NetAn: A Python Toolbox Leveraging Network Topology for Comprehensive Gene Annotation Enrichments

Rasmus Magnusson

School of Bioscience, Systems Biology Research Center, University of Skövde, Skövde, Sweden

Abstract

Summary: Gene annotation enrichment analysis is the gold standard for studying the biological context of a set of genes, but available tools often overlook important network properties of the underlying gene regulatory system. This paper introduces the NETwork ANnotation enrichment package, NetAn, which augments annotation analysis with approaches such as the inference of closely related genes to include local neighbors in the analysis, and the separation of over-representation analyses based on network clustering, cliques, and the extraction of separate components. By using NetAn, an example of how such approaches enhance the identification of pertinent annotations in human gene sets is demonstrated.
Availability and Implementation: The software is implemented in Python under a GNU GENERAL PUBLIC LICENSE V3, and can be downloaded from https://github.com/rasma774/netan.
Contact: rasmus.magnusson@his.se
Supplementary Information: The code relevant to this article is available in the doc/article_material folder of the NetAn repository.

Introduction

Gene annotation enrichment analysis is the cornerstone approach to establishing the association of biological functionality with a set of investigated genes. While several tools for identifying annotations from gene sets have been developed [1]–[3], most of them encounter challenges related to at least two key factors. Firstly, if the studied gene set contains genes associated with two or more independent annotations, the signal will effectively be obscured in each individual statistical test, thereby increasing the risk of false negative identifications. Secondly, genes form a complex network; yet, network topologies are frequently overlooked when analyzing gene set annotation enrichments. To address these issues, the Network Annotation Enrichment package (NetAn) was designed, a Python toolkit intended to integrate a network perspective into annotation enrichment analyses.

Implementation

NetAn was developed as a Python 3 package, and is available under a GNU General Public License V3. The package, along with a Jupyter Notebook tutorial, is freely available at https://github.com/rasma774/netan. NetAn is installed locally using the Python package management system pip, to allow for a simple and fast install. While NetAn uses GO annotations [2] and a processed STRINGdb network [4] as the default option, there is also an option for using any user-supplied annotation set or network.

The core of NetAn regards how gene annotation enrichment analyses can be augmented by a priori defined biological networks. The first functionality regards the option to expand the set of studied genes. With a small starting set of genes, NetAn can add statistical power by using an algorithm that calculates the sum of edge weights for all connected nodes, and norms by an estimated background. NetAn then appends the top M*N associated genes, where N is a factor of the number of input genes M (default N=2.0).

While such leverages of network topology in annotation enrichment analyses have been used before, [5]–[7], few approaches have used networks to disentangle convoluted cellular mechanisms in annotation analyses. If the studied gene set involves genes from two or more moderately independent cellular processes, the signal is obscured, and false negative identifications will be more likely. An optimal approach would instead separate the set of genes into clear-cut fractions. To this end, NetAn first draws the adjacency matrix of the input gene set from the loaded network, and applies either K-means clustering, maximal clique identification, or the extraction of separated network components to sort genes into individual sets. Next, the genes in each set are tested using a Fisher's exact test-based over-representation analysis (Fig. 1a).

To analyze the interdependence between identified gene subsets, NetAn has a functionality where the average shortest path length between all gene cluster pairs is computed and compared to the average path length of the loaded network. The significance of this difference in distance is tested using a non-parametric Mann-Whitney U test. For efficiency, if the number of pairs between the clusters is too large to calculate efficiently, NetAn randomly samples pairs in batches until the mean converges.

Lastly, NetAn provides a functionality to export results in the form of a csv-textfile of annotation and odds ratios (ORs), together with nominal an multiple testing-corrected p-values. The package also offers a built-in plotting tool. In summary, a user of NetAn can go from a list of identified genes to a plot of sets of functionally independent gene sets with an annotation-based biological context, all in <5 lines of Python code.

Results

NetAn disentangles unrelated cellular processes: a multiple sclerosis case study

To demonstrate NetAn's application to relevant disease genes, associated gene sets from the Diseases database [8] were used. Specifically, genes linked to multiple sclerosis (MS), a disorder where the immune system damages the insulating myelin covering of nerve cells, were selected. In the Diseases set, MS was associated with 111 genes. Initial over-representation analysis showed a strong prevalence of immunologically-related annotations in the top results (Fig. 1a).

However, employing NetAn to segregate genes into functionally related groups using network clustering revealed a second cluster of genes related to nerve cells and myelin sheets. This set had 8 significant terms (p_{adj} <0.05, Benjamini-Hochberg-adjusted), whereof 6 were also significant in the non-clustered analysis. However, their rankings were notably lower in the original analysis (mean position 45.25), potentially leading to oversight, showcasing how a network stratification can greatly benefit the study of convoluted diseases.

The case study continued by studying the average shortest gene-gene path distances, both between and within clusters, as implemented in NetAn. The neuron-related cluster exhibited a 38% closer average distance between genes than that of the full network's average path, which further suggested that the neuron-related genes form a separate sub-cluster on the human interactome map.

Lastly, NetAn was used to strengthen the disease signal by doubling the 111 MS genes before the over-representation analysis. This augmentation with topologically close genes yielded 65 new significant annotations (p_{adj} <0.05, Benjamini-Hochberg-adjusted), with annotations heavily tied to the immune system. These 65 annotations are available in Supplementary Material 1.

NetAn increases statistical power in a broad array of diseases

The next test related to NetAn's ability to infer annotations from a broader gene set. To this end the full Diseases database [8], containing 477,770 gene-disease associations related to 5,776 distinct diseases, was considered. From this dataset, 100 diseases with more than 10 but fewer than 100 associated genes were randomly drawn. Subsequently, NetAn was applied to predict annotations based on two scenarios: i) the associated genes, and ii) the associated genes along with their local neighbors within the STRINGdb network [4], as integrated into NetAn.

Each analysis was typically executed within less than 60 seconds. This process was performed for various gene fold expansions, ranging from 0.5 to 2.5 in intervals of 0.5. To illustrate, a fold expansion of 0.5 using an initial set of 10 genes would yield an additional 5 genes. The results indicated that NetAn consistently augmented the number of significant annotations (p_{adj} <0.05, Benjamini-Hochberg corrected). For the five

fold expansion inputs between 0.5 and 2.5, the number of significant annotations increased 1.3, 1.5, 2.7, 2.2, and 3.5-fold, respectively.

Importantly, annotations that were exclusively significant after gene expansion were near significance in the over-representation analysis based on the non-expanded gene set. Median nominal p-values for these annotations were all significant, with median odds ratios (ORs) of 18.0, 16.1, 13.8, 11.1 and 11.7, respectively (Fig. 1b).

To compare NetAn's performance with other network-based annotation enrichment tools, it was also applied to the example dataset provided with the netGO tool [9]. This dataset comprises nearly 330,000 gene-annotation pairs and includes 30 genes associated with BRCA breast cancer.

It is noteworthy that netGO takes 5-25 minutes for this application, while NetAn completed the task in seconds. Moreover, netGO identified 9 annotations after multiple testing correction, all of which were associated with cancers in various tissues, including breast, colon, and lungs. In parallel, the standard application of NetAn identified two annotations, both of which were related to breast cancer and were also identified by netGO.

Acknowledgments

This work was supported by the Systems Biology Research Centre at the University of Skövde under grants from the Knowledge Foundation [20200014], Petrus och Augusta Hedlunds Stiftelse [M-2023-2054], and Stiftelsen B von Beskows stipendiefond as administered by The Royal Swedish Academy of Sciences [BS2023-0031]. I want to thank Senior Lecturer Zelmina Lubovac for her input on the manuscript.

References

[1] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, 2016, ISSN: 13624962. DOI: 10.1093/nar/gkv1070.

[2] T. G. O. Consortium, S. A. Aleksander, J. Balhoff, *et al.*, "The Gene Ontology knowledgebase in 2023," *Genetics*, vol. 224, no. 1, iyad031, May 2023, ISSN: 1943-2631. DOI: 10.1093/genetics/iyad031. [Online]. Available: https://doi.org/10.1093/genetics/iyad031.

[3] R. Magnusson and Z. Lubovac-Pilav, "TFTenricher: a python toolbox for annotation enrichment analysis of transcription factor target genes," *BMC Bioinformatics*, vol. 22, no. 1, pp. 1–6, 2021, ISSN: 14712105. DOI: 10.1186/s12859-021-04357-4. [Online]. Available: https://doi.org/10.1186/s12859-021-04357-4.

[4] D. Szklarczyk, A. Franceschini, S. Wyder, *et al.*, "STRING v10: Protein-protein interaction networks, integrated over the tree of life," *Nucleic Acids Research*, vol. 43, no. D1, pp. D447–D452, 2015, ISSN: 13624962. DOI: 10.1093/nar/gku1003.

[5] E. Glaab, A. Baudot, N. Krasnogor, R. Schneider, and A. Valencia, "EnrichNet: network-based gene set enrichment analysis," *Bioinformatics*, vol. 28, no. 18, pp. i451–i457, Sep. 2012, ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bts389. [Online]. Available: https://doi.org/10.1093/bioinformatics/bts389.

[6] S. Bovo, P. L. Martelli, P. Di Lena, and R. Casadio, "NETGE-PLUS: Standard and Network-Based Gene Enrichment Analysis in Human and Model Organisms," *Journal of Proteome Research*, vol. 19, no. 7, pp. 2873–2878, Jul. 2020, ISSN: 1535-3893. DOI: 10.1021/acs.jproteome.9b00749. [Online]. Available: https://doi.org/10.1021/acs.jproteome.9b00749.

[7] H. Cousins, T. Hall, Y. Guo, *et al.*, "Gene set proximity analysis: expanding gene set enrichment analysis through learned geometric embeddings, with drug-repurposing applications in COVID-19," *Bioinformatics*, vol. 39, no. 1, btac735, Jan. 2023, ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btac735. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac735.

[8] S. Pletscher-Frankild, A. Pallejà, K. Tsafou, J. X. Binder, and L. J. Jensen, "DISEASES: Text mining and data integration of disease-gene associations," *Methods*, vol. 74, no. 2015, pp. 83–89, 2015, ISSN: 10959130. DOI: 10.1016/j.ymeth.2014.11.020. [Online]. Available: http://dx.doi.org/10.1016/j.ymeth.2014.11.020.

[9] J. Kim, S. Yoon, and D. Nam, "NetGO: R-Shiny package for network-integrated pathway enrichment analysis," *Bioinformatics*, vol. 36, no. 10, pp. 3283–3285, 2020, ISSN: 14602059. DOI: 10.1093/bioinformatics/btaa077.

Figures

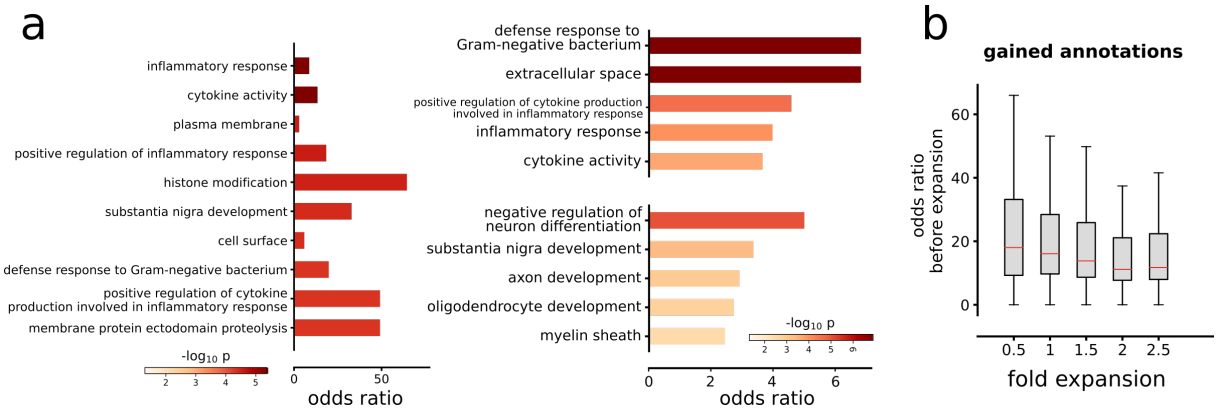

Figure 1: Figure 1: Examples of NetAn functionalities. In a), the 111 MS-related genes were used as input to NetAn, with standard settings. The over-representation analysis showed a clear immune-related response. However, using NetAn's built-in network clustering functionality, a second annotation group related to nerve cells emerges. b) By expanding the gene set tested in the over-representation analysis, as implemented in NetAn, there was a clear increase in significant annotations. These annotations were near significance before expansion. However, the original OR of the added annotations tended to decrease the more the gene set was expanded.