2011

Identifying modular function via edge annotation in gene correlation networks using Gene Ontology search

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

Ishwor Thapa
University of Nebraska at Omaha, ithapa@unomaha.edu

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

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

Identifying Modular Function via Edge Annotation in Gene Correlation Networks using Gene Ontology Search

Kathryn Dempsey, Ishwor Thapa, Dhundy Bastola, Hesham Ali
College of Information Science & Technology, University of Nebraska at Omaha
Department of Pathology & Microbiology, University of Nebraska Medical Center
hali@mail.unomaha.edu

***Abstract*-Correlation networks provide a powerful tool for analyzing large sets of biological information. This method of high-throughput data modeling has important implications in uncovering novel knowledge of cellular function. Previous studies on other types of network modeling (protein-protein interaction networks, metabolomes, etc.) have demonstrated the presence of relationships between network structures and organization of cellular function. Studies with correlation network further confirm the existence of such network structure and biological function relationship. However, correlation networks are typically noisy and the identified network structures, such as clusters, must be further investigated to verify actual cellular function. This is traditionally done using Gene Ontology enrichment of the genes in that cluster. In this study a novel method to identify common cluster functions in correlation networks is proposed, which uses annotations of edges as opposed to the traditional annotation of node analysis. The results obtained using proposed method reveals functional relationships in clusters not visible by the traditional approach.**

***Keywords*-Ontology, edge enrichment**

I. INTRODUCTION

Network models are gaining popularity as a tool for modeling large-scale biological data from a holist view: a network can represent the entire transcriptome or interactome of the cell at different environmental conditions. Thus, application of network theory can be a powerful model for identifying shifts in function between varying states, such as aging and disease. Specifically, it is advantageous to examine high throughput gene expression data using network model to identify relationships among groups of genes, which was traditionally done using statistical analyses such as Gene Set Enrichment Analysis that allowed identification of functionally related genes. In this network model, commonly known as a correlation network, nodes represent genes and edges represent the correlation, or strength of the relationship between expression patterns of two genes over multiple samples in the same environment. Models made in this way are particularly helpful in identifying temporal changes in the biological system.

Further, multiple types of biological networks have been found to contain specific network structures that reveal knowledge about critical functions occurring within the cell. For example, nodes in the network with the most connections have been found to correspond with essential genes, and clusters of well-connected nodes corresponded to cellular substructures such as protein complexes or regulatory cohorts (Barabasi 2004, Dong *et al*. 2007). The power of the correlation network model is palpable; however, implementation of this type of network analysis requires expertise with graph theory, high-performance computing, and intimate knowledge of the target domain, leaving the process of modeling somewhat tedious and overwhelming for individual or small laboratories with limited resources. Particularly, the current method uses Gene Ontology (GO) Enrichment to identify functions within the substructures of the network, which does not always reveal clear functions due to a variety of issues such as high levels of noise and database incompleteness. Further, Gene Ontology Enrichment is performed on the genes in the cluster. This does not tell us about the relationships between genes, only about the gene population of the cluster. For example, consider performing GO Enrichment on two subgraphs: set 1, with 10 nodes and 10 edges, and set 2, with the same 10 nodes but 35 edges. GO Enrichment on node sets of both subgraphs will return the same result, when these two subgraphs are inherently different in the *relationships* that exists between the genes in the cluster.

As such, it is critical to also examine the GO functionality of the edges in a network, just as it is critical to identify the GO functionality of the nodes. In this vein we propose that the process of identifying functional clusters can be improved and automated such that more confidence can be placed in cluster annotation. We present our method for annotating edges of a correlation network with Gene Ontology function where the parent-child nature of the GO tree is used in associating a score with each gene relationship (edge). Furthermore, we examine the node population versus the edge population of the networks and show that the edge population identified by our method reveals more significant intra-cluster functions than the node population. At the present state, this work is not comprehensive but serves as a proof of concept demonstrating the need to examine relationships in clusters of correlation networks, which has unfortunately been overlooked.

The method used in the current study takes two genes with a high correlation from a filtered correlation network (a

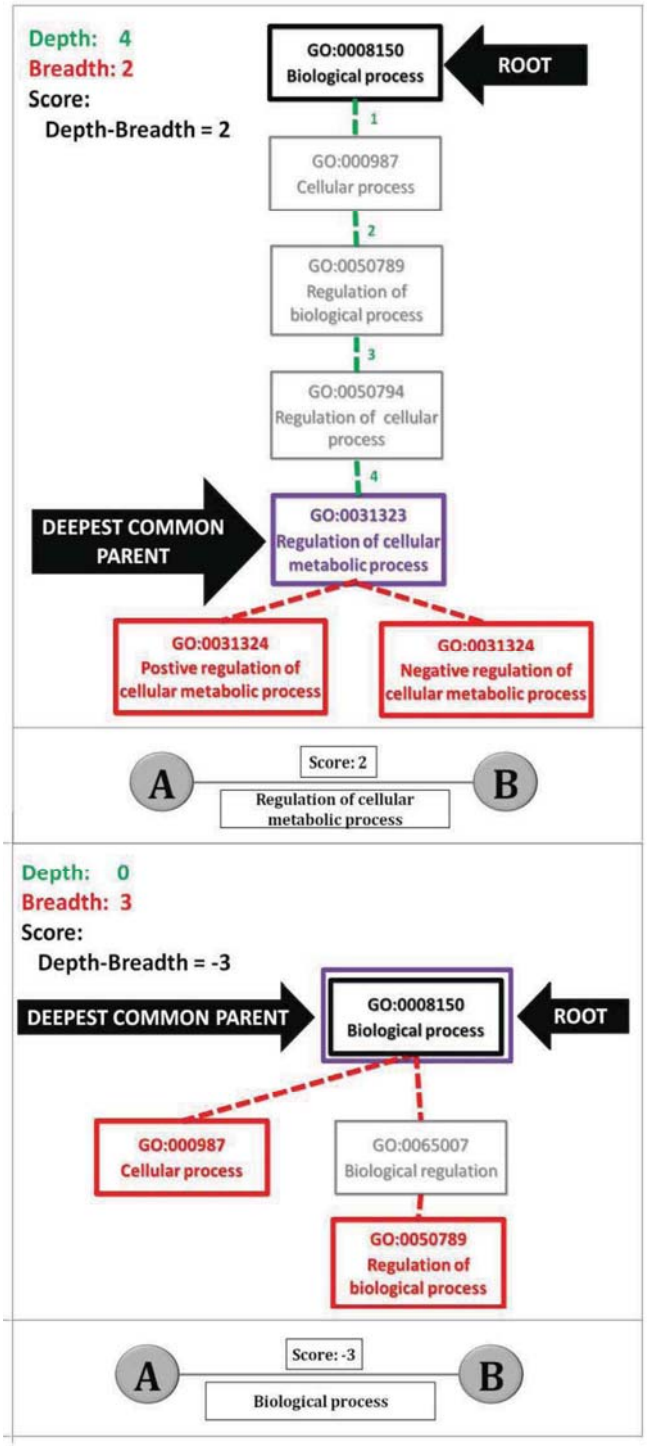

FIGURE 1. Our method of GO Scoring. Overall scheme for terms A and B (in red), the deepest common parent (DCP, in purple) and the root node (in green at top, Biological Process). Top: Depth of 4, breadth of 2 resulting in score of 2 (4-2=2). Bottom: Depth of 0 (root node is DCP), breadth of 3 for a score of -3 (0- ¯3=¯3)

method described by Dempsey *et al* 2011) and identifies any Gene Ontology (GO) terms associated with those genes, and additionally identifies the position of those terms in the tree structure of the Gene Ontology. There are three possible trees in the Gene Ontology: Biological Process, Molecular Function, and Cellular Component. Here we focus only on Biological Process solely to leave the process as uncomplicated as possible; however, this method can be applied to any of the trees. Once the terms are identified, we implement the algorithm for finding the *deepest common parent* of those terms within the network. The deeper the common parent of two genes, the better the score associated with that relationship. This score takes advantage of the increasing specificity of function with depth-based tree traversal in the GO; terms that are closer to the root have broader definitions of function, and terms that are deeper are more definite. Our method, explained in the Methods section, uncovers true function of relationships (edges) within correlation networks by application of tree-traversal and ontological enrichment of relationships instead of node population.

II. Previous Work

Many previous studies have focused on functional predictions of proteins or gene products using the Gene Ontology (Schwikowski *et al*., 2000, Hishigaki *et al*., 2001, Vazquez *et al*. 2003, Karaoz *et al*. 2004) but few methods have focused on the annotation of edges or relationships in clusters. More recently, Zhu 2007 proposed a method for functional prediction based upon GO's Biological Process tree that predicts specific functional classes for proteins with unknown function based on three measures, including gene co-expression data as additional evidence. Further, Deng *et al.* 2004 defined a method for predicting the probability of Gene Ontological function on a protein based on a Markov Random Field model. Cakmak *et al.* 2007 proposed a method for identifying novel functional metabolic pathways by exploiting the true path quality of the Gene Ontology (Gene Ontology Consortium, 2000). These methods, while effective, have not caught on as a standard for identifying cluster function in correlation networks, and further, few of these methods have been applied to the correlation network (and remain applicable for networks such as protein-protein interactions and metabolome).

The Gene Ontology (http://www.geneontology.org/) has been used by many groups performing biological network analysis to measure functional similarity. These approaches focused mainly in finding (a) similarity score between two sets of gene annotation (terms) based on Information Content of the terms (probability of the term occurring in the Gene Ontology) (Lord *et al.* 2003), and (b) similarity score between gene pairs based on the number of common GO Terms (Chabalier *et al.* 2007, Huang *et al.* 2007, Mistry *et al.* 2008). The first group of studies uses the similarity measure as a function of the Information Content of Lowest Common Ancestor between the terms, but do not consider how far the terms are with respect to each other. Mistry *et al.* 2008 proposed a simpler method to find functional similarity between genes. The method is based on the Term

Overlap Measure as the cardinality of the common terms in two sets of gene annotation.

A. Correlation Networks and Ontology

The main thrust of the method presented in this study is to combine the power of correlation network with the parent-child nature of the Gene Ontology. We define a correlation network in this study as a network model composed of nodes and edges, where nodes represent genes or gene products and edges are drawn between them if the expression pattern of two particular genes over a series of homologous samples is correlated (correlation represented as ρ). Generally, a high correlation is considered to be $-1.00 \leq \rho \leq -0.70$ and $0.70 \leq \rho \leq 1.00$ (Dempsey *et al.* 2011, Carter *et al.* 2004) and thus networks made from this model are filtered to contain only correlations of this value. The resulting network typically has thousands of nodes and can have upwards of a million edges, although reducing the number of edges by correlation filtering and hypothesis testing is typical (Zhang, Horvath 2005). The networks used within our study are filtered to correlations of $0.91 \leq \rho \leq 1.00$ and p-value ≤ 0.0005 unless otherwise noted. Additionally, these networks tend to be "scale-free", which means that their degree distribution is indicative of a few nodes that are well connected (Albert *et al*. 2005, Barabasi *et al.* 2004).

The Gene Ontology works off the premise that the deeper the tree traversal, the deeper the specificity of the function obtained. Each child node in the GO tree can be considered a sub-function of its parent. This particular ontology was made to clarify functionalities of genes and is maintained and updated by the community and the GO Consortium to act as a repository for cellular functions, but it remains incomplete as the functionalities of genes and gene products are continually being added. Further, as the complexities of the mechanisms behind observed functions continued to be uncovered, it may be necessary to modify the structure of the Gene Ontology and other data repositories to match new findings.

Previous work by Dempsey *et al*. 2011 revealed that identification of high-degree nodes, or hubs, in correlation networks can be performed and that hubs identified indeed corresponded to lethal or essential genes at a rate of 50-70% (similar studies of essentiality in protein protein interaction networks can be found in Jeong *et al*. 2001). This implies that a filtered network created from gene expression data follows a power-law degree distribution, (correlations follow a normal distribution), and from these characteristics hubs can be identified. Hubs were verified for essentiality using the JAX Mouse Genome Informatics database (http://www.informatics.jax.org/) and out of the 29 hubs identified that had been tested for *in vivo* knockouts, only three were found to have no observable phenotype when tested. Eleven of hubs tested were lethal for mice pre- to post-natally, and 27 had some effect on one of the 24 major systems identified by the MGI database (growth, lifespan, etc.). The rate of hub essentiality is important because it suggests that the structures (hubs, clusters) found to be important in other biological networks (metabolome, interactome, etc) also have potential importance in correlation network studies. Thus, in this study we identify those critical network structures and use the Gene Ontology to aide in identifying their functional importance in the cell.

```
Input: Correlation network
Output: Returns the deepest common parent
and depth score of two genes connected by
an edge

For each edge in the correlation network{
 Get the set of all GO terms for gene1
 Get the set of all GO terms for gene2
 For each term1 in GO_set1
    For each term2 in GO_set2
    Get the GO depth score for term1 and
    term2 via function Get_GOScore (term1,
    term2)
}

Get_GOScore (term1, term2){
 For term t1, find all of its parent terms
 For each parent term (P1)
   Find minimum distance from P1 to term t1
   Find depth (distance from root)of parent

 For term t2, find all of its parent terms
 For each parent term (P2)
   Find minimum distance from P2 to term
   t2
   Find depth of parent
    Compare current parent terms P1 and P2
    For each common parent term, P1 == P2:
      Path length = sum of distance of
      term1 to P1 & distance of term2 to P2
      GO Score = Difference of common
      parent term (P1 == P2) to path length
    Return the highest GO score and
    corresponding parent term for term1 and
    term2
}
```

III. Methods

Network creation and filtering. Data was obtained from NCBI's Gene Expression Omnibus (GEO – see Barrett *et al*.) database in January 2011. Dataset used was GSE5078 (Verbitsky *et al*. 2004), which was originally designed to identify changes in gene expression relating to learning in the hippocampus of Young mice (YNG) versus middle-aged mice (MID). Networks were created as stated with Pearson Correlation coefficient and filtered to only the highest of correlations ($\rho = 1.00$) and only correlations passing P-value < 0.0005 were kept.

Young Clusters

Cluster 1		Cluster 6	
regulation of cellular process	**6.16%**	**metabolic process**	**21.05%**
cellular biosynthetic process	**6.16%**	**signal transduction**	**21.05%**
primary metabolic process	5.57%	protein modification process	10.53%
Cluster 2		Cluster 7	
primary metabolic process	**6.40%**	**metabolic process**	**13.33%**
multicell. organismal development	5.54%	**cellular macromolecule metabol**	**13.33%**
organ development	5.12%	**cellular protein metab.process**	**13.33%**
Cluster 3		Cluster 8	
regulation of cellular process	**7.48%**	**reg. of transcription, DNA-dep.**	**46.67%**
biosynthetic process	5.61%	regulation of cellular process	13.33%
regulation of biological quality	4.67%	neg. regulation of transcription, DN	13.33%
Cluster 4		Cluster 9	
regulation of cellular process	**16.22%**	**pos. regulation of cellular proces**	**7.14%**
biological_process	13.51%	**regulation of cellular process**	**7.14%**
cell proliferation	8.11%	metabolic process	5.36%
Cluster 5		Cluster 10	
macromole metabol. process	**9.84%**	**reg. of biological process**	**10.34%**
cellular biosynthetic process	8.20%	signal transduction	6.90%
organ development	4.92%	multicellular organismal process	6.90%

Mid Clusters

Cluster 1		Cluster 6	
multicellular organismal proc.	**7.13%**	**signal transduction**	**9.09%**
response to stimulus	4.99%	**response to chemical stimulus**	**9.09%**
regulation of cellular process	4.51%	**regulation of cell proliferation**	**9.09%**
Cluster 2		Cluster 7	
transport	**11.54%**	**metabolic process**	**6.09%**
regulation of cellular process	7.69%	cell adhesion	5.22%
cellular metabolic process	4.81%	primary metabolic process	5.22%
Cluster 3		Cluster 8	
signal transduction	**9.43%**	**pos. reg. of transcrip.-RNA polyn**	**15.38%**
regulation of cellular process	8.81%	**primary metabolic process**	**15.38%**
cell surface recep. linked signaling	6.29%	macromolecule metabolic process	11.54%
Cluster 4		Cluster 9	
transport	**8.51%**	**regulation of cellular process**	**10.48%**
regulation of cellular process	7.09%	metabolic process	6.67%
signal transduction	6.38%	signal transduction	4.76%
Cluster 5		Cluster 10	
regulation of cellular process	**25.93%**	**regulation of cellular process**	**12.50%**
signal transduction	14.81%	**cellular protein metabolic proc.**	**12.50%**
transport	14.81%	**cellular macromolec. biosynthet**	**12.50%**

FIGURE 2. Top 3 Edge Annotations for YNG and MID clusters: Terms are ranked according to percentage. Percentage represents number of edges in the cluster annotated with that term versus total edges in the GO depth filtered network. Bolded terms are the top term(s) by percentage for that cluster.

A. Algorithm

We describe the algorithm used in this study in page 3 and visually in Figure 1. Correlation networks are notorious for having noise – indeed; a major concern for those using correlation networks is filtering coincidental edges from causative relationships. Two genes connected in the correlation network who also share a deep relationship in the Gene Ontology are more likely to be important that two genes who have a high correlation but no GO relationship.

Our algorithm examines each edge in the network, where nodes represent genes and an edge represents the correlation in gene expression of those two genes. The algorithm then loads a local version of the GO association table, which links genes to its known GO terms (it is possible and probable that a gene will be associated with more than one term). Each gene then has a list of GO terms associated with it in memory, at which point the algorithm identifies matches between these two lists. It is possible for there to be no match in the lists (thus, this edge will not be present in the result network) or it is possible for the algorithm to identify multiple common terms, thus the need to identify the deepest (and therefore most specific) common GO term in the two sets. Once the deepest common parent (DCP) is identified, the nodes, score, and GO term id are output as a result network.

Caveats. Caveats associated with our method include the fact that the GO is incomplete and can contain false or misleading data. Future work will focus on statistical testing of our method for quality control and also integration of other knowledge-based information to improve result confidence scores.

IV. Experimental Results

The initial results of our method reveal a number of things about examining relationships in correlation networks and the completeness of the Gene Ontology itself. We highlight the most intriguing results here:

Networks filtered using our method were found to maintain the integrity of the scale-free biological network – resulting networks have a power-law node degree distribution and contain critical high-degree nodes, or hubs. Out of the 100 nodes in the network with the highest degree, more than half of the nodes tested for in vivo knockout are involved in mortality/aging as determined by the MGI database (62.82% - YNG, 67.14% MID). This is in agreement with previous studies investigating correlation networks.

Application of our method resulted in drastic reduction of node and edge number in each network. Edges with no relationship or whose deepest common parent was the root of the tree (GO:0008150, biological process) were removed. 50.22% of edges were removed from the YNG original network and 51.32% of edges were removed from the MID original network. Even with the drastic reduction of edges, modularity of the network was maintained. We identified the top 10 clusters within the filtered networks using AllegroMCODE under default settings (www.allegroviva.com/allegromcode) by Jun *et al.* 2011. After cluster identification, we represented the edge annotations of our GO method as a percentage of the total edges. A portion of these results are presented in Figure 2. The top GO term per cluster for the YNG network was responsible for 14.46% of the edges on average; in the MID network the average was 11.61%. These numbers do not network the average was 11.61%. These numbers do not take into account overlapping or parent-child terms. Edges annotated as "metabolic process" (depth of 1) were considered separate from "primary metabolic process" (depth of 2). Combination of parent-child terms is planned for future studies.

We compared our method to the node populations of each cluster by identifying all GO terms for each gene in the cluster and then identifying the most common terms among the union of all gene GO term sets. The top results from the

						Our Method			Traditional Method		
Cluster	NC	EC	TPEs	Density	Avg. Depth	Annotation	Anno. Depth	%	Annotation	Anno. Depth	%
yng 1	43	782	903	0.8660	2.8056	regulation of cellular process cellular biosynthetic process	3 3	6.16% 6.16%	metabolic process	1	6.54%
yng 2	62	878	1891	0.4643	2.7335	primary metabolic process	2	6.40%	metabolic process	1	6.70%
yng 3	21	189	210	0.9000	2.8677	regulation of cellular process	3	7.48%	metabolic process	1	8.92%
yng 4	15	60	105	0.5714	2.4667	cellular biosynthetic process	3	17.78%	primary metabolic process metabolic process	2 1	8.93% 8.93%
yng 5	25	94	300	0.3133	2.5851	macromolecule metabolic process	2	9.84%	primary metabolic process metabolic process	2 1	7.41% 7.41%
yng 6	8	28	28	1.0000	2.6429	metabolic process signal transduction	1 3	5.61% 5.61%	metabolic process	1	6.54%
yng 7	12	33	66	0.5000	2.7576	metabolic process cellular macromolecule metabolic process cellular protein metabolic process	1 3 4	13.33% 13.33% 13.33%	metabolic process primary metabolic process	1 2	8.79% 8.79%
yng 8	6	15	15	1.0000	4.6667	regulation of transcription, DNA-dependent	7	46.67%	primary metabolic process metabolic process	2 1	6.67% 6.67%
yng 9	41	102	820	0.1244	3.2549	positive regulation of cellular process regulation of cellular process	4 3	7.14% 7.14%	developmental process	1	4.43%
yng 10	19	44	171	0.2573	2.3182	response to stimulus	1	12.12%	cell communication	2	4.05%

FIGURE 3: YNG Network Results: NC = Node count, EC = Edge Count, TPEs = Total Possible Edges. Density is the density of the cluster based on actual versus possible edges, and average depth is the average of the DCP of edges annotated with a depth of 0 or higher. The annotations from our method and the traditional method are the most common edge annotations (our method) and node annotations (traditional method) per each cluster.

						Our Method			Traditional Method		
Cluster	NC	EC	TPEs	Density	Avg. Depth	Annotation	Anno. Depth	%	Annotation	Anno. Depth	%
mid 1	41	645	820	0.7866	3.0512	multicellular organismal process	1	7.13%	metabolic process	1	8.14%
mid 2	28	184	378	0.4868	2.7554	transport	3	11.54%	metabolic process	1	5.24%
mid 3	41	239	820	0.2915	2.5565	signal transduction	3	9.43%	cell communication	2	5.01%
mid 4	41	208	820	0.2537	2.8750	transport	3	8.51%	primary metabolic process	2	5.45%
mid 5	10	43	45	0.9556	2.6047	regulation of cellular process	3	25.93%	primary metabolic process	2	5.26%
mid 6	11	47	55	0.8545	2.8511	signal transduction response to chemical stimulus regulation of cell proliferation	3 2 4	9.09% 9.09% 9.09%	cell communication	2	8.54%
mid 7	65	225	2080	0.1082	2.9733	metabolic process	1	6.09%	metabolic process	1	6.65%
mid 8	10	30	45	0.6667	3.3000	positive regulation of transcription from RNA polymerase II promoter primary metabolic process	8 2	15.38% 15.38%	primary metabolic process metabolic process	2 1	8.97% 8.97%
mid 9	58	171	1653	0.1034	2.6199	regulation of cellular process	3	10.48%	primary metabolic process	2	5.35%
mid 10	12	33	66	0.5000	2.3333	regulation of cellular process cellular protein metabolic cellular macromolecule biosynthetic process	3 4 3	12.50% 12.50% 12.50%	primary metabolic process cell communication signal transduction metabolic process	2 2 3 1	7.55% 7.55% 7.55% 7.55%

FIGURE 4. MID Network Results: NC = Node count, EC = Edge Count, TPEs = Total Possible Edges. Density is the density of the cluster based on actual versus possible edges, and average depth is the average of the DCP of edges annotated with a depth of 0 or higher. The annotations from our method and the traditional method are the most common edge annotations (our method) and node annotations (traditional method) per each cluster.

node population analysis are shown in Figures 3 and 4. Some common themes we have identified from our method are the following:

Density does not necessarily imply likelihood of common or true function. YNG Cluster 1 and YNG Cluster 9 both have a similar node count (41 and 43 respectively) but YNG C1 is much more dense. However, average density and annotation with our method *and* traditional methods indicate that the YNG C9 cluster has more defined function. The difference in annotation depth is subtle, but the difference in density decides the rank of YNG C9 in clustering. Using our method, it would be suggested that YNG C9 is more likely to be representative of true cellular function and C1.

Traditional methods may overlook the relationships that edges represent. YNG Cluster 8 (Fig. 2) is a small, complete clique (K_6) and 7 of its edges were found to be associated with DNA-dependent regulation of transcription. Our method readily identifies this set of edges.

Clusters likely cannot be annotated with just one function; instead, they should be represented as a distribution of functions. While the reason behind functional overlaps in clusters remains unclear, it is evident that oftentimes a cluster cannot be associated with only one function. A 'pleiotropic' nature of a cluster is entirely possible within the concept of a cellular network.

V. Discussion

In this study, an algorithm has been presented for identifying the deepest common relationship between two nodes in a network using the parent-child structure of the Gene Ontology. This method allows for the functional annotations for edges in correlation networks. We show that this method enhances the functional relationships within clusters, and provides a stepping stone for future work by allowing us to better identify clusters of interest with higher likelihood of actual cellular impact. In the future, we hope to apply this work to the automation of functional cluster annotation in correlation networks, resulting in the ability to visually and computationally reduce the size and complexity of the correlation network while maintaining biological relevance of network structure. Further, this method could be supplemental in identifying unknown gene function by exposing the common shared functions of genes in common clusters of the unknown gene.

As the Gene Ontology and other publicly available data warehouses continue to expand and improve, so will the need for tools to incorporate the data within these stores while sorting biological signal from noise. It is critical, then, that the ability to automate these processes of identifying functions of network structures becomes available to the scientific community. Our method specifically has shown that examining the ontological enrichment of edges in the correlation network is critical for understanding function of network structures.

Acknowledgement

This publication was made possible by Grant Number P20 RR16469 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH) and its contents are the sole responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

References

[1] Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium (2000) Nature Genet. 25: 25-29

[2] Albert, R. (2005). Scale-free networks in cell biology. Journal of Cell Science, 118(Pt 21), 4947-4957.

[3] Barabasi, A. L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. Nature Reviews.Genetics, 5(2), 101-113.

[4] Bender A, Beckers J, Schneider I, Hölter SM et al.. Creatine improves health and survival of mice. Neurobiol Aging 2008 Sep;29(9):1404-11. PMID: 17416441

[5] Cakmak A et al. Gene ontology-based annotation analysis and categorization of metabolic pathways. Proceedings of SSDBM 2007.

[6] Carter, S. L., Brechbuhler, C. M., Griffin, M., & Bond, A. T. (2004). Gene co-expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics (Oxford, England), 20(14), 2242-2250.

[7] Chabalier J, Mosser J, Burgun A: **A transversal approach to predict gene product networks from ontology-based similarity.** *BMC Bioinformatics* 2007, **8:**235

[8] Dempsey, K., Bonasera, S., Bastola, D., Ali, H.. (2011) A Novel Correlation Networks Approach for the Identification of Gene Targets. System Sciences (HICSS), 2011 44th Hawaii International Conference onSystem Sciences , vol., no., pp.1-8, 4-7

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

[10] Hishigaki H, Nakai K, Ono T, Tanigami A, Takagi T (2001) Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18: 523–531

[11] Huang da W, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA: **The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists.** *Genome Biol* 2007, **8**(9)**:**R183

[12] Jeong, H., Mason, S. P., Barabasi, A. L., & Oltvai, Z. N. (2001). Lethality and centrality in protein networks. Nature, 411(6833), 41-42.

[13] Jun S. Yoon and Won-Hyung Jung, "A GPU-accelerated bioinformatics application for large-scale protein interaction networks", APBC poster presentation, 2011. (http://www.allegroviva.com/allegromcode)

[14] Karaoz U, Murali TM, Letovsky S, Zheng Y, Ding C, Cantor CR, Kasif S (2004) Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA 101: 2888–2893

[15] Lord PW, Stevens RD, Brass A, Goble CA: **Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.** *Bioinformatics* 2003, **19**(10)**:**1275-1283.

[16] Mistry, M., and Pavlidis, P. (2008). Gene Ontology term overlap as a measure of gene functional similarity. BMC Bioinformatics, 9:327.

[17] Opgen-Rhein, R., & Strimmer, K. (2007). From correlation to causation networks: A simple approximate learning algorithm and its application to high-dimensional plant gene expression data. BMC Systems Biology, 1, 37.

[18] Paul D. Thomas, Michael J. Campbell, Anish Kejariwal, Huaiyu Mi, Brian Karlak, Robin Daverman, Karen Diemer, Anushya Muruganujan, Apurva Narechania. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res., 13: 2129-2141 .

[19] Reverter, A., & Chan, E. K. (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.

[20] Schwikowski B, Uetz P, Fields S (2000) A network of protein–protein interactions in yeast. Nat Biotechnol 18: 1257–1261

[21] Vazquez A, Flammini A, Maritan A, Vespignani A (2003) Global protein function prediction from protein–protein interaction networks. Nat Biotechnol 21: 697–700

[22] Verbitsky, M., Yonan, A. L., Malleret, G., Kandel, E. R., 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.

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

[24] Zhu, M, Gao, L., Guo, Z., Li, Y., Wang, D, Wang, J. Wang, C. (2007) Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities. Gene 2007b Apr 15; 391(1-2):113-9. Epub 2006 Dec. 22.