HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2020 February 26.

Published in final edited form as:

Pac Symp Biocomput. 2020 ; 25: 439–450.

Graph-based information diffusion method for prioritizing functionally related genes in protein-protein interaction networks

Minh Pham,

Integrative Molecular and Biomedical Sciences Graduate Program, and Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA

Olivier Lichtarge

Departments of Molecular and Human Genetics, Structural and Computational Biology and Molecular Biophysics, Biochemistry and Molecular Biology, and Pharmacology, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA

Abstract

Shortest path length methods are routinely used to validate whether genes of interest are functionally related to each other based on biological network information. However, the methods are computationally intensive, impeding extensive utilization of network information. In addition, non-weighted shortest path length approach, which is more frequently used, often treat all network connections equally without taking into account of confidence levels of the associations. On the other hand, graph-based information diffusion method, which employs both the presence and confidence weights of network edges, can efficiently explore large networks and has previously detected meaningful biological patterns. Therefore, in this study, we hypothesized that the graph-based information diffusion method could prioritize genes with relevant functions more efficiently and accurately than the shortest path length approaches. We demonstrated that the graph-based information diffusion method substantially differentiated not only genes participating in same biological pathways ($p << 0.0001$) but also genes associated with specific human drug-induced clinical symptoms ($p << 0.0001$) from random. Furthermore, the diffusion method prioritized these functionally related genes faster and more accurately than the shortest path length approaches (pathways: $p = 2.7e-28$, clinical symptoms: $p = 0.032$). These data show the graph-based information diffusion method can be routinely used for robust prioritization of functionally related genes, facilitating efficient network validation and hypothesis generation, especially for human phenotype-specific genes.

Keywords

Network diffusion; Network validation; Gene function annotation

minh.pham@bcm.edu.

1. Introduction

Biological networks, such as protein-protein interaction (PPI) networks, facilitate functional interpretation of large omics data[1] and knowledge discovery of disease genes[2] and drug targets[3]. One of the major applications of biological network validation is validating functionally related genes, in which genes of interest that are highly connected to genes annotated with specific functions in the networks are more likely to have the same functions. Biological networks extensively support this application because they aggregate biological associations of a large number of genes[1,4], thus allowing exploration of functionality of uncharacterized genes in a context of other genes. Biological networks also characterize the complexity of biology as they support integrating information of different types of biological processes from multiple data sources. For example, STRING[4], a PPI network database, provides network information of different biological processes, such as physical protein-protein interaction, protein fusion, and co-expression. The network information comes from experimental data, computational predictions, and text mining, adding different levels of confidence for the network associations. Biological networks, therefore, are often very complex with thousand nodes and million edges, often with confidence weight features. Methods that can handle the complicated nature of biological networks and efficiently explore network information are necessary to speed up knowledge discovery.

Shortest path length methods are routinely used to validate functionally related genes using biological network information[5]. Non-weighted shortest path is the path that requires smallest number of edges to travel between two nodes. On the other hand, weighted shortest path is the path with smallest sum of edge weights between two nodes. The general idea is that genes that are in closer distance or *have shorter paths* are often more likely to be involved in same biological processes. Non-weighted shortest path length is more often used than weighted shortest path length because it is easier to interpret how genes of interest interact directly with each other. However, without considering confidence weights of edges, the method could prioritize the interactions that are not supported by many evidences. The edge weights demonstrate how strongly genes are interacted with each other based on experimentally derived data[1] and/or the number of supporting publications from text mining[4] for given associations. Therefore, edge weights contain useful information to interpret biologically associations better and should be integrated.

A problem with shortest path length approaches is that they are computationally expensive. Multiple methods have been proposed yet it is still challenging, especially when computing for weighted graphs. For example, Dijkstra's algorithm [6] is a popular method to compute shortest path length, both weighted and non-weighted. To determine shortest path, Dijkstra's algorithm goes through unvisited nodes with the smallest distance from the starting node, continue to other unvisited nodes and update the neighbor's distance[6]. For a network of $|V|$ nodes and $|E|$ edges, the time to compute a given shortest path length can take up to $O(|E| + |V|\log|V|)$[7]. For the application of prioritizing and validating functionally related genes, shortest path length will have to be computed for every pair of a validated gene and a gold standard gene of known functions, increasing computational time. Because the shortest path length approaches need extensive resources, they hinder full exploration of network information and knowledge discovery.

Graph-based information diffusion method offers a solution. Graph-based information diffusion method[8,9] simulates the flow of liquid or *information*, starting from nodes with certain information or *known functional annotations*, and spreading the *information* throughout the network to other nodes. Nodes that are closer to the starting nodes, meaning that they are few edges away and the edges have higher confidence weights, will receive more information signals and thus, more likely to share similar functions. The graph-based information diffusion method performs fast on large networks, allowing quick exploration of network information and knowledge discovery. Previously, graph-based information diffusion has been applied to biological networks and accurately predict functional annotations of uncharacterized protein structures[9] and novel antigen for antimalarial drug[10]. This suggests that the diffusion method may robustly prioritize genes associated with similar biological processes and even human phenotypes.

Because the graph-based information diffusion method employs both the presence and confidence weights of network edges, and the method has robustly predicted protein function, we hypothesized that the diffusion method could prioritize functionally related genes more accurately than the shortest path length approaches. Our data validated that the diffusion method robustly prioritized genes participating in same biological pathways and gene ontologies from random. We further demonstrated that the predictions for pathway genes of the diffusion method outperformed the shortest path length approaches. Finally, we showed that the diffusion method can predict genes associated with human-like clinical phenotypes in mice with statistically better performance than the shortest path length measures. Overall, our study advocated the use of graph-based information diffusion for efficient prioritization of functionally related genes, supporting robust validation of omics data and hypothesis generation of novel disease and drug mechanisms.

2. Materials and Methods

2.1. Data sources

2.1.1. Biological network information—The biological network that we used was the protein-protein interaction (PPI) STRING network[11] (version 10.0), which can be downloaded from http://version10.string-db.org/. For our analyses, we used only *Homo sapiens* protein interaction network data, which consists of 19,236 proteins and 4,272,402 edges. In order to construct a weighted graph, we used combined confidence scores of edges. Therefore, the constructed graph considered combined probabilities of predicted associations from different evidence channels, i.e. conserved neighborhood, gene fusion, phylogenetic co-occurrence, co-expression, large-scale experiments, literature co-occurrence, and databases of biological pathways and physical protein interactions. Predictions from pathway database imports account for 5% predicted associations (7,938 genes and 212,370 edges) in the combined network, indicating that the network is not restricted to only pathway information. Edges with greater weights have higher confidence levels. Methods that can leverage edges with higher confidence weights can prioritize more functionally relevant genes that have higher associative probabilities predicted by multiple evidence channels.

2.1.2. References for pathway and ontology data—In order to validate functional gene prioritization abilities of different approaches, we selected a number of popular manually curated pathway and ontology data that have been pre-processed by Enrichr database[12] (https://amp.pharm.mssm.edu/Enrichr). Pathway references used were Reactome[13] (version 2016), KEGG[14] (version 2016), and WikiPathways[15] (version 2016). Gene Ontology Annotation (GOA) for aspects of Biological Process (version 2017), Cellular Component (version 2017), and Molecular Function (version 2017)[16,17] were also examined. The numbers of gene sets and total gene coverages of the validated pathways and ontologies are summarized in Table 1. There are only 3 gene sets that are present in all of the three selected pathway databases, suggesting that these pathway databases are overall distinct from each other.

2.1.3. References for genes associated with human drug-induced clinical symptoms—The genes associated with mouse phenotypes are compiled from Mouse Genome Informatics database[18] (MGI: http://www.informatics.jax.org). The genes selected were those that when being knocked out, yield substantial mouse phenotypes. We were interested in gene sets for relevant human clinical phenotypes, yet the information was not readily available. Therefore, we selected gene sets for mouse phenotypes that resemble drug-induced side effect symptoms in human (e.g. “parotid gland inflammation” and “joint swelling”), assuming that the genetics behind these phenotypes are similar in human and mice. The human drug-induced side effect symptoms are annotated in SIDER[19] (version 4.1) (http://sideeffects.embl.de). Combining the two databases gave us 266 human-like clinical phenotypes in mice and their gene sets cover in total 2,856 genes.

2.2. Network analysis methods

2.2.1. Graph-based information diffusion method—Graph-based information diffusion method was previously applied on biological networks[8,9] using the following formula:

$$f = (I + \alpha L)^{-1} y \quad (1)$$

where L = the Laplacian matrix of the combined STRING protein network

I = the identity matrix

y = a vector of labels prior to diffusion

f = the vector labeled after diffusion

$\alpha = 1/\| L \|_1$(ensuring convexity of the cost function[8])

Every node or genes in the network was considered with a label. Diffusion was performed throughout the whole constructed STRING network. For the vector *y*, we initialized the diffusion process by setting the *source nodes* or genes with known functional annotations to 1 and all other network nodes or *recipient nodes* to 0. After diffusion, the diffused signals or *diffusion values* that the *recipient nodes* received, as represented in the vector f, were ranked, with higher values suggesting that they had higher probability to share similar functions with

the *source nodes*. The known functional annotations of the source nodes or genes can be whether these genes participate in known biological pathways and ontologies and/or are associated with specific phenotypes. The method was run on a processor of 2.9 GHz Intel Core i5 and memory of 16 GB 1867 MHz DDR3.

2.2.2. Shortest path length (SPL) approaches—Dijkstra's algorithm[6] was utilized. The running time could take[7]:

$$O(|E| + |V|\log|V|) \quad (2)$$

where $|V|$ = the number of nodes

$|E|$ = the number of edges

We applied networkx python package[20] to process the network data and compute shortest path length, both non-weighted and weighted. The codes were run on the same computational system used for the diffusion method. Non-weighted shortest path length method prioritizes the path with fewest steps or edges while weighted shortest path method prioritizes the paths with the lowest sum of edge weights. The STRING network that we used associates a higher edge weight with a higher confidence level. Therefore, in order to prioritize the path with highest confidence using the shortest path length method, we constructed another graph with the inversed values for edge weights. The transformed graph still has the same edge connections with the originally constructed STRING network but with inversed edge weight values. Both non-weighted and weighted shortest path length calculations were applied on the transformed network.

2.3. Diffusion method to validate genes in same pathways and ontologies

We tested whether the diffusion method could detect genes that are functionally related more than random. We used references of biological pathways and gene ontologies, as described in Section 2.1, for this analysis. Each gene set was randomly split into half. Diffusion signals would start from either of the halves (*source nodes*) and propagate throughout the entire network. We would compare the signals received by the other genes in the gene set and by random genes. Genes that are more connected to the diffusion *source nodes* would receive more diffusion signals. The random genes were selected either uniformly in the network or by matching degrees with the recipient genes in the gene set. This whole process was repeated with the other half of the gene set as the *source nodes* for diffusion. Therefore, there were two experiments for each gene set in the references. Kolmogorov–Smirnov test was performed to compare the distributions of diffusion signals received by pathway genes and random genes.

2.4. Comparisons of predictive performance for prioritizing functionally related genes

We evaluated whether diffusion method could prioritize genes of same functions from random genes more robustly than the shortest path length methods. Because the shortest path length methods are computationally intensive, we had to arbitrarily limit our analyses to only Reactome pathways with 6 to 20 genes, which gave us 591 pathways covering in total 3,242 genes. These empirically selected sizes of Reactome pathway let us to finish the

shortest path length calculations in a week. We randomly split each of these pathways into halves. Diffusion signals started from one half and the received signals were used to predict the other half of the same pathway. Average shortest path length to one half of the pathways was calculated for the other half of the pathway and random genes. Genes that are closer to the known pathway genes, either through diffusion or shortest path length methods, were more likely to be in the same pathways. We measured area under receiver operating characteristic (AUROC) to evaluate predictive performance of different methods. For the diffusion method, the ranking was based on signals of the recipient nodes after diffusion. For the shortest path length approaches, genes that were ranked higher were those that have shorter average shortest path lengths. The truth table was whether those genes were in the same pathways with the initial source genes. We could not perform shortest path length predictions over every node of the network due to limited time and resources, thus we randomly selected ($3 \times n$) random genes in the network, in which n is the number of pathway recipient genes, to evaluate AUROC for these methods. Finally, the distributions of predictive AUROC values for the diffusion and shortest path length methods were compared by Kolmogorov–Smirnov test.

2.5. Diffusion method to prioritize genes associated with drug-induced clinical symptoms

Going beyond genetic and molecular processes, we explored whether the diffusion method could explore genes associated with human phenotypes. Specifically, we tested whether the diffusion method could detect genes that were linked to human drug-induced clinical symptoms. Similar to the approaches described in sections 2.3 and 2.4, we first explored whether the diffusion method could differentiate genes associated with specific clinical symptoms from random and compared the predictive performance of the method against the weighted and non-weighted shortest path length approaches. For comparing the diffusion values between pathway genes and random genes, we performed the experiments on the whole 266 gene sets associated with human-like clinical phenotypes in mice from MGI and SIDER. For the performance comparisons with shortest path length approaches, we limited the analysis to only 128 symptom-related gene sets with 6 to 60 genes, covering 1,496 genes in total. The empirically selected size range of the gene sets allowed us to finish shortest path length calculations in a week.

3. Results and Discussions

3.1. The diffusion method robustly prioritized functionally related genes

3.1.1. The diffusion method robustly prioritized pathway-specific genes—We explored whether the diffusion method detected genes participating in same biological pathways, i.e. whether genes in the same pathways diffused to each other more than to random genes. Fig. 1 shows that genes in the same pathways statistically diffused to each other more than random (KS test: $p << 0.0001$ for both degree-matched and uniformly selected random). Pathway genes often have higher degrees because they are studied more, thus more likely to connect to other in the PPI network than lower degreed genes. This is demonstrated as the distributions of the degree-matched random genes were skewed to higher diffusion values than the distributions of uniformly selected random genes (Fig. 1).

However, even when controlling for node degrees, the diffusion method still substantially differentiated pathway genes from degree-matched genes.

It is worth noting that the observed pattern was consistent across multiple pathway references (i.e. Reactome, KEGG, and WikiPathways), which have different numbers of gene sets and gene coverages (Table 1), suggesting that the observation is global. In addition, interestingly, the distributions of recipient diffusion signals for biological pathways seemed to close to unimodal, centering at larger diffusion values, while distributions for random genes were bimodal, spreading over larger ranges of values. Because selected random genes are involved in multiple biological processes, this data suggests the diffusion method specifically prioritized genes participating in same biological pathways.

3.1.2. The diffusion method robustly prioritized gene ontology-specific genes—Similar to pathway-specific genes, the diffusion method robustly detected genes linked to same gene ontologies. For diffusion initialized from a portion of gene ontologies, genes in the same gene ontologies received significantly higher diffusion signals than random genes, whether they were degree-matched or not (Fig. 2; KS test: $p << 0.0001$). Interestingly, the distributions of recipient diffusion values for ontology-related genes seemed to closer to bimodal with more smaller signal values, instead of unimodal distributions centered at larger diffusion values like pathway-specific genes. This is potentially because ontology-specific genes participate in multiple biological processes, thus making the predictive performance of the diffusion method less robust. Overall, these data demonstrate the usability of diffusion method in detecting functionally similar genes in biological networks.

3.2. The diffusion method outperformed the shortest path length approaches in prioritizing functionally related genes

Because the diffusion method employs both the number of edges and edge confidence weights for measuring distance, we hypothesized that the diffusion method can detect functionally related genes better than both non-weighted and weighted shortest path length approaches. Because shortest path length detection requires intensive computational time, we limited our analyses to small pathways, specifically Reactome pathways with 6 to 20 gene members. Overall, we observed that all three methods performed fairly well, in which for the majority cases, AUROC can be achieved up to 1.0, confirming that genes that are functionally similar diffused better to each other and were closer in distance as measured by both weighted and non-weighted shortest path length (Fig. 3). However, the diffusion method stood out to be the best performing method overall (Fig 3). The AUROC distribution for the diffusion method was statistically skewed more to higher AUROC values than those of the non-weighted and weighted shortest path length approaches (KS test: $p_{\text{diffusion vs non-weighted SPL}} = 2.7e\text{-}28$, $p_{\text{diffusion vs weighted SPL}} = 2.8e\text{-}11$). Non-weighted shortest path length performed slightly better than weighted shortest path length ($p_{\text{non-weighted vs weighted SPL}} = 2.7e\text{-}10$), suggesting that the number of edges between genes was probably more important than the edge confidence weight, at least in the context of small pathways. However, by employing both of these elements, diffusion could predict functionally related genes the best.

3.3. The diffusion method robustly predicted human phenotype-related genes

3.3.1. The diffusion method robustly prioritized genes linked to specific human drug-induced clinical symptoms—

Because the diffusion method robustly predicted functionally similar genes, we explored the possibility of using the diffusion method to detect phenotype-related genes in biological networks. We compiled genes that, when being knocked out, give rise to human-like drug-induced clinical symptoms in mice from Mouse Genomics Informatics (MGI) database. We observed that genes associated with similar symptoms diffused to each other statistically more than to random genes, whether they were degree-matched or uniformly selected (Fig. 4, KS test: $p << 0.0001$). Interestingly, the distribution of diffusion values for symptom-related genes is bimodal, similar to what we observed in Gene Ontologies. This is consistent with the fact that clinical symptoms are often involved with multiple biological processes. These data show that the diffusion method robustly utilized biological network information to detect genes that are involved in not only fundamental biological processes but also human phenotypes.

3.3.2. The diffusion method outperformed the shortest path length approaches in prioritizing clinical symptom-specific genes—

Because the diffusion method predicted genes participating in same biological processes more robustly than the shortest path length approaches, we hypothesized that the diffusion method could also outperform in predicting genes associated with specific human drug-induced clinical symptoms. Overall, the predictive performances for symptom-associated genes of all methods were not as good as their predictions for pathway-related genes (Fig. 3 and 5). However, the diffusion method still statistically outperformed the shortest path length methods (Fig. 5, KS test: $p_{\text{diffusion vs non-weighted SPL}} = 0.032$, $p_{\text{diffusion vs weighted SPL}} = 5.1e\text{-}07$), with 48.8% of predictions had AUROC above 0.70. On the other hand, the mean AUROC of predictions by the non-weighted shortest path length method is 0.62 while the mean AUROC of the weighted shortest path length method is slightly higher at 0.66 (Fig. 5, KS test: $p_{\text{non-weighted vs weighted SPL}} = 3.1e\text{-}03$). These data show that the diffusion method, by combining both the number of steps like the non-weighted shortest path length approach and the edge weight like the weighted shortest path length, robustly prioritized relevant genes for specific human phenotypes.

4. Conclusions

Validating functionally related genes is one of major tasks of biological network analysis. In this study, we proposed using the graph-based information diffusion method, instead of the routine shortest path length approaches, in order to prioritize functionally similar genes faster and more accurately. While shortest path length methods employ either a single shortest path (non-weighted) or purely confidence weights of network edges (weighted), the diffusion method considers both edge confidence weights and multiple paths that genes are connected to each other in the networks. We demonstrated that the diffusion method prioritized *pathway*-, *ontology*-, and *clinical symptom*-specific genes more robustly than the shortest path length methods. These data suggest that the diffusion method may detect functionally related genes that the shortest path length methods miss. In addition, because the diffusion method can quickly explore the whole network, it allows full utilization of

network characteristics, such as global topology and local structure, in making predictions. The method also supports investigation of more candidate genes simultaneously in the networks, up to the maximum of all network nodes, thus generating a greater number of hypotheses for novel gene functionality, such as discovery of disease genes and drug targets. A limitation of the diffusion method is that it is not as easy to interpret how genes of interest interact directly with each other as for using the non-weighted shortest path length method. Detailed investigations of the multiple connected paths of genes of interest are necessary to fully understand their functional relations.

Acknowledgments

The authors would like to thank Daniel Konecki, Kwanghyuk Lee, and Jennifer Asmussen for their technical support and discussions. This work has been supported by the National Institutes of Health [GM079656, GM066099, AG061105]. Conflict of Interest: none declared.

References

1. Li T, Wernersson R, Hansen RB, et al. A scored human protein–protein interaction network to catalyze genomic interpretation. Nat Methods. 2017;14(1):61–64. doi:10.1038/nmeth.4083 [PubMed: 27892958]
2. Suratanee A, Plaimas K. Network-based association analysis to infer new disease-gene relationships using large-scale protein interactions. Kanungo J, ed. PLoS One. 2018;13(6):e0199435. doi:10.1371/journal.pone.0199435 [PubMed: 29949603]
3. Ji X, Freudenberg JM, Agarwal P. Integrating Biological Networks for Drug Target Prediction and Prioritization. In: ; 2019:203–218. doi:10.1007/978-1-4939-8955-3_12
4. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019;47(D1):D607–D613. doi:10.1093/nar/gky1131 [PubMed: 30476243]
5. Al-Ramahi I, Lu B, Di Paola S, et al. High-Throughput Functional Analysis Distinguishes Pathogenic, Nonpathogenic, and Compensatory Transcriptional Changes in Neurodegeneration. Cell Syst. 2018;7(1):28–40.e4. doi:10.1016/j.cels.2018.05.010 [PubMed: 29936182]
6. Dijkstra EW. A note on two problems in connexion with graphs. Numer Math. 1959;1(1):269–271. doi:10.1007/BF01386390
7. Fredman ML, Tarjan RE. Fibonacci Heaps And Their Uses In Improved Network Optimization Algorithms. In: 25th Annual Symposium OnFoundations of Computer Science, 1984 IEEE; :338–346. doi:10.1109/SFCS.1984.715934
8. Lisewski AM, Lichtarge O. Untangling complex networks: risk minimization in financial markets through accessible spin glass ground states. Physica A. 2010;389(16):3250–3253. doi:10.1016/j.physa.2010.04.005 [PubMed: 20625477]
9. Venner E, Lisewski AM, Erdin S, Matthew Ward R, Amin SR, Lichtarge O. Accurate protein structure annotation through competitive diffusion of enzymatic functions over a network of local evolutionary similarities. PLoS One. 2010;5(12). doi:10.1371/journal.pone.0014286
10. Lisewski AM, Quiros JP, Ng CL, et al. Supergenomic network compression and the discovery of exp1 as a glutathione transferase inhibited by artesunate. Cell. 2014;158(4):916–928. doi:10.1016/j.cell.2014.07.011 [PubMed: 25126794]
11. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43. doi:10.1093/nar/gku1003
12. Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44(W1):W90–7. doi:10.1093/nar/gkw377 [PubMed: 27141961]
13. Fabregat A, Jupe S, Matthews L, et al. The Reactome Pathway Knowledgebase. Nucleic Acids Res. 2018;46(D1):D649–D655. doi:10.1093/nar/gkx1132 [PubMed: 29145629]

14. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30. http://www.ncbi.nlm.nih.gov/pubmed/10592173. Accessed January 22, 2019. [PubMed: 10592173]

15. Kelder T, Van Iersel MP, Hanspers K, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012;40(Database):D1301–D1307. doi:10.1093/nar/gkr1074 [PubMed: 22096230]

16. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017;45(D1):D331–D338. doi:10.1093/nar/gkw1108 [PubMed: 27899567]

17. Huntley RP, Sawford T, Mutowo-Meullenet P, et al. The GOA database: Gene Ontology annotation updates for 2015. Nucleic Acids Res. 2015;43(D1):D1057–D1063. doi:10.1093/nar/gku1113 [PubMed: 25378336]

18. Eppig JT, Smith CL, Blake JA, et al. Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. In: Methods in Molecular Biology (Clifton, N.J.). Vol 1488 ; 2017:47–73. doi:10.1007/978-1-4939-6427-7_3

19. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Res. 2016;44(D1):D1075–9. doi:10.1093/nar/gkv1075 [PubMed: 26481350]

20. Hagberg AA, Schult DA, Swart PJ. Exploring Network Structure, Dynamics, and Function Using NetworkX.; in Proceedings of the 7th Python in Science Conference (SciPy2008), Varoquaux Gäel, Vaught Travis, and Millman Jarrod (Eds), (Pasadena, CA USA), pp. 11–15, 2008

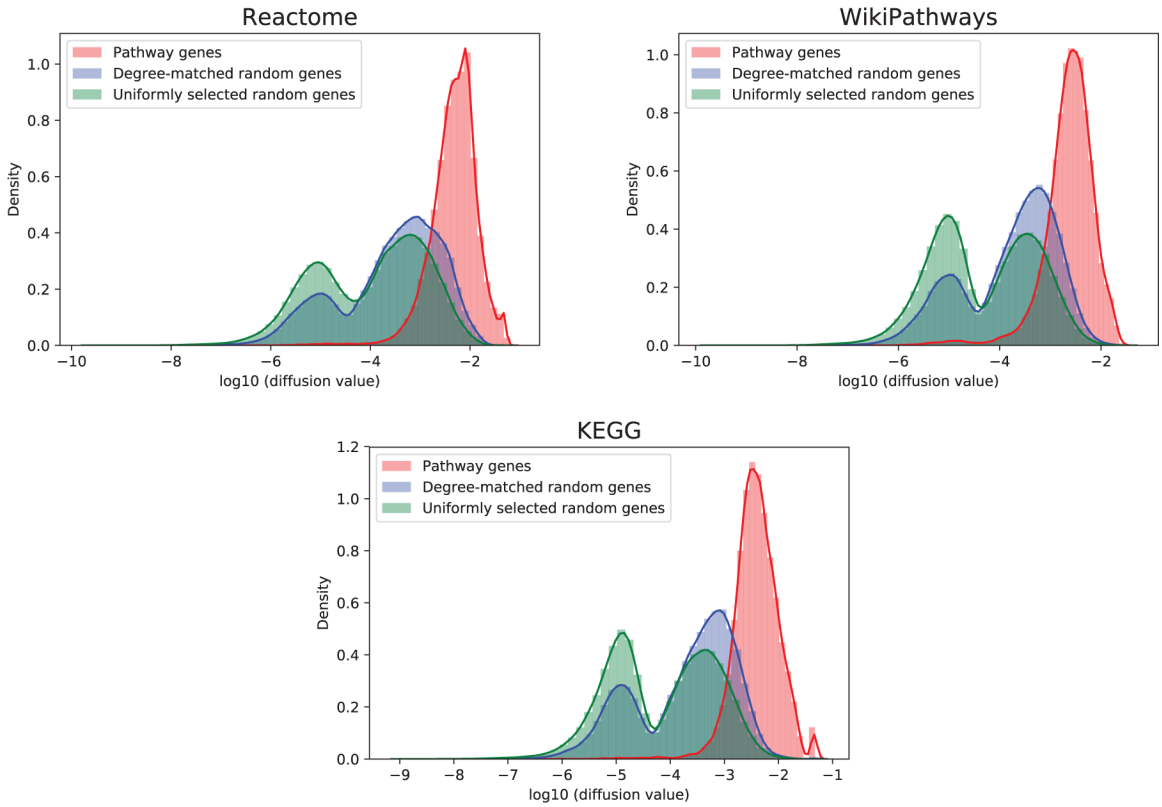

Fig. 1.

The diffusion method robustly prioritized pathway-specific genes. Pathway genes (red) are more connected to each other than to degree-matched random genes (blue) (KS test: $p << 0.0001$) or uniformly selected random genes (green) (KS test: $p << 0.0001$) in the STRING PPI network.

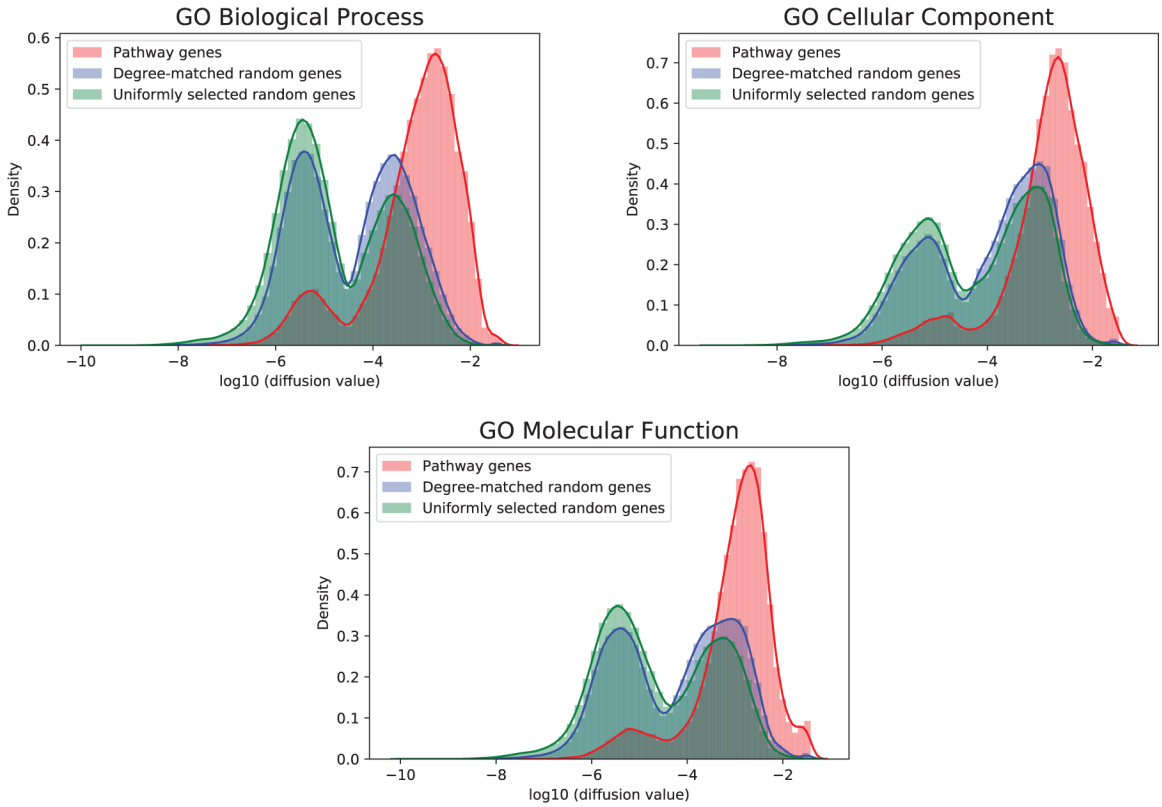

Fig. 2.

The diffusion method robustly prioritized ontology-specific genes. Pathway genes (red) are more connected to each other than to degree-matched random genes (blue) (KS test: $p << 0.0001$) or uniformly selected random genes (green) (KS test: $p << 0.0001$) in the PPI network.

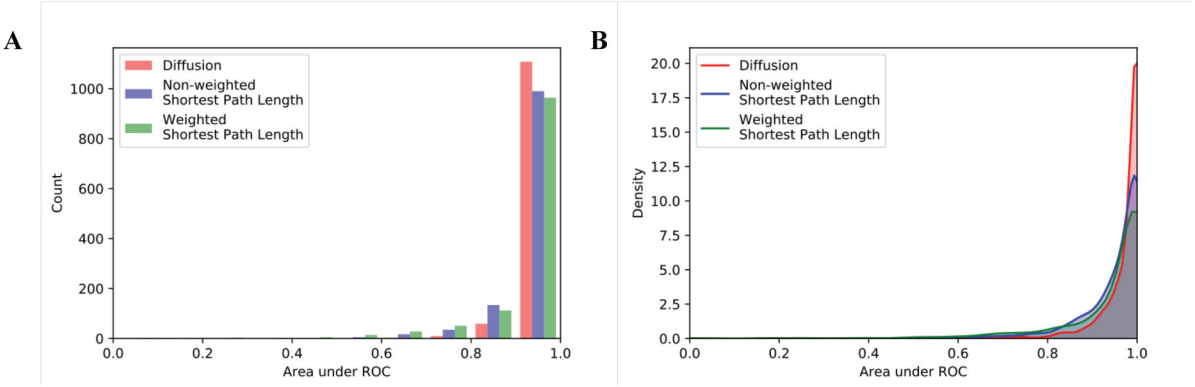

Fig. 3.

The diffusion method (red) detected functionally related genes statistically better than the non-weighted (blue) and weighted (green) shortest path length approaches, as shown in a histogram plot (A) and a kernel density estimation plot (B) (KS test: $p_{\text{diffusion vs non-weighted SPL}} = 2.7e\text{-}28$, $p_{\text{diffusion vs weighted SPL}} = 2.8e\text{-}11$, $p_{\text{non-weighted vs weighted SPL}} = 2.7e\text{-}10$).

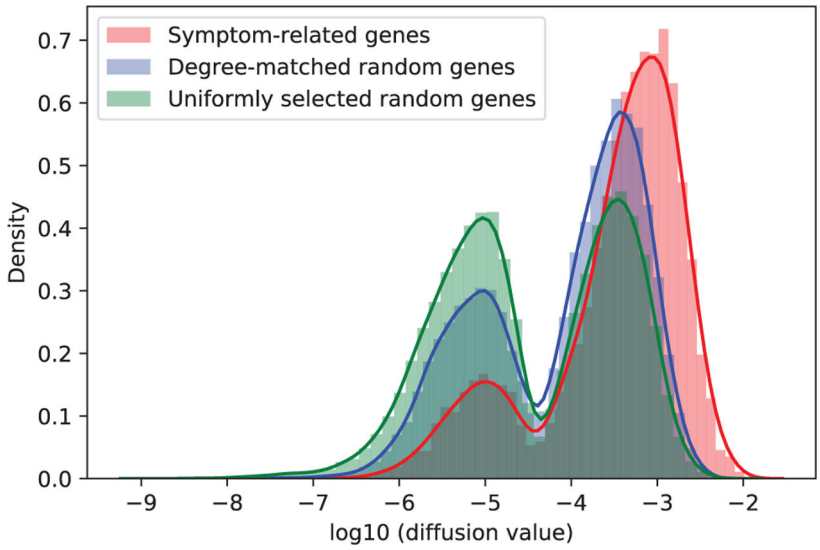

Fig. 4.

The diffusion method robustly prioritized human clinical symptom-related genes (red) from degree-matched (blue) and uniformly selected (green) random genes (KS test: $p << 0.0001$).

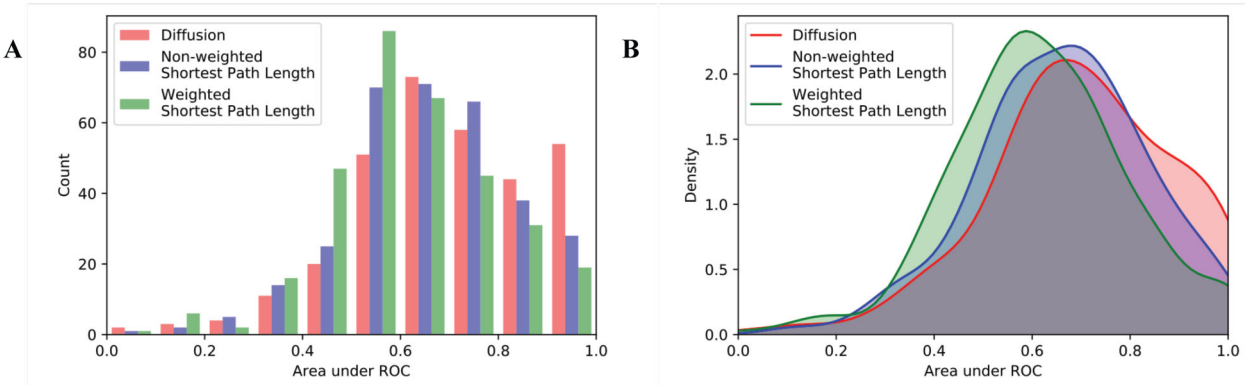

Fig. 5.

The diffusion method (red) detected functionally related genes significantly better than the non-weighted (blue) and weighted (green) shortest path length approaches as shown in a histogram plot (A) and a kernel density estimation plot (B) (KS test: $p_{\text{diffusion vs non-weighted SPL}} = 0.032$, $p_{\text{diffusion vs weighted SPL}} = 5.1\text{e-}07$, $p_{\text{non-weighted vs weighted SPL}} = 3.1\text{e-}03$).

Table 1.

Statistics of pathway and ontology data for validation.

Pathway/Ontology	# gene sets	Total gene coverage
Reactome	1,530	8,973
KEGG	293	7,010
WikiPathways	437	5,966
GO Biological Process	3,166	13,822
GO Cellular Component	636	10,427
GO Molecular Function	972	10,601