On the Design of Advanced Filters for Biological Networks using Graph Theoretic Properties

Kathryn Dempsey[0]*, Tzu-Yi Chen[•], Sanjukta Bhowmick[0], Hesham Ali[0]*
[0]College of Information Science and Technology, University of Nebraska at Omaha
*Department of Pathology & Microbiology, University of Nebraska Medical Center
[•]Department of Computer Science, Pomona College
Contact Email: hali@mail.unomaha.edu

***Abstract*-**Network modeling of biological systems is a powerful tool for analysis of high-throughput datasets by computational systems biologists. Integration of networks to form a heterogeneous model requires that each network be as noise-free as possible while still containing relevant biological information. In earlier work, we have shown that the graph theoretic properties of gene correlation networks can be used to highlight and maintain important structures such as high degree nodes, clusters, and critical links between sparse network branches while reducing noise. In this paper, we propose the design of advanced network filters using structurally related graph theoretic properties. While spanning trees and chordal subgraphs provide filters with special advantages, we hypothesize that a hybrid subgraph sampling method will allow for the design of a more effective filter preserving key properties in biological networks. That the proposed approach allows us to optimize a number of parameters associated with the filtering process which in turn improves upon the identification of essential genes in mouse aging networks.

Keywords-biological networks, network filters, chordal graphs, spanning tree, lethal genes, hubs, clusters

I. INTRODUCTION

A network model that surveys the cellular landscape can contain tens of thousands of probes for multiple states; thus, complexity in can quickly stretch computational limits. Models inherently include noise which must be filtered for accurate analysis where causative structures (nodes/edge groups with biological function) are more easily identified. Previous studies [1, 7, 11, 12] on biological network structure show that structures such as high-degree nodes (hubs), clusters, motifs, and spanning trees all can reveal biological function. Our previous work [2-4] has found that removing noise by identifying redundant structures in the network can further help to reduce noise and improve the biological impact of the model. Figure 1 highlights some of these findings. In our earlier work [2] we demonstrated that finding chordal subnetworks [5]) allows for reduction in network size, maintenance of biological signal, and discovery of previously masked signal. Figure 1.B highlights previous work that a spanning tree (a subnetwork touching all nodes in the network but contains no cycles) maintains hub nodes in the network while removing ~50% of edges; in fact, the biological signal of finding essential genes in these networks is enhanced using the spanning tree filter.

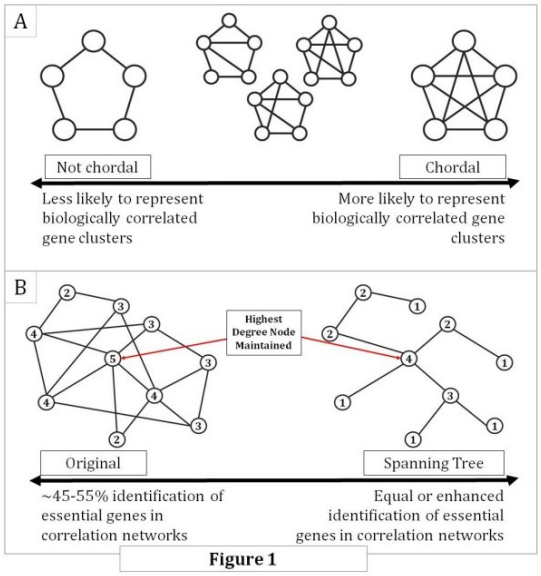
A. Proposed Method

We have examined the lethality of central nodes (hubs, high betweenness nodes, etc) as a measure of the biological impact of hubs, and further we have used Gene Ontology Enrichment of clusters in the network to measure the biological impact of network structure. These studies have identified that spanning trees optimize the identification of lethal (also known as essential) genes in the network, while chordal-filtered networks readily identify clusters that contain critical relationships; however, chordal networks are not particularly adept at identifying lethal nodes [2, 4] and spanning trees are not able to identify clusters of any kind [3]. Thus, the proposed algorithm identifies a spanning tree within the network, then connects clusters of nodes with high density in the network using a chordal re-addition scheme. We suspect that by combining the "best of both worlds" we will be able to optimize on the good characteristics of both the spanning tree and chordal graph filters by identifying lethal hub nodes and conserving critical clusters from the network. This work describes our experimental study to verify our hypothesis.

II. METHODS

We describe our method in the following format: A. Network Creation, or the description of how data were obtained and the creation and filtering of networks, B. Network Analysis and Enrichment, which describes how we define a hub node, and how we assess the biological impact of those hub nodes via integration of biological data and enrichment score, and C. Method Description, in which we describe the algorithm used to identify our enhanced networks that incorporate our previous spanning tree and chordal graph work.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE



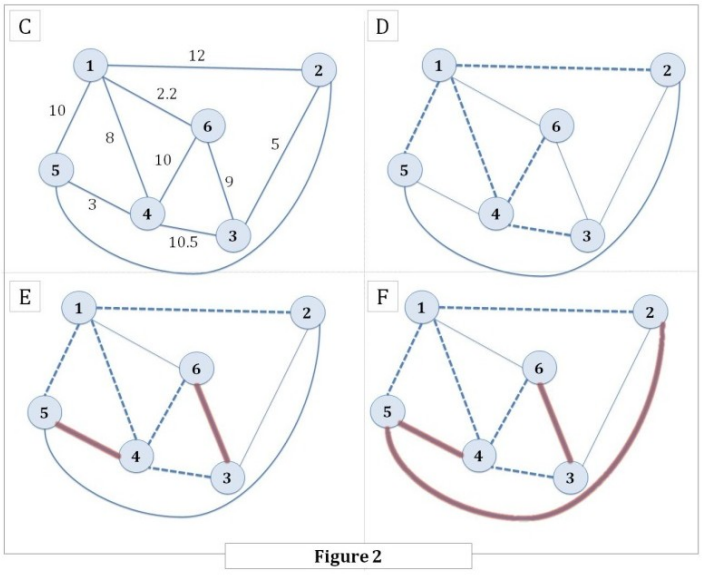


Figure 1 (left): A. Chordal subgraphs are more likely to represent causative relationships. If $gene_A$ regulates $gene_B$ and $gene_A$ regulates $gene_C$, then correlation of $gene_B$ and $_C$ is also likely to be high. B. Spanning trees of biological network maintain high degree nodes [6], and reduce edge count by up to 50% while maintaining essential hubs. **Figure 2 (right):** Our method for identifying a spanning tree with chordal properties, as described in section 3.1. (A) Original network. (B) Maximum spanning tree. (C). Spanning tree augmented with the sibling line. (D) Spanning tree augmented with sibling all.

A. Network Creation

The networks used in this research were correlation networks derived from microarray dataset GSE5078 taken from NCBI's Gene Expression Omnibus (GEO) website [9, 19]. GSE5078 was designed to examine mechanisms behind the processes of aging in the murine brain. The YNG and MID networks represent expression data created from mice at 2 months and 18 months, respectively. Networks were created in parallel by pairwise computation of Pearson Correlation (ρ) [13, 14, and 15] of gene expression values for all genes versus each other on the University of Nebraska at Omaha's Black forest computing cluster. The network was built such that nodes represent genes and edges represent the weighted correlation of two gene's expression. Edges with correlations with p-value <0.005 were not considered statistically significant and were discarded. Networks were then filtered to a correlation threshold of $0.90 \leq \rho \leq 1.00$ to capture only very highly correlated expression values. Both networks created were found to adhere to a power-law degree distribution and exhibit qualities of a modular network.

B. Network Analysis and Enrichment

In correlation networks, there has been no formal definition of a hub node in terms of degree, but it has been agreed that hubs represent those few nodes that are very highly connected in a power-law degree distributed network. Previous studies of centrality in correlation networks [14] have found that examining many thresholds for hub definition identifies an optimal threshold; as such, we use thresholds of top 1, 5, 10 15, 20, and 25% of degree-ranked nodes in the network were used to identify the optimal threshold. For each node in the network, we determined if an *in vivo* knockout mutation had been performed on that gene. If that mutation had been performed, we determined if the mutation was lethal/essential. Then we perform an enrichment analysis to determine the log-odds ratio enrichment of lethal genes in hub nodes versus the rest of the network, referred to as the background set:

$$Enrichment = log_2\left(\frac{b/n}{B/N}\right)$$

Where b = count of lethal genes in hub set, n = total count of genes in the hub set, B = count of lethal genes in background set, and N = total count of genes in the background set. P-value was determined by performing hyper-geometric distribution on the enrichment scores.

C. Spanning Tree Chordal Method

Our method for identifying a spanning tree with chordal properties is described below in 3.1 and 3.2. The proposed algorithm, Spanning Tree Chordal (STC), adds edges to a constructed spanning tree to obtain a chordal subgraph of the network. The final filtered network depends on the order and frequency of edge selection and whether edge weights (strength

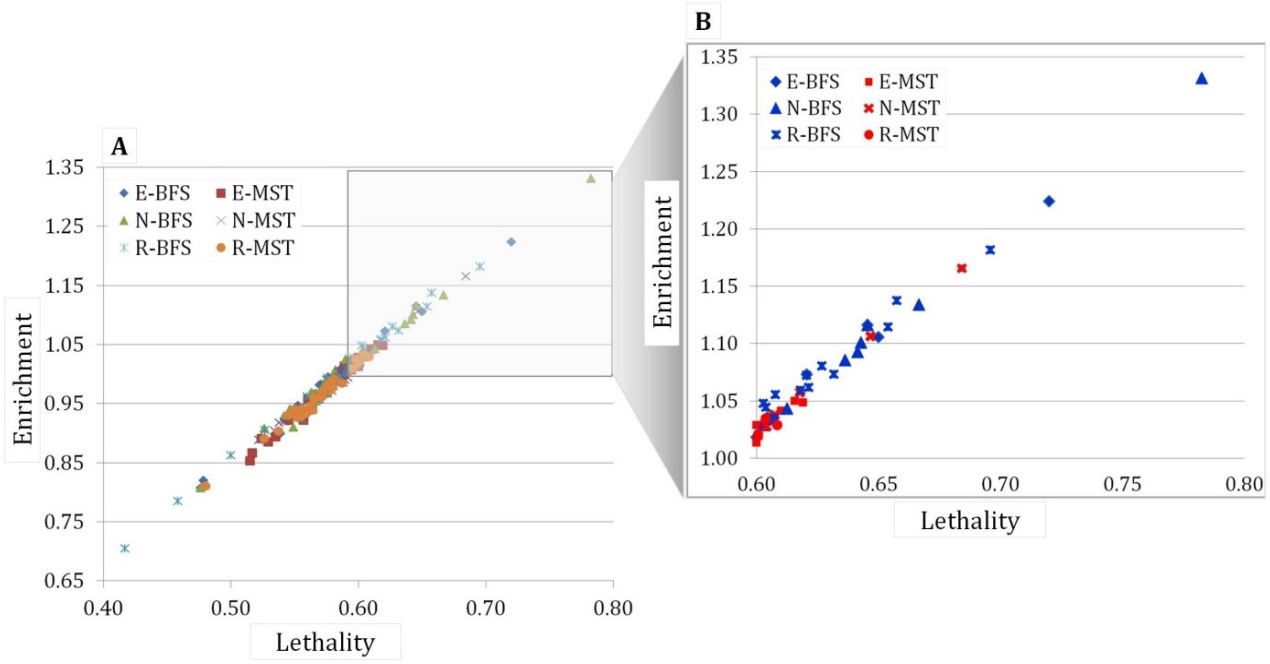


Figure 3: (A) Lethality vs. Enrichment for all filtered networks comparing BFS to MST . (B) Lethality vs. Enrichment for filtered networks above the optimal value of 56% lethality and 1.00 enrichment comparing BFS to MST .

of the relationship between genes) are included in the network.

C.1) Algorithm and parameters: All of the algorithms for network creation begin by computing a spanning tree of the network. We experimented with two types of spanning trees: a breadth-first-search tree (BFS), which is computable in linear time. The second is a maximum spanning tree (MST), which is more expensive to compute but constructs the spanning tree to maximize the sum of the edge weights and thus is more likely to maintain edges between highly correlated genes. We developed two algorithms for augmenting the spanning tree: the first (SL for "sibling line") roots the spanning tree at a node, then looks at children of each node and adds a subset of the edges between those children. The edges are chosen so that no subgraph induced by the children of a node contains a cycle; this condition is sufficient to ensure that the overall network maintains chordality. The second method (SA for "sibling all") adds every edge between siblings of a node to maintain the highly connected nature of hubs and capture dense subgraphs. However, perfect chordality may no longer be maintained.

Our algorithm is described further visually in Figure 2: Let the network (C) be the network after thresholding. The dashed edges in (D) indicate the maximum spanning tree (MST); i.e. the set of edges that connects every vertex without creating a cycle and maximizes the sum of the edge weights in the tree. This is denoted MST-NO. The highlighted edges in (E) are edges added through the SL algorithm. The MST has been rooted at vertex 1, which means vertices 2, 4, and 5 are now siblings. The algorithm checks for an edges between 2 and 4, then 4 and 5 (but does not check for an edge between 2 and 5 as it might create a chordless cycle). Since only edge 4 to 5 is present in the initial network, only that edge is added. Similarly, vertices 3 and 6 are children of vertex 4, and the edge between them is added. This is the MST-SL network. Finally, the network in (F) is the MST-SA network. Now, when looking at the vertices 2, 4, and 5, every edge that exists between those three vertices is added. In particular, both the edge from 4 to 5 and the edge from 2 to 5 are present.

C.2) Edge weights: Three types of edge weights were used in our analysis. The Normal edge type (N) was defined as used where each edge was set to the weight of the correlation in the original network. Random edges (R) denote randomly chosen edge weights between -1.00 and 1.00. The Equal edge type (E) denotes when each edge was set to an equal value (in this case, that value was 1.00). The normal edge type is expected to be the best performer as it includes the weight of expression correlation, a biological bias, where the E and R edge weights are meant to act more as controls for what we can expect from a network not biased by prior information. For the equal weights option the MST and BFS have the

same objective—filtering a spanning tree of the network. The difference lies in the algorithm. The BFS traverses through connected components while the MST maintains a forest of trees that are ultimately connected, as per Kruskal's algorithm. The difference in the approaches results in different trees even with the equal weights on the edges.

Using the three algorithm parameters and the three types of edge weights resulted in a total of six networks to analyze per YNG and MID dataset, or 36 (2 networks * 3 filters * 3 weights) networks total in addition to the 6 ORIG networks. The ORIG set includes the YNG network without any filtering of edges, as well as our original filters using the chordal subgraph and maximum spanning trees. In the next section, we test the following hypothesis:

H_0: Given a graph that represents gene expression correlation, the proposed algorithm STC produces a filtered network that conserves all hub nodes representing lethal genes in the network. Moreover, STC uncovers new lethal hub nodes previously hidden in the network, and outperforms both spanning tree filters and chordal subgraph filters in terms of identification of new lethal hubs.

III. EXPERIMENTAL RESULTS

We performed analysis of all 42 networks in terms of structural and biological properties.

A. Lethality Results

We plot lethality versus enrichment in one analysis and use this to determine the success or failure of a filter. Any network whose top X% degree target set scores over 54.12% lethality of the original network and 1.00 enrichment is a positive result. We find that for all results, the hub thresholds of 10, 15, 20, and 25% perform best among all contenders.

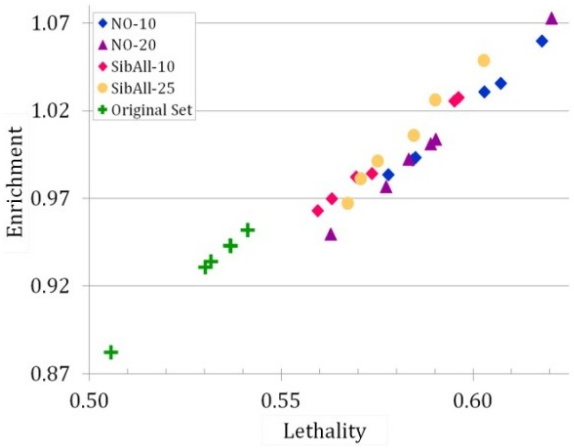


Figure 4: Comparison of optimal filters vs. ORIG networks.

To compare BFS vs. MST, we group data according to BFS or MST to determine the optimal parameter (Figure 3) and find that the majority of our "positive results" are found in the BFS parameter at any threshold. Comparing NO vs. SL. vs. SA at the optimal threshold values of 10-25% yields a slightly better performance of NO and SA over SL. For NO and SA parameters at optimal threshold values using BFS selection, we find that NO performs best at parameters 10 and 20% and SA performs best at parameters 10 and 25, further shown in Figure 4, where ORIG networks are compared to filtered with a visible gap in lethality and enrichment. Evaluation of all networks resulted in 222 experiments, with six of those being evaluations of the original networks. Out of the remaining 216 experiments, 73 of those performed well enough to be qualified in the "positive result" region, with 23 of the 73 falling within the optimal degree threshold and node choice method. We summarize these 23 optimal experimental results in Figure 5, and find that the distribution of these results are almost equally distributed between the None and Sibling All chordal selection methods. Of these two "optimal" parameter/threshold choices, there appeared to be no advantage over others in terms of edge weight choice.

IV. DISCUSSION

We have proposed a new graph theoretic methodology to enhance the biological signal from a noisy network model using chordal and spanning tree filter application. This method enhances networks by removing noise from correlation networks and making essential hub nodes more readily identifiable. We tested the impact of degree threshold, node selection method, and chordal network identification within the modular portions of the network. Per the results, we have found that BFS method outperforms the MST node selection, and thresholds of 10- 25% are optimal for stable definition of hub nodes in the filtered networks. Finally, we find that while all filtering methods perform as well as or better than the original networks, the NO and SibAll filters are the best for identifying lethal hub nodes. This suggests that structural filters, specifically filters based on combined graph theoretic techniques, can significantly reduce network size while maintaining and even enhancing biological signal. Returning to our original hypothesis, we have shown that our method is able to identify lethal hub nodes as well as, or better than, previously studied. This work constitutes an important step towards the construction

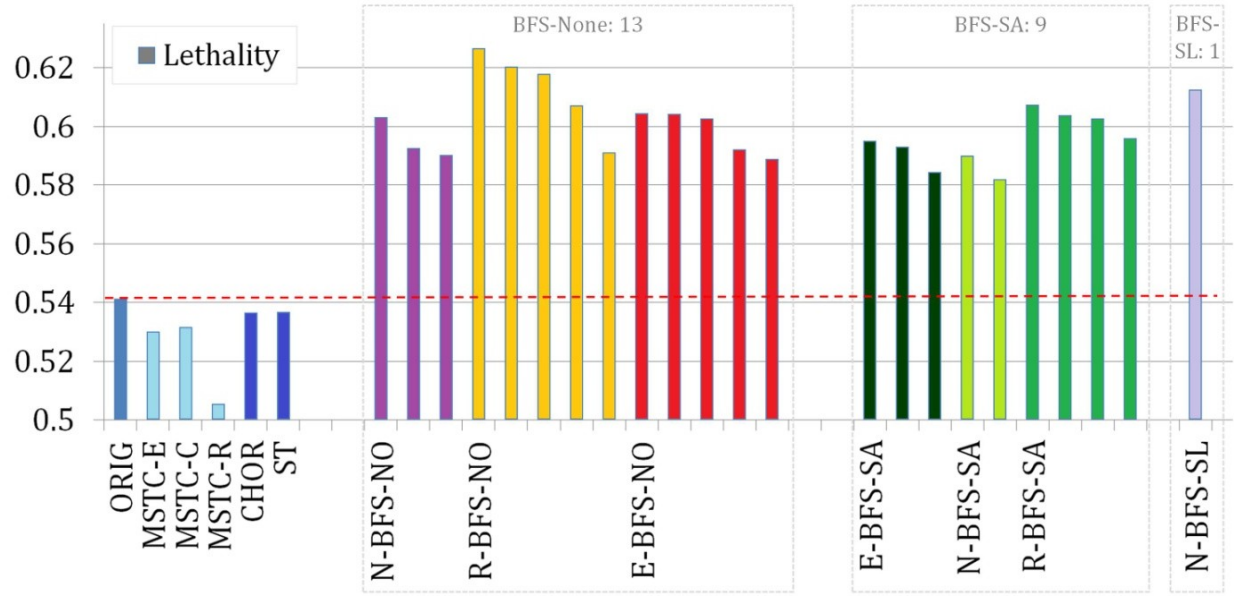


Figure 5: Lethality distribution for the optimal performance experiments compared to ORIG networks (ORIG, MSTC-E|C|R, CHOR, and ST). The red dotted line indicates the baseline expected lethality for the enhanced networks, which each experiment exceeds by at least 0.04.

of advanced filters for the purpose of analyzing large-scale biological networks.

ACKNOWLEDGEMENT

The authors acknowledge the NIH grants number P20 RR16469 from the INBRE Program of the National Center for Research Resources.

REFERENCES

[1] Barabasi, AL, & Oltvai, ZN (2004). Network biology: Understanding the cell's functional organization. Nature Reviews.Genetics, 5(2), 101-113.

[2] Duraisamy, K, Dempsey, K, Ali, H, and Bhowmick, S. (2011). A noise reducing sampling approach for uncovering critical properties in large scale biological networks (2011). *High Performance Computing and Simulation 2011 International Conference (HPCS):* July 4-8. Istanbul, Turkey.

[3] Dempsey K, Duraisamy, K, Ali, H, and Bhowmick S. (2011). A parallel graph sampling algorithm for analyzing gene correlation networks (2011). *International Conference on Computational Science 2011*. June 1-3. Singapore.

[4] Dempsey K,, Duraisamy, K, Bhowmick S, and H Ali. (2012). The development of parallel adaptive sampling algorithms for analyzing biological networks. *11th IEEE International Workshop on High Performance Computational Biology (HiCOMB 2012)*. May 21, 2012. Shanghai, China.

[5] Dearing, PM, Shier, DR, and Warner, DD. (1988). Maximal Chordal Subgraphs. Discrete Applied Mathematics 20(3):181-190.

[6] Edgar, R, Domrachev, M, & AE Lash (2002). Gene Expression Omnibus: NCBI gene expression & hybridization array data repository. Nuc Acid Res 30(1):207-10.

[7] Dong, J, & Horvath, S. (2007). Understanding network concepts in modules. BMC Systems Biology, 1, 24.

[8] Reverter, A, & Chan, EK. (2008). Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks. Bioinformatics (Oxford, England), 24(21), 2491-2497.

[9] Watson-Haigh, NS, Kadarmideen, HN, & Reverter, A. (2010). PCIT: An R package for weighted gene co-expression networks based on partial correlation and information theory approaches. Bioinformatics (Oxford, England), 26(3), 411-413.

[10] Ewens, WJ, & Grant, GR. (2005). Statistical methods in bioinformatics (Second Edition ed.). New York, NY: Springer.

[11] Zhang, B, & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology, 4, Article17.

[12] Carter, SL, Brechbuhler, CM, Griffin, M, & Bond, AT. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics (Oxford, England), 20(14), 2242-2250.

[13] Verbitsky, M, Yonan, AL, Malleret, G, Kandel, ER, Gilliam, T C, & Pavlidis, P. (2004). Altered hippocampal transcript profile accompanies an age-related spatial memory deficit in mice. Learning & Memory (Cold Spring Harbor, N.Y.), 11(3), 253-260.

[14] Dempsey K, and H. Ali. (2012. On the Discovery of Cellular Subsystems in Correlation Networks using Centrality Measures. *Current Bioinformatics*. Publication in Summer 2012.