2011

A Novel Correlation Networks Approach for the Identification of Gene Targets

Kathryn Dempsey Cooper
University of Nebraska at Omaha, kdempsey@unomaha.edu

Stephen Bonasera
University of Nebraska at Omaha

Dhundy Raj Bastola
University of Nebraska at Omaha, dkbastola@unomaha.edu

Hesham Ali
University of Nebraska at Omaha, hali@unomaha.edu

Follow this and additional works at: https://digitalcommons.unomaha.edu/interdiscipinformaticsfacproc

Part of the Bioinformatics Commons, Genetics and Genomics Commons, Metaphysics Commons, and the Numerical Analysis and Scientific Computing Commons

Please take our feedback survey at: https://unomaha.az1.qualtrics.com/jfe/form/SV_8cchtFmpDyGfBLE

A Novel Correlation Networks Approach for the Identification of Gene Targets

Kathryn Dempsey, Stephen Bonasera, Dhundy Bastola and Hesham Ali

College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE 68182
Department of Pathology and Microbiology, University of Nebraska Medical Center, Omaha, NE
68198 *hali@mail.unomaha.edu*

Abstract

Correlation networks are emerging as a powerful tool for modeling temporal mechanisms within the cell. Particularly useful in examining co-expression within microarray data, studies have determined that correlation networks follow a power law degree distribution and thus manifest properties such as the existence of "hub" nodes and semi-cliques that potentially correspond to critical cellular structures. Difficulty lies in filtering coincidental relationships from causative structures in these large, noise-heavy networks. As such, computational expenses and algorithm availability limit accurate comparison, making it difficult to identify changes between networks. In this vein, we present our work identifying temporal relationships from microarray data obtained from mice in three stages of life. We examine the characteristics of mouse networks, including correlation and node degree distributions. Further, we identify high degree nodes ("hubs") within networks and define their essentiality. Finally, we associate Gene Ontology annotations to network structures to deduce relationships between structure and cellular functions.

1. Introduction

The advent of high-throughput "omics" technologies has allowed for surveillance of cellular mechanisms on the genome-wide scale. With the development of these approaches comes the need for thorough systems biology analysis. In 2006, Bruggeman *et al.* highlighted three main requirements for a characterizing a biological system that can be applied on the cellular level. These steps include developing a knowledge-base of all interactions and interactors within the system, the mechanisms of those interactions in various experimental conditions, and responses to both internal and external stimuli. This complete analysis combined with appropriate system modeling will allow for the prediction of cellular response, the identification of unknown regulatory mechanisms, and eventually the ability to guide the treatment of cellular system to a desired response [3]. While the technology required to orchestrate and model such a thorough analysis may not yet exist in entirety, the study of systems biology continues to mature and realize novel insights about cellular systems.

There are many approaches to systems biology; indeed, the integration of network models in the creation and analysis of biological relationships has become a powerful took for representing spatio-temporal changes on a whole-genome scale. In particular, the correlation network stands out as a tool for measuring linear relationships such as gene expression and protein concentration. Threshold-filtered correlation networks created from gene expression data have been found in *S. cerevisiae* [6, 13], *A. thaliana* [10, 11], *D. melanogaster* [14], and *M. musculus* [7]. All of these networks fall within a broad network structure described as a 'scale-free' network topology that has certain defining characteristics such as adherences to a power-law node degree distribution, a lower than average clustering co-efficient, and lower than average path-length [2, 3]. This scale-free definition has been found to apply not only to other biological networks (interactome, metabolome) but also in a wide range of applications such as social networks and citation databases [1].

1.1. Network Construction

In a network drawn from microarray expression data, genes are represented as nodes, with an edge drawn between two nodes if some relationship exists between them. One of the most straightforward relationships to identify between two genes is correlation of linear expression, commonly measured by Pearson correlation. The resulting network is a graph of all genes and their correlations to all other genes in the dataset; in graph theory this is known as the complete or K_n network, where n is equal to the number of genes. The nodes and edges in this

network must then be weighted or directed for the user to be able to discern critical patterns; this can be done by imposing weights on the edges (where the weight is determined by the level of correlation between two genes) or by removing edges outside of some threshold t as shown in Figure 1. Either method is suitable for the correlation network in theory, but for large networks representing genome-wide transcripts, the method of removing edges creates a more computationally manageable network. For example, in a complete graph K_n, the number of edges is known to be equal to $^{n*(n-1)}/_2$. The K_{1000} graph will contain some 499,500 edges, a number that is relatively manageable by current analysis and visualization methods (but likely requiring the availability of multiple processing cores). When one investigates the entire set of genes for an organism, volume becomes a problem once edges are created. For example, one microarray analysis for a BalbC mouse contains over 41,000 gene probes resulting in the creation of a complete network with ~840 million edges. While construction and storage of this network is relatively simple, the analytic complexity far exceeds current computational resources for laboratories without access to supercomputing resources. Thus, for large networks, it is prudent to use the method of edge removal over edge weighting for network management.

When looking at the correlation distribution of the complete K_n network, previous studies have shown that correlations in the complete network tend toward a normal distribution, with the majority of edges having an undiscernable or random expression correlation around 0.00, and distribution of correlations becomes increasingly smaller as the extreme correlations (±1.00) are approached. With the non-linear distribution of correlations in a complete network, the threshold filtering removal of edges cuts the network down considerably.

1.2. Hubs, Clusters and Pathways

When examining the filtered correlation network, one of the most interesting and readily identifiable structures are the network "hubs," or nodes with a high degree (number of edges) compared to the rest of the network. This is apparent when examining the degree distribution in networks that follow a power law distribution, meaning that there are many nodes that are poorly connected and a few nodes that are very well connected [1]. Studies have shown that nodes identified as hubs are critical for network structure, and their removal results in the breakdown of passage of signals and network robustness [2]. In protein-protein interaction networks, the hubs noted as proteins are encoded by genes that are known as

'essential'. When such genes are knocked-out or knocked-down, lethality of the organism results [8, 12]. Similar observations have been reported with gene correlation networks in 2010 by Mutwil *et al.*,
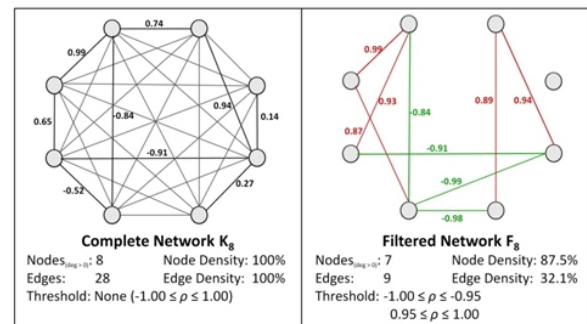

Figure 1. Correlations in the complete network (left) have been removed for clarity. Green edges in the filtered network (right) correspond to negatively correlated nodes, and red edges correspond to positively correlated nodes.

where 20 essential hub genes were identified in *A. thaliana* and the resulting tDNA knock-out mutation of those genes resulted in a lethal phenotype in 5 cases, and a reduction of major system functions (size, coloring) in one case.

1.3. Proposed Approach

In this work we provide a pipeline for network characterization with a proof of concept using temporal microarray expression data from the aging mouse. We highlight four steps for identifying some critical structures within large networks (>40,000 nodes) that have a high likelihood for corresponding to real biological function. These steps, shown in Figure 2, correspond to increasingly complex processes that are currently being investigated and applied in systems biology research. The first filtering step involves examining size and density of large networks to determine an appropriate threshold for reducing the size of the network. Second, we suggest a characterization step to verify that the resulting filtered networks indeed adhere to a power law distribution and are sized appropriately for current visualization and analysis programs. Third, we suggest a sweep of the network to identify basic structures that are known to correspond with potentially critical genes and gene modules. Finally, we address the need for integration of graph theory and current systems biology techniques to filter noise from causative structures, and align those structures under different spatio-temporal conditions.

To illustrate this pipeline we propose a proof of concept using hypothalamic gene expression data from mice from three different age groups: young, middle-aged, and aged. Through the application of
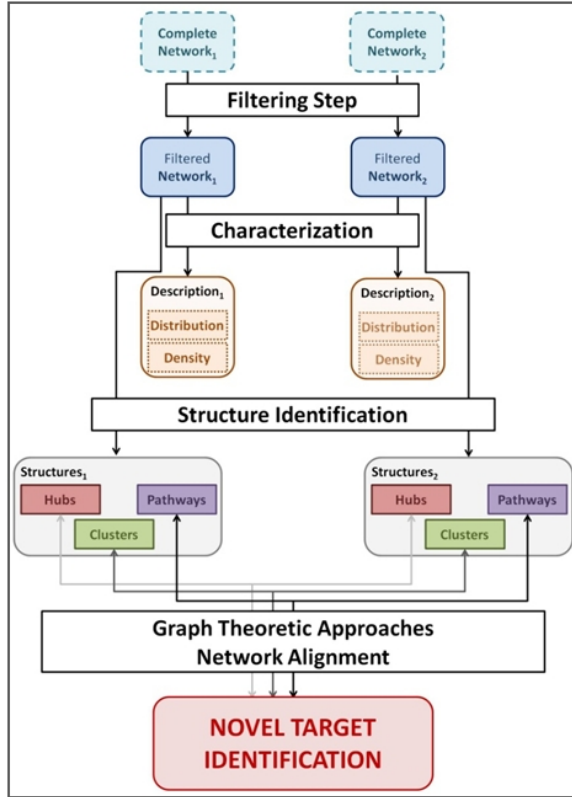
Figure 2. Pipeline for identifying important network structures with real biological function within correlation networks

this pipeline we are able to highlight the results for each step and propose a putative target gene list for further study in aging based solely on network analysis. The datasets used in this study are highlighted in Table 1 (right). We created and filtered complete correlation networks for each dataset by performing a Pearson correlation between each possible pair of gene expression vectors in the dataset. We then decide upon a network correlation threshold of ±0.95-1.00 by examining the correlation distribution for a random 10,000 node sample from each network. By filtering networks to solely the very correlated and very anti-correlated edges, we observe a power-law node degree distribution typical of the scale-free network [1]. We then rank nodes according to their degree, and are then able to identify the top 'hub' nodes most likely to be involved in essential interactions for that particular temporal state. Finally, we use the publicly available network analysis software NetworkBLAST to identify clusters of nodes and provide an example of a cluster that exhibits a common function. The availability of temporal whole-genome expression data allows us to exploit and observe the evolution of hub nodes from the young to the middle aged to the aged mouse.

1.4. Correlation vs. causation

It is important to note that correlation networks can reveal direct correlations between two genes or interactors, but indirect interactions can be and are often lost [11]. For example, in a regulation cascade, the first gene in the pathway may be expressed at low levels, but that gene's products may go on to regulate transcription of a second gene which will be expressed at exponentially higher levels or the third gene that may be expressed at a exponentially lower levels. These relationships are unlikely to be captured by the application of the correlation network, and other means are needed to measure these non-linear relationships (such as the Spearman Rank correlation).

Table 1. Age and type of hypothalamus gene expression data used for male mice

Dataset	Age	Type	Probes
MB_Young	2-3 months	Balb/C	41,174
MB_Midage	12-13 months	Balb/C	41,174
MB_Aged	13-24 months	Balb/C	41,174
MC_Young	2-3 months	CBA	43,675
MC_Midage	12-13 months	CBA	43,675
MC_Aged	13-24 months	CBA	43,675

2. Methods

2.1. Data & Pearson correlation

Complete microarray expression data for 6 sets of mice at various stages in their lives, denoted as young, middle aged, and aged mice (see Table 1) was obtained from the Bonasera lab. Each gene in each dataset had at least $n = 5$ sample expression values; gene comparisons with uneven sample numbers or $n < 5$ samples were thrown out. Each microarray is represented in our network creation pipeline as a set of gene expression vector objects, $gene_i$, where $gene_i$ contains a gene identifier id_i and a set of expression values for each sample exp_i, exp_{i+1}, and so on. The Pearson correlation coefficient was then determined for each gene expression vector versus all other gene expression vectors in the dataset, using Equation 1:

$$\rho(x,y) = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i-1}^{m}x_i^2\ \sum_{i-1}^{m}y_i^2}}, \qquad (1)$$

Where x = the set of expression values for $gene_x$ and y = the set of expression values for $gene_y$.
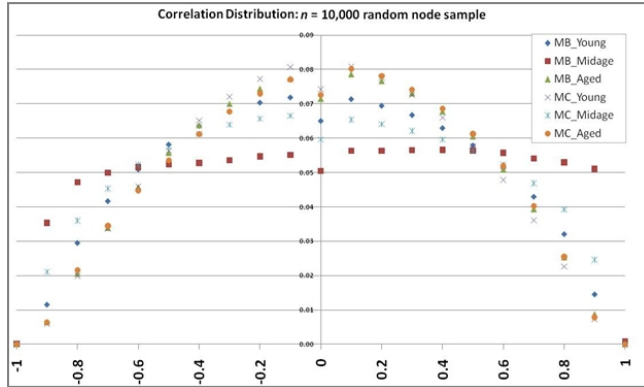
Figure 4. Correlation distribution for random 10,000 node samples per dataset

2.2. Network creation, threshold filtering, and densities

For each network, we took a random 10,000 node sample and created a complete network, $K_{\sim 41k}$, from each to determine their respective correlation distribution. Based on the distribution we chose to examine only those extremely correlated values where $-1.00 \le \rho \le -0.95$ and $0.95 \le \rho \le 1.00$ to create manageable networks containing very tightly correlated genes. This pairwise correlation was calculated for all genes under the previously specified positive ($0.95 \le \rho \le 1.00$) and negative ($-1.00 \le \rho \le -0.95$) thresholds to create filtered networks. We note that not all important edges will be captured by this choice of threshold and likewise that not all edges captured will be important; indeed the filtering of noise from causative relationships remains an issue with the correlation network model.

2.3. Structure identification

One of the most easily recognizable characteristics about a network is the node degree distribution. At this point there is no true thresholding method for differentiating between hub and non-hub nodes. A study in network robustness performed by Giaver *et al.* in 2002 determined that the disruption of 73% of randomly chosen genes in the yeast network does not cause organism lethality (by means of network structural collapse), however a targeted attack of hub nodes will likely result in network failure [6]. We confirmed the adherence of the networks to a power law node degree

distribution and identified hub nodes for each network by ranking nodes in descending order by degree. The top 20% of nodes per network (8,236 nodes in Balb/c; 8,735 nodes in CBA) were labeled as "hub" nodes, with the 20% arbitrarily chosen (there is no current definition or threshold for what distinguishes hub versus non-hub nodes). In addition, the essentiality of the top 20 hubs was examined using publicly available databases. The theory of centrality-essentiality states that hub nodes have a higher likelihood to be lethal when disrupted, highlighting the vulnerability of the scale-free network. The disruption of a non-hub node has less potential for affecting organism survival, but an intelligent attack, or disruption of a hub, could potentially cause the death of the cell and in turn the organism if it is a critical point in development [8, 12]. To test the essentiality of hubs, we integrated data from the Mouse Genome Informatics (MGI) database, and defined gene essentiality by the following criteria: (1) the gene had been tested by an *in vivo* knockout and (2) the resulting phenotype from knockout was lethal or severely affected a major body system (growth, reproduction, etc). Lethal or system-affecting phenotypes resulting from mutations that could not be directly attributed to a disruption *only* in the gene of interest were not considered essential or system-affecting.

Finally, we considered the clusters of nodes in the network for their possibility in corresponding to gene modules. It is possible for complete or "almost complete" K_n networks to exist within large correlation network, and it has been proposed that these networks can correspond to a set of genes working toward some function, for example, in a regulatory manner or as a protein complex. In graph theory, these complete or almost complete subnetworks are be referred to as cliques and semi-
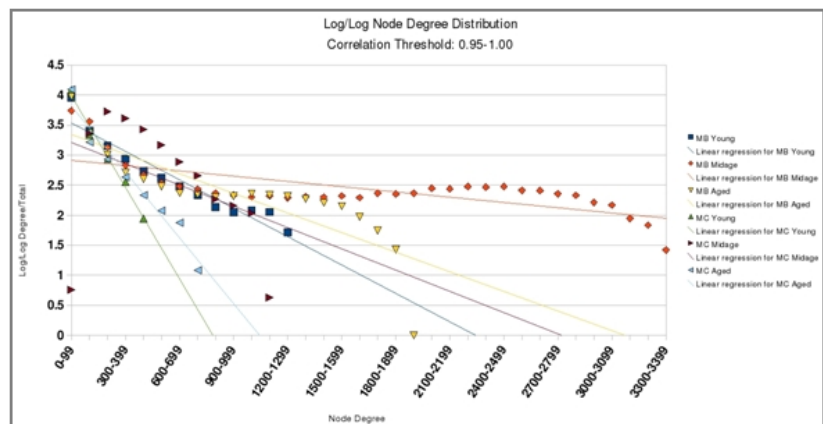

Figure 3. Log/log node degree distribution for each dataset. Degree counts are grouped for clarity. Linear regression lines are included. Node degree is grouped for clarity; however, note that this skews the linear nature of the distribution in log/log format.

cliques, and many algorithms exist to find them within larger graph models. However, the size of correlation network produced by our initial filter was still too large for current methods to handle. We chose, then, to filter our networks to only edges with positive correlations of value 1, producing networks with only a few thousand edges. Then we executed the NetworkBLAST software developed by Kalaev *et al.* in 2009 [9] under default parameters for one species to identify top clusters within our twice-filtered network. The top resulting clusters were analyzed by Gene Ontology term enrichment to determine if any functional annotation was common among nodes [5].

3. Results

The following results are for correlation networks created from hypothalamic gene expression data from two types of male mice, Balb\C and CBA, at three stages of life: young, middle-aged, and aged.

3.1. Correlation Distribution

We created complete $K_{10,000}$ networks for each dataset using a random 10,000 node sample and we present the resulting correlation distribution in Figure 3. In all networks described, the majority of correlations fall at or near 0.00, indicating there is little to no correlation between linear expression patterns. Thus we present that for the size of network, choosing a tight threshold for correlation at the very extreme values will allow for the creation of a manageable complete network.

3.2. Power-law degree distribution

Figure 4 describes the log/log node degree distribution for all datasets and indeed follows a power-law distribution (which is a straight line in log/log form). The hub nodes are identified as the top 20% of nodes in a ranked list according to degree. One point worth mentioning observed from Figure 4

Table 2. Essentiality descriptions for known in vivo mutations of hub nodes

Dataset	*in vivo* KO/KD	Essential	System-affecting	No Phenotype
MB Young	6	0.50	0.33	0.17
MB Midage	5	0.40	0.60	0.20
MB Aged	3	0.67	0.33	0.00
MC Young	1	0.00	1.00	0.00
MC Midage	7	0.43	0.43	0.14
MC Aged	7	0.14	0.86	0.00

is that the maximum node degree rises and then falls again from young to middle-aged to aged mice. Until further testing is performed an explanation is speculation at best, but this could be due to compensatory efforts of cells as mechanisms die out over time. The cell is designed to maintain a state of homeostasis and will attempt to return to a state of homeostasis after perturbation. It is known that over time with the accumulation of damages to nuclear DNA that some mechanisms are bound to become irreversibly damaged; evolution has designed a robust system such that these disruptions will not cause the death of the organism. The cell is designed to adapt and compensate for these losses – as such, if one gene were to become mutated, the cell might signal to other genes to compensate for the loss. This is one possible explanation for the above phenomenon that may be of interest in future studies.

Figure 5. Hub genes per dataset identified in the MGI database as having an *in vivo* knock-down/out/in mutation and the resulting phenotype.

3.3. Centrality-essentiality

The centrality-essentiality concept states that those genes identified as hubs within a network are more likely to be known as essential genes [8]. To verify this hypothesis within our data, we identified the top 20 hubs for each datasets and turned to the Mouse Genome Informatics database (MGI). For each hub node (gene) we identified two factors from the MGI database: (1) Has a knock-out, knock-down, or knock-in mutation had been performed for that gene *in vivo*? and (2) if the mutation had been
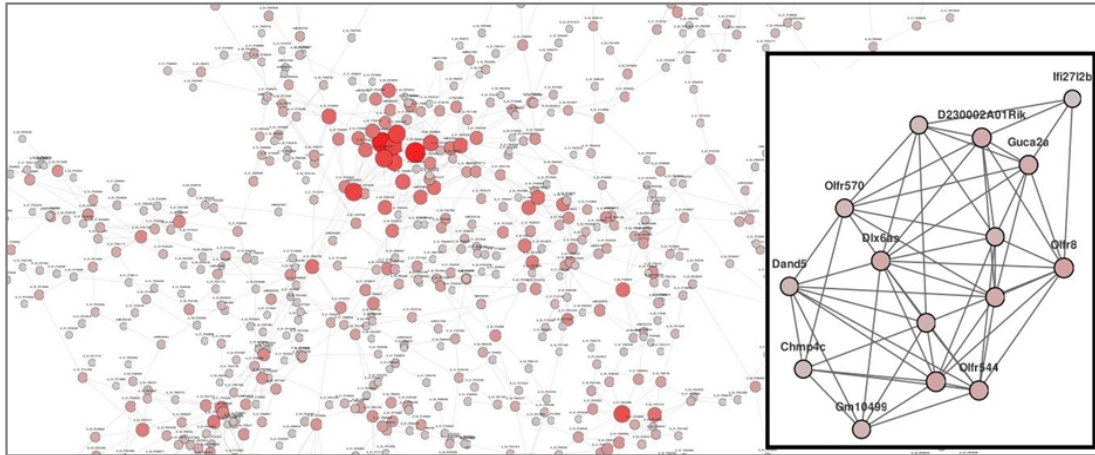
Figure 6. A portion of the MC Young correlation network at 1.00 only filter, where node size and color correspond to degree. Insert: The top cluster identified from the MC Young network (right, inset) with nodes annotated with Entrez Gene names where applicable.

performed, is it detrimental to some system or result in lethality of the mouse pre- to post-natally? Table 2 defines how many genes had been mutated *in vivo* and classifies those mutations as "essential" (pre-postnatal lethality), "system-affecting" (detrimental to a major system), or "no phenotype" (no observable phenotype observed with gene mutation). We also present the genes identified in Table 2 with their gene name, description, and characterization (Figure 5). The system descriptions accompanying essential mutations are as they are described in the MGI database. This reaffirms the idea that hub nodes are likely to be essential for life or for "normal" system functions, this also provides a list of targets for future experimental validation. In addition, the remainder of untested genes are candidates for essential genes in the aging process.

3.4. Cluster analysis and Gene Ontology enrichment

For each network we filtered the top 20% of nodes from the original network to create a smaller subnetwork (8,236 nodes in MB and 8,735 nodes in MC) and applied a correlation filter to only see those genes that had a correlation near 1.00 to hasten runtime (as the version used had not been implemented in parallel). In the MC Young dataset we identified a cluster (or semi-clique) of genes with 15 nodes and 60 edges existing out of a possible 105. This cluster had 10 nodes with known Gene Ontology annotation, and of those, four were significantly related by ontology. Those were nodes for Olfr570, Olfr544, Dand5, and Olfr8, which play roles in the following common functionalities based on GO annotation: cell surface receptor linked signaling pathway, signal transduction, sensory

perception, cognition, signal transmission, signaling process, and G-protein coupled receptor protein signaling pathway.

One method for filtering noise from causation in a network is to use Gene Ontology enrichment. A study in *D. melanogaster* in 2006 by Xia *et al.* provided a list of Gene Ontology annotations grouped according to their general cellular function (proliferation, differentiation, proteolysis, and immunity) that were found to be enriched when studying differential expression of clusters in the proteome. To determine if this differential expression occurred within our networks, we annotated nodes in the MB Young and MB Aged networks with all known Gene Ontology (GO) annotations, downloaded in October 2009 from http://www.geneontology.org. We then colored the nodes according to their GO classification (proteolysis, differentiation, proliferation, or immunity) and examined the results (Figure 7).

Young Network	
# Nodes	23,156
# Edges	70,060

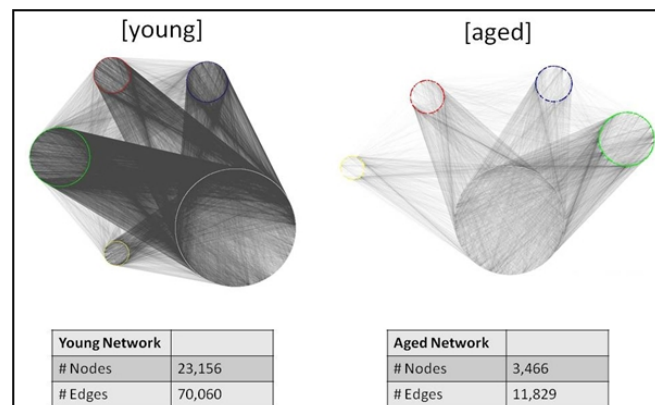| Aged Network	
# Nodes	3,466
# Edges	11,829

Figure 7. Young and aged BalbC networks filtered at 0.99 correlation and by Gene Ontology for differentiation (red), proliferation (green), proteolysis (blue), immunity (yellow), and unknown genes (gray).
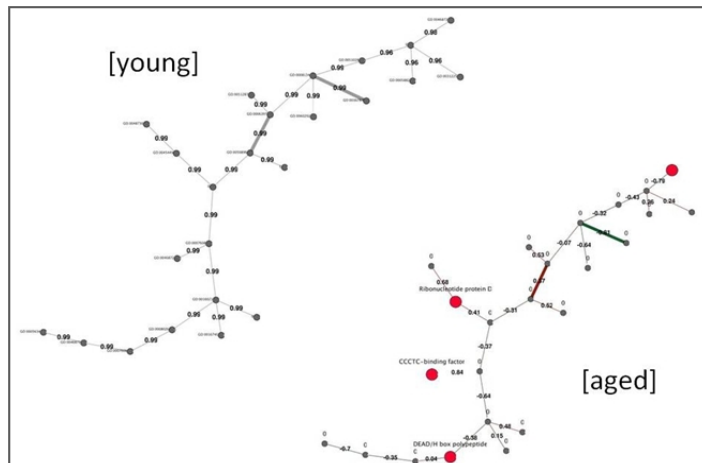
Figure 8. Subnetwork found in the young mouse network and aged mouse networks (BalbC) with differing correlations, indicating that if this were a structure that corresponded to true biological function, this structure would be a potential target of interest.

What we observe is that while we know the overall edge density of the original filtered aged mouse network to be relatively close to the edge density of the original filtered young network, the young mouse network is actually more enriched in the GO annotations listed, and thus denser. This could further reinforce the idea of compensation – while some interactions are strong and tightly correlated in the young mouse, as disruptions and mutations occur these signals must find other avenues or stop altogether, resulting in the loss of the tight gene correlation seen in the young.

To ensure that this apparent loss of edges was not occurring due to thresholding (i.e. genes were highly correlated around 0.90 but missed by the 0.95-1.00 threshold) we examined subgraphs from the MB Young and MB Aged networks to determine what was happening. For the MB Young subgraph in Figure 8, all of the edges exist with a correlation value of 0.99. The correlations in the MB Aged subgraph with the same nodes have correlations occurring across all values possible from -1.00 to 1.00, suggesting that the tight correlations observed in the MB Young subnetwork were coincidental and therefore are likely noise. Further investigation revealed that this structure had little common Gene Ontology enrichment. This subnetwork is one of many from the original filtered network, but highlights the high "noise" component of correlation network analysis and the need for further tools to discern coincidental relationships from true biological modules.

4. Discussion

We have provided a proof of concept and approach for examining temporal gene expression data using the correlation network. It was observed that correlations follow a distribution that resembles a standard normal distribution, and as such a threshold for large networks can be chosen according to this distribution. It should be emphasized that for large networks (>40,000 nodes) that are difficult to visualize these types of characterizations and analyses are needed to know what the network "looks" like, and to ensure that the network has been constructed and filtered appropriately. We have verified the scale-free nature of the filtered gene correlation network over several datasets and used that distribution to identify resulting hub structures within each network, in addition to the essentiality of those hub structures where applicable. Finally, we present an example of how one might use current network analysis tools on a smaller network to identify modular structures that potentially correspond to true biological complexes. This work only briefly touches on graph theoretic approaches that can be employed to identify causative structures within the noisy correlation network; indeed there are a number of more complex analyses that can be performed on a network, including graph matching, network alignment, pathway searches, etc.. Since the field of interaction networks is relatively new, there is opportunity for novel methodologies and discoveries to be made for a variety of types of interaction networks.

Acknowledgements

References

[1] Albert, R. Scale-free networks in cell biology (2005) Journal of Cell Science 118, 4947-4957.

[2] Albert, R. and Barabasi, A. Statistical mechanics of complex networks. (2002) Rev Modern Phys 74, 47-97.

[3] Barabasi, A. L., Oltvai, Z. N. Network biology: understanding the cell's functional organization (2004) Nat Rev Genet. 5 101-113.

[4] Bruggeman, F. J., Westerhoff, H. V. The nature of systems biology 2007.

[5] Eden, E., Navon, R., Steinfeld, I., Lipson, D. and Yakhini, Z. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists 2009.

[6] Giaver, G., Chu, A.M., Ni, L., Connelly, C. Riles, L. Veronneau, S. Dow, S. Lucau-Danila, A., Anderson, K, Andre, B. et al (2002). Fucntional profiling of the Saccharomyces cerevisae genome. Nature 418 387-391.

[7] Gill, R., Datta, S. and Datta, S. A statistical framework for differential network analysis from microarray data 2010.

[8] Jeong, H. Mason, SP, Barabasi AL, Oltvai, ZN. 2001. Lethality and centrality in protein networks. Nature 411, 41-42.

[9] Kalaev, M., Smoot, M., Ideker, T. and Sharan, R. NetworkBLAST: comparative analysis of protein networks 2008.

[10] Mutwil, M., Usadel, B., Schutte, M., Loraine, A., Ebenhoh, O. and Persson, S. Assembly of an interactive correlation network for the Arabidopsis genome using a novel heuristic clustering algorithm 2010.

[11] Opgen-Rhein, R., Strimmer, K. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data 2007.

[12] Reverter, A., Chan, E. K. Combining partial correlation and an information theory approach to the reversed engineering of gene co-expression networks 2008.

[13] Said, MR. Begley, TJ., Oppenheim AV, Lauffenburger DA, and Samson LD. 2004. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. Proc Natl Acad Sci USA 100, 1128-1133.

[14] Xia, K., Xue, H., Dong, D., et al. Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms 2006.