that longer paths contribute less than shorter ones (Estrada and Hatano, 2008). Using the adjacency matrix $\mathbf{A}$, where the i,j entry is 1 if nodes are connected and 0 otherwise, communicability is defined:

$$C_{i,j} = \sum_{k=1}^{\infty} \frac{(\mathrm{A}^k)_{i,j}}{k!} = (e^{\mathbf{A}})_{i,j} \qquad F_{i,j} = \sum_{k=1}^{l} \frac{(\mathrm{A}^k)_{i,j}}{k!} \tag{1}$$

Evaluating $e^{\mathbf{A}}$ for large networks is possible (Estrada and Hatano, 2008). However, when we investigated use of conventional network communicability in genome-wide networks, we found size and interconnectedness resulted in large values and loss of variability between nodes. Thus, we investigated a metric we deem finite network communicability $F_{i,j}$ considering finite path lengths from 1 to l (Equation 1). Preliminary analysis revealed our metric was vulnerable to node degree. Thus, we

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

normalized by subtracting the mean and dividing by the standard deviation in that row:

$$F_{i,j}=\sum_{k=1}^{l}\frac{\frac{(A^k)_{i,j}-\overline{(A^k)_i}}{\sigma_{i,k}}}{k!}\quad F_{i,\mathbf{T}}=\sum_{k=1}^{l}\frac{\frac{\overline{(A^k)_{i,\mathbf{T}}}-\overline{(A^k)_i}}{\sigma_{i,k}}}{k!}\tag{2}$$

where $\sigma_{i,k}$ is the standard deviation of the i-th row of the k-th power of the adjacency matrix. To compute the metric $F_{i,\boldsymbol{T}}$ to a set of target nodes $\boldsymbol{T}$, we computed the mean across the set of nodes for each power (Equation 2). Other metrics of center and scale such as median and interquartile range can also be used.

Human disease network and genome-wide association studies (GWAS)catalog

Phenotypic classes for Online Mendelian Inheritance in Man (OMIM) disease genes were extracted computationally from the Supplementary Tables from Goh *et al.* (2007). Genes assigned to multiple classes because different mutations of the same gene cause different phenotypes (allelic heterogeneity) were allowed; however, each gene could contribute to the same class only once. For shortest path analysis, we computed the mean shortest path for each node to each disease class and centered and scaled in a manner analogous to network communicability.

Gene–disease associations from GWAS were downloaded from the National Human Genome Research Institute Catalogue of Published GWAS (Welter *et al.*, 2014). Diseases ‘type 1 diabetes’, ‘type 2 diabetes’, ‘breast cancer’, ‘age-related macular degeneration’ and ‘coronary heart disease’ were selected because of a large quantity of associations and diverse etiologies. Associations met a significance of 5×10^{-8} and corresponded to a single gene.

3 RESULTS

We assessed connectedness of 3724 proteins causative of Mendelian diseases that have been previously classified to 18 phenotypic classes (Goh *et al.*, 2007). We used the InWeb (Lage *et al.*, 2007) network to calculate mean communicabilities of proteins causing neurological and cardiovascular disease to those of other classes. Our metric identified increased connectedness among proteins within classes than between classes and outperformed shortest path (Fig. 1A, Supplementary Fig. S1). Our method similarly identified interactions among gene products associated with type 1 diabetes (Fig. 1B).

4 CONCLUSIONS

Network communicability provides advantages over alternative metrics because it retains topology information, lends itself to set-based analysis and is easy to represent with univariate scores. It outperforms shortest path in a variety of situations. Overall, our metric may prove useful in the analysis of a variety of biological networks, and our package provides a straightforward approach to computation even on large networks.

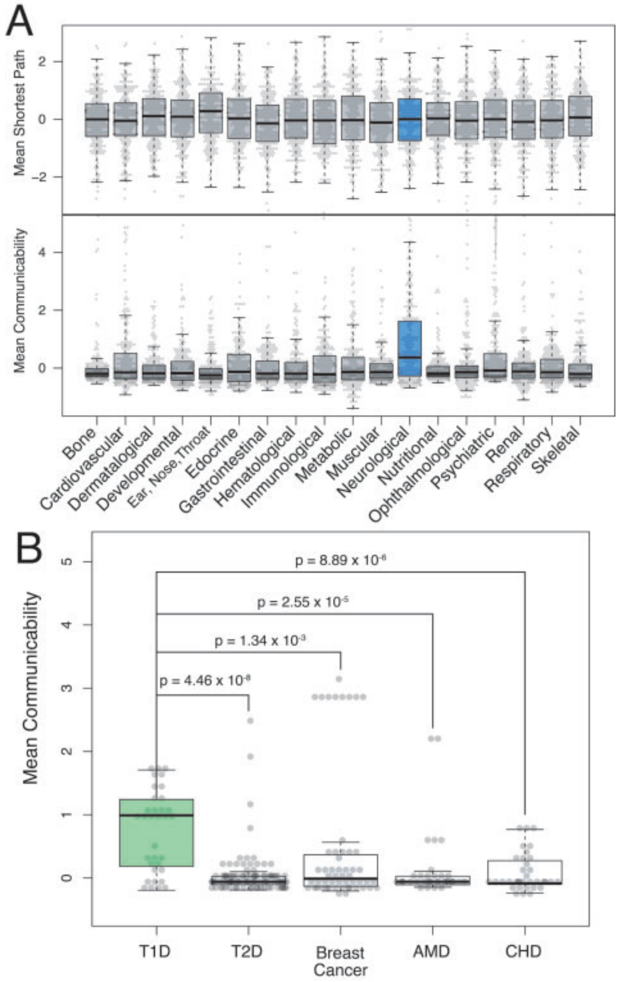

Fig. 1. Finite network communicability performance. Same class disease proteins are more connected within than to proteins in other classes. (**A**) Connectedness among proteins causing neurologic phenotypes does not appear greater than to proteins in other classes by shortest path; however, on average, these same proteins are more connected to each other than to other class proteins by finite network communicability. Interestingly, proteins of the psychiatric class form the second strongest connections. (**B**) Protein products of genes with significant associations to type 1 diabetes are more connected than to proteins associated with other diseases. *P*-values are Bonferroni corrected Wilcox post hoc tests following a significant Kruskal–Wallis test. T1D, type 1 diabetes; T2D, type 2 diabetes; AMD, age-related macular degeneration; CHD, coronary heart disease

ACKNOWLEDGEMENTS

The authors thank the Lupski laboratory at BCM for valuable feedback.

Funding: I.M.C. is a fellow of the BCM Medical Scientist Training Program (T32 GM007330) and was supported by a fellowship from the National Institute of Neurological Disorders and Stroke (F31 NS083159).

Conflict of interest: none declared.

REFERENCES

Brohée,S. *et al.* (2008) Network analysis tools: from biological networks to clusters and pathways. *Nat. Protoc.*, **3**, 1616–1629.

Campbell,I.M. *et al.* (2013) Fusion of large-scale genomic knowledge and frequency data computationally prioritizes variants in Epilepsy. *PLoS Genet.*, **9**, e1003797.

Estrada,E. and Hatano,N. (2008) Communicability in complex networks. *Phys. Rev. E*, **77**, 036111.

Goh,K.-I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Jaccard,P. (1912) The distribution of the flora in the alpine zone. *New Phytol.*, **11**, 37–50.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Noh,J.D. and Rieger,H. (2004) Random walks on complex networks. *Phys. Rev. Lett.*, **92**, 118701.

Welter,D. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.