View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by The University of Nebraska, Omaha

University of Nebraska at Omaha
DigitalCommons@UNO

Interdisciplinary Informatics Faculty Proceedings & Presentations

School of Interdisciplinary Informatics

2015

On the Comparison of State- and Transition-based Analysis of Biological Relevance in Gene Co-expression Networks

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

Prasuna Vemuri
University of Nebraska at Omaha

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

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

Part of the Bioinformatics Commons, and the Genetics and Genomics Commons

Recommended Citation

Cooper, Kathryn Dempsey; Vemuri, Prasuna; and Ali, Hesham, "On the Comparison of State- and Transition-based Analysis of Biological Relevance in Gene Co-expression Networks" (2015). *Interdisciplinary Informatics Faculty Proceedings & Presentations*. 25.
https://digitalcommons.unomaha.edu/interdisciplinformaticsfacproc/25

This Conference Proceeding is brought to you for free and open access by the School of Interdisciplinary Informatics at DigitalCommons@UNO. It has been accepted for inclusion in Interdisciplinary Informatics Faculty Proceedings & Presentations by an authorized administrator of DigitalCommons@UNO. For more information, please contact unodigitalcommons@unomaha.edu.

On the Comparison of State- and Transition-based Analysis of Biological Relevance in Gene Co-expression Networks

Kathryn Cooper, Prasuna Vemuri, and Hesham Ali
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE
hali@unomaha.edu

Abstract— **Traditional correlation network analysis typically involves creating a network using gene expression data and then identifying biologically relevant clusters from that network by enrichment with Gene Ontology or pathway information. When one wants to examine these networks in a dynamic way – such as between controls versus treatment or over time – a "snapshot" approach is taken by comparing network structures at each time point. The biological relevance of these structures are then reported and compared. In this research, we examine the same "snapshot" networks but focus on the enrichment of changes in structure to determine if these results give any more insight into the mechanisms behind observed phenotypes. Our main hypothesis is that more information, particularly related to potential dynamic changes, can be obtained through transition-based analysis of biological networks. To test this hypothesis, we compare gene expression data from the mouse hippocampus at three different time points: young, middle-aged, and aged, and compare the traditional state-based approach to the dynamic transition-based enrichment approach. In this study we use a clustering approach (SPICi) designed specifically for clustering of large biological networks. The results of this study verify an inconsistency between traditional and dynamic structure identification approaches through biological enrichment. These results highlight an intriguing issue for those performing, critiquing, and using network based approaches in their research – that a black box or workflow type of approach typically used in network based research can be supplemented with a transition-based approach to support movement from in silico to in vivo experimentation of target genes.**

***Keywords*— state-based, transition-based, network analysis, gene expression**

I. INTRODUCTION

With explosive rise in availability of biological data, issues associated with stability of methods and consistency of computational results continue to be of major concern to biomedical researchers [1-5]. Along with this data influx, there has been a call for marked increase in activities to develop computational tools for big bioinformatics data analytics to extract useful information from the available data. The first generation of bioinformatics tools have primarily focused on using simple methods for basic analysis of homogenous biological data and extracting the glaring signals from the available datasets. Now with the availability of many types of biological datasets obtained under different experimental conditions and carrying various levels of accuracy and quality, new advanced tools are needed to extract different types of relationships associated with signals of varying levels of strength. Hence, using simple standard workflow approaches with generic "black box" computational steps may no longer be as effective in the analysis of current big bioinformatics data. In this work, we conduct as in-depth study of network models in analyzing biological data and illustrate how important it is to employ advanced data analytical approaches to extract accurate and trustworthy signals for the available biological datasets.

In this research, we defined a network as set of nodes or vertices connected by edges defined by correlation of gene expression. There are different types of networks span a variety of domains, namely, technological, biological, social, economic networks, and information networks [6-13]. Visual comparison of large networks can be misleading, and typical networks when dealing with "big" biomedical data are large and difficult to manage with a traditional graphic user interface (GUI) approach. Therefore, we define set of quantitative measures using graph theory such as clustering, degree distributions and so on to compare them. A standard approach to basic biological network analysis involves network creation, identification of structures such as high-degree or hub nodes, clusters, bottlenecks, or pathways, and then evaluation of the biological relevance of these identified structures [6,8,14-16]. Comparison of networks at different states, control versus treatment, for example, is typically done by comparison of these structures and their biological relevance using a "snapshot" view of the system: the network model for a given state is created and after cleaning and pre-processing, is assumed to be representative of the observed phenotype, state, or condition. It has been suggested in other non-biological network analyses that this snapshot approach may not be the best indicator of network dynamics or mechanistic change [17,18], but rather, examining ***changes*** across time (biologically, this would likely imply short time-series data) can be as, if not more, informative of the goings-on within a cellular environment. We define these two viewpoints as state-based or transition-based: state-based implying a snapshot evaluation of network structure and function, and transition-based implying a dynamic evaluation of network structure and function. Currently, publicly shared data that lends itself towards a true transition-based approach is not readily available – this would require short-term time series gene expression analyses with

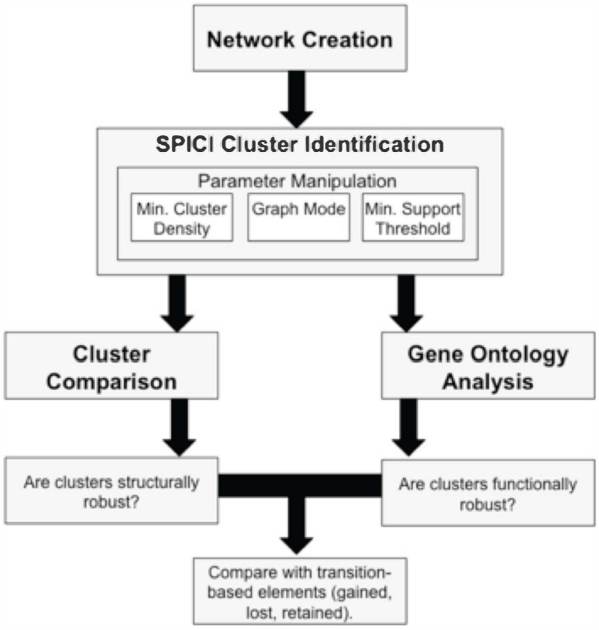

Figure 1. SPICi flowchart for parameter manipulation.

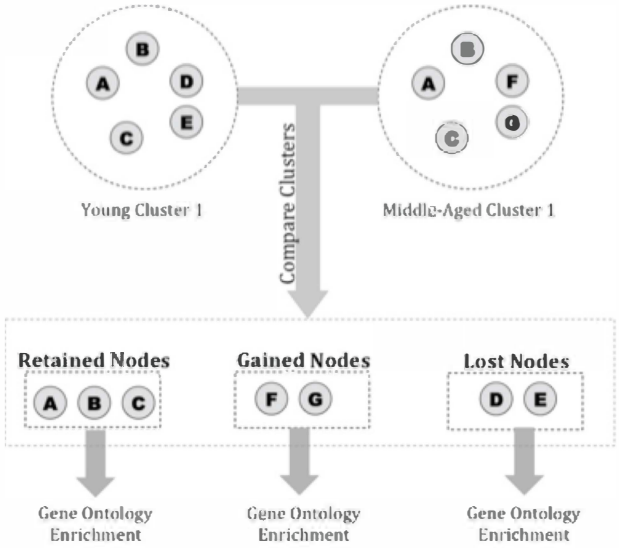

Figure 2. Our transition-based approach to identifying three sets of important nodes: gained, lost, and retained. For any given clluster from the first network and for any given cluster from the second network, the node content of the two clusters is compared. Nodes that are unique to the first cluster are considered lost, nodes that are unique to the second cluster are considered gained, and nodes common to both clusters are considered retained. This pairwise cluster comparison is performed for every cluster in network 1 and every cluster in network 2. Each individual set of nodes (lost, gained, and retained) is then evaluated for Gene Ontology enrichment

high sample counts per state or treatment. However, the data that are currently available can be analyzed in a crude transition-based format by examining which structures appear, disappear, or are retained from one state to the next. Examining the biological relevance of network structures identified in this way is the focus of this research. Briefly then, the outline of this paper is as follows. In Section II we describe in detail the implementation of our methods and how these clusters are analyzed. In Section III the findings obtained from the approach are analyzed by performing gene ontology on the clusters. In Section IV the results are discussed and hypothesis is concluded.

Our main objective in this paper is to compare traditional state-based network analysis of correlation networks to the transition-based network analysis to determine if there is consistency of biological relevance between the two. The first objective will provide insight into a potentially novel way to assess the biological relevance of a network and show whether or not using tools such as enrichment yield reproducible annotations of function and the second objective serves to investigate our speculation that a "black-box" type of approach is not appropriate for biological network analysis.

II. Methods

A. Network Creation

Gene expression data was collected from the mouse brain of male Balb/C mice at ages 2 months (young), 12 months (middle-aged) and 18 months (aged) with no treatment. A total of 41,174 probes were analyzed per dataset and each of these three datasets contained 6 samples. Three networks (called Young, Middle, and Aged) were created by performing Pearson correlation between pairwise vectors of gene expression data where each gene represents a vector over a number of samples. Correlations with a threshold of 0.85 to 1.00 were kept with p-value less than or equal to 0.005. The publicly available R language [19] and accompanying library *igraph* [20] was used to identify the number of nodes, edges, density, cluster distribution, degree distribution, and cluster sizes for each network (using the *V(g), E(g), clusters*, and *degree.distribution* functions of *igraph*).

B. Clustering and Analysis

Clustering was performed using a recent biological network adapted algorithm called "Speed and Performance in Clustering," or SPICi for short [21]. SPICi is a self-described "fast local network clustering algorithm" designed for biological networks [21], which tend to be large and sparse [6]. SPICi has a time complexity of O(V log V + E) where V includes the set of all nodes contained within a network or graph and E includes the set of all edges contained in the graph or network. SPICi was originally tested for use in protein-protein interaction networks (PPIs) which are built by identifying binary or weighted affinity scores (edges) between physically interacting proteins (nodes) [8]. In 2010, SPICi developers Jiang and Singh reported that the algorithm performed competitively at identifying protein complexes within protein-protein interaction networks as well as modules of biological function [21]. Typically, this indicates that the network structure lends itself toward clusters that are on the smaller side (approximately 2-20 nodes or protein components) and they are relatively dense (all proteins in a complex may not physically interact but clusters will largely have a high density due to multiple domains of interaction). In summary, we have

Transition-based Cluster Comparison Code

```
Input: C1 = set of clusters from state 1 of
       size n
       C2 = set of clusters from state 2 of
       size m
       Assume no node duplicates in any
       cluster
  Output:Nodes gained, lost, &
retained for
       each pairwise cluster comparison

  Call cluster comparison(C1,C2):
for each cluster ci in C1 where i = 1 to n
       nodesx = list of nodes in current
       cluster ci
       for each node nk in nodesx
              hash{nk} = 1;
              for each cluster cj in C2
              where j = 1 to m
                     nodesy = list of nodes
                     in current cluster ci
              for each node n1 in nodesy
                     if(hash{nk} = 1)
                            hash{nk} = 2;
              else if (hash{nk} does not
              exist)
                     hash{nk} = 3;
  for each key k and value v in hash{}
       if v = 1, output k and "lost"
       if v = 2, output k and "retained"
```

chosen a clustering algorithm designed specifically for networks. The rationale behind choosing this algorithm is intended to examine a key objective of this research – are the structural identities of clusters inherently easy or robust to different clustering algorithms? Our speculation is that the nature and complexity of the biological network model would lend itself to small, dense clusters, but that these clusters are not necessarily readily apparent and as such may not be consistently found from one algorithm to the next, or even within different parameterizations of one algorithm.

To first assess the consistency of SPICi results, we chose to manipulate a number of parameters offered by the algorithm as shown in Figure 1: minimum density, minimum support, and graph mode (or how sparse/dense the network is). To perform this, we performed 75 different iterations of clustering on all three networks: where minimum density ranged from 0.0 to 1.0 in 0.25 increments, minimum support ranged from 0.0 to 1.0 in 0.25 increments, and graph mode was given as 0,1, or 2. Using the graph mode value of 2, clustering took over 1 hour and as such these processes were terminated. Other parameterizations took >10 minutes and gave a variety of different results, shown as Average Cluster size and Number of Clusters below in Table 1. Ultimately, the default parameters for SPICi were used in the final analysis and implementation.

After clustering was performed, two different paths to analysis were taken: the traditional **state-based** approach, where clusters are analyzed by Gene Ontology (GO) enrichment, and the proposed **transition-based** approach, where clusters are compared between states (i.e. young versus middle-aged, middle-aged versus aged). In the transition-based

TABLE I. TABLE 1. PARAMETERIZATION RESULTS

			# of Clusters		
Min Density	Min Support	Graph Mode	Yng	Mid	Aged
0	0	0	2	2	5
0.25	0	0	676	644	695
0.5	0	0	1409	1380	1442
0.75	0	0	2427	2378	2429
1	0	0	2423	3534	2587
0	0.25	0	74	62	42
0.25	0.25	0	685	662	709
0.5	0.25	0	1410	1383	1443
0.75	0.25	0	2427	2378	2429
1	0.25	0	2423	3534	2587
0	0.5	0	1166	1161	1150
0.25	0.5	0	1222	1218	1220
0.5	0.5	0	1576	1585	1579
0.75	0.5	0	2446	2397	2444
1	0.5	0	2423	3534	2587
0	0.75	0	10152	12902	10471
0.25	0.75	0	10152	12902	10471
0.5	0.75	0	10152	12902	10471
0.75	0.75	0	10152	12902	10471
1	0.75	0	2423	3534	2587
0	1	0	10152	12902	2587
0.25	1	0	10152	12902	10471
0.5	1	0	10152	12902	10471
0.75	1	0	10152	12902	10471
1	1	0	2423	3534	2587
0	0	1	2	2	5
0.25	0	1	672	643	695
0.5	0	1	1415	1385	1442
0.75	0	1	2423	2381	2429
1	0	1	2377	3592	2582
0	0.25	1	74	62	42
0.25	0.25	1	686	662	710
0.5	0.25	1	1417	1388	1443
0.75	0.25	1	2423	2381	2429
1	0.25	1	2377	3592	2582
0	0.5	1	1168	1158	1151
0.25	0.5	1	1239	1220	1222
0.5	0.5	1	1579	1585	1584
0.75	0.5	1	2439	2392	2440
1	0.5	1	2377	3592	2582
0	0.75	1	10169	12893	10484
0.25	0.75	1	10169	12893	10484
0.5	0.75	1	10169	12893	10484
0.75	0.75	1	10169	12893	10484
1	0.75	1	2377	3592	2582
0	1	1	10169	12893	2582
0.25	1	1	10169	12893	10484
0.5	1	1	10169	12893	10484
0.75	1	1	10169	12893	10484
1	1	1	2377	3592	2582

approach, the set of clusters $C_{state1} = \{c_{state1.1}, c_{state1.2}, \ldots, c_{state1.n}\}$ is compared to the set of clusters $C_{state2} = \{c_{state2.1}, c_{state2.2}, \ldots, c_{state2.m}\}$ by comparing node content. Briefly, we want to compare each cluster in from network at state 1 (for example, young) to each cluster from the network at state 2 (for example, middle-aged) such that for each cluster comparison, three outputs are generated: the nodes that are unique to the cluster at state 1 (lost nodes), the nodes that are unique to the cluster at state 2 (gained nodes), and the nodes in common to both clusters (retained). An example of this is shown in Figure 1. This approach was implemented in Java and the pseudocode is described above. The result of the above program would be the list of nodes that are retained, gained and lost between any two stages of clusters.

TABLE II. BASIC NETWORK CHARACTERISTICS FOR EACH OF THE THREE NETWORKS DEFINED: YOUNG, MIDDLE, AND AGED. NODE AND EDGE COUNTS WERE DETERMINED IN R USING IGRAPH. DENSITY IS CALCULATED AS THE EDGE COUNT DIVIDED BY THE TOTAL NUMBER OF POSSIBLE EDGES, OR (N*(N-1))/2. THE FINAL TWO COLUMNS REPRESENT THE NUMBER OF CLUSTERS FOUND USING BOTH CLUSTERING ALGORITHMS; ALREADY WE SEE A VERY LARGE DIFFERENCE IN APPROACHES. THE TABLE DOES NOT ACCURATELY SHOW IT, BUT THE CLUSTERS FOUND BY SPICI TEND TO BE SMALLER (<50 NODES) THE CLUSTERS IDENTIFIED BY IGRAPH CLUSTERS TEND TO BE LARGE AND SPARSE.

Network	Node Count	Edge Count	Density	# of Clusters found SPICi	# of Clusters found *igraph clusters*
Young	38,618	1,083,522	0.128%	659	1
Middle	39,964	4,937,369	0.582%	654	1
Aged	39,728	1,925,865	0.227%	644	1

Finally, Gene Ontology overrepresentation was performed using PANTHERDB.org's Statistical Overrepresentation test using the *Mus musculus* organism background set and the Gene Ontology's biological process tree annotation set (with *only* experimentally verified terms being used). Gene Ontology overrepresentation was performed on both traditional state-based clusters (named in results as Young, Middle, and Aged) and proposed transition-based cluster dynamics (named in results as Young vs. Middle – Lost, Young vs. Middle – Gained, Young vs. Middle – Retained, Middle vs. Aged– Lost, Middle vs. Aged– Gained, Middle vs. Aged– Retained).

The study is designed to test our main hypothesis that more biologically relevant information can be extracted from biological networks created using the transition-based model as compared to the traditional state-based approach. We also attempt to test a secondary hypothesis that the obtained results are influenced by which computational tools, like clustering algorithms, are used to analyze the created networks. Positive affirmation that a simple black box usage of computational tools or a standard way of employing a generic workflow may fall short in providing stable and reproducible data analytics tools for many complex bioinformatics applications.

III. RESULTS

A. CLUSTERING PARAMETERIZATIONS

We found that manipulating graph parameterizations resulted in multiple types of clusters being found, with a sample of these results shown in Table 1. We found that using graph mode = 2, for a large sparse graph, runtimes exceeded an hour and were terminated. For other parameterizations, for example, the extremes (minimum support = 1 or minimum density = 1) clusters typically ran aground of their "norm" for that particular parameter set. We found that while cluster parameterizations could potentially result in very different networks, this is a reflection of the abilities of the algorithm itself to distinguish between different graph properties. It is not fair to ask the algorithm to provide consistent results between each parameterization, and ultimately, it is up to the user to determine the optimal settings for their usage. As a result, we used default parameters in our analysis of clusters (minimum density of 0.5, graph mode of sparse, and a minimum support threshold of 0.5) and are confident that these results reflect the networks generated.

B. FUNCTIONAL RELEVANCE.

Initial results examining networks highlight some basic network statistics in Table 2; Node counts were similar for each network containing 93.8%, 97.1%, and 96.5% (Young, Middle, and Aged, respectively) of the original probes from each dataset. Edge counts varied much more and all were found to be very sparse networks. With an original probe count (n) of 41,174, the total possible number of edges for each individual network would be (n*(n - 1))/ 2, or 847,628,551 edges. The density then reflects the percentage of observed edges (Edge Count) to the number of total possible edges. All networks are very sparse, with the Middle-Aged or Middle network being the most dense (used loosely) and the Young network being most sparse. The final columns in Table 1 represent the number of clusters found by each clustering algorithm as represented by the algorithm itself; SPICi found on average many smaller clusters and *igraph clusters* found only one (very large) cluster.

The traditional state-based approach overall examines clusters, hubs, and other structures in individual networks from a snapshot point of view. The transition-based approach determines which structures are gained, lost, and missing from each network to incorporate a more dynamic point of view.

Table 3 shows the results for the top 5 SPiCI cluster results (the results all 600+ clusters per network were too large to show) using state-based and transition-based approaches. In Table 3, an "X" is marked for that approach if the cluster was found to be enriched with the term under the "Biological Process" column. Rows with annotations occurring in more than one result are *italicized* and in this table, due to a higher number of annotations, groups of annotations found with similar transitions are grouped together to save space. It is found that total out of 116 annotations, 76 (65.5%) were found in more than one result, spanning either State-based or Transition-based approaches (not listed due to size, please refer to Table 3). Of these groups, four major "patterns" emerged: (1) Terms present in the state-based approach (Young) and transition-based approach (Y_M Lost); (2) Terms present in the state-based approach (Middle) and transition-based approach (M_A Lost); (3) Terms present in the Aged network only, and (4) terms present in the M_A Lost network only. For both Y_M Gained/Retained and M_A Gained/Retained, no biologically significant terms were found.

TABLE 3. THE STATE-BASED AND TRANSITION-BASED APPROACH RESULTS FOR THE SPICI ALGORITHM. COLUMN 2 DENOTES THE BIOLOGICAL PROCESS ANNOTATION WAS FOUND, AND THE REMAINDER OF THE COLUMNS ARE SPLIT INTO STATE-BASED AND TRANSITION-BASED. THE STATE-BASED COLUMNS REPRESENT CLUSTERS IDENTIFIED IN THREE NETWORKS (YOUNG, MIDDLE, AND AGED) VIA THE *CREATE NETWORK* → *FIND CLUSTERS* → *PERFORM ENRICHMENT* APPROACH. THE TRANSITION-BASED COLUMNS REPRESENT CLUSTER TRANSITIONS IDENTIFIED IN TWO NETWORK COMPARISONS OF *YOUNG VERSUS MIDDLE* (Y_M) AND *MIDDLE VERSUS AGED* (M_A). EACH INDIVIDUAL COLUMN REPRESENTS NODES LOST, GAINED, OR RETAINED

	SPiCI								
	Traditional			Dynamic					
Biological Process	**Young**	**Middle**	**Aged**	**Y_M Gained**	**Y_M Lost**	**Y_M Retain**	**M_A Gained**	**M_A Lost**	**M_A Retained**
cell proliferation in forebrain, cellular glucan metabolic process, positive regulation of T cell differentiation in thymus, positive regulation of thymocyte aggregation, positive regulation of toll-like receptor signaling pathway, regulation of alpha-beta T cell proliferation, regulation of toll-like receptor signaling pathway	*x*				*x*				
cardiac chamber formation, eye morphogenesis, eye photoreceptor cell development, eye photoreceptor cell differentiation, negative regulation of peptidyl-serine phosphorylation, photoreceptor cell development, photoreceptor cell differentiation, post-embryonic development, regulation of release of sequestered calcium ion into cytosol by sarcoplasmic reticulum, regulation of ryanodine-sensitive calcium-release channel activity, sensory organ morphogenesis	*x*				*x*				
amide biosynthetic process, cardiac ventricle formation, deoxyribonucleoside triphosphate metabolic process, mesonephros development, purine deoxyribonucleoside triphosphate metabolic process, regulation of DNA repair, regulation of lymphocyte mediated immunity, regulation of response to DNA damage stimulus, regulation of skeletal muscle contraction, regulation of skeletal muscle contraction by calcium ion signaling	*x*				*x*				
DNA alkylation, DNA methylation, DNA methylation or demethylation, establishment of epithelial cell polarity, glial cell apoptotic process, lipid homeostasis, lipoprotein biosynthetic process, lipoprotein metabolic process, protein lipidation,	*x*				*x*				
B cell proliferation, CD4-positive, alpha-beta T cell differentiation, digestion, import into cell, leukocyte apoptotic process, peptidyl-glutamic acid modification, positive regulation of mitotic nuclear division, regulation of neuron migration, regulation of peptidyl-cysteine S-nitrosylation, regulation of transcription involved in primary germ layer cell fate commitment, response to ATP		*x*						*x*	
cell migration involved in gastrulation, negative regulation of extrinsic apoptotic signaling pathway in absence of ligand, negative regulation of signal transduction in absence of ligand, positive regulation of cartilage development, positive regulation of chondrocyte differentiation, positive regulation of natural killer cell mediated cytotoxicity, positive regulation of natural killer cell mediated immunity, protein deglutamylation, protein side chain deglutamylation		*x*						*x*	
midbrain-hindbrain boundary development, positive regulation of appetite, positive regulation of NIK/NF-kappaB signaling, positive regulation of response to food, regulation of systemic arterial blood pressure by hormone, regulation of systemic arterial blood pressure by renin-angiotensin, rostrocaudal neural tube patterning, toxin metabolic process, vasoconstriction		*x*						*x*	
cation transmembrane transport, inorganic anion transport, ion transmembrane transport, organic hydroxy compound transport, regulation of histone H3-K9 methylation, regulation of protein kinase A signaling, response to drug, serotonin transport, spermatid nucleus elongation, synaptic transmission, cholinergic		*x*						*x*	
interleukin-33-mediated signaling pathway		x							
immune system process			x						
biological adhesion, cellular macromolecule localization, developmental maturation, negative regulation of cell development, positive regulation of cell development, positive regulation of nervous system development, regulation of cell development, regulation of growth, regulation of nervous system development, secretion, secretion by cell			x						
cell adhesion, cell-cell adhesion			x						
canonical Wnt signaling pathway, determination of bilateral symmetry, determination of left/right symmetry, extracellular matrix-cell signaling, heart looping, specification of symmetry			x						
calcium-dependent cell-cell adhesion via plasma membrane cell adhesion molecules, cellular response to xenobiotic stimulus, lateral ventricle development, membrane fusion, positive regulation of circadian rhythm, positive regulation of humoral immune response mediated by circulating immunoglobulin, regulation of humoral immune response, regulation of humoral immune response mediated by circulating immunoglobulin, synapse assembly, ventricular system development			x						
cerebellar cortex development, cerebellar granular layer development, cerebellar granular layer formation, cerebellar granular layer morphogenesis, cerebellar granule cell differentiation, dentate gyrus development, hippocampus development, negative regulation of proteasomal protein catabolic process, negative regulation of proteasomal ubiquitin-dependent protein catabolic process, regulation of stem cell maintenance								x	

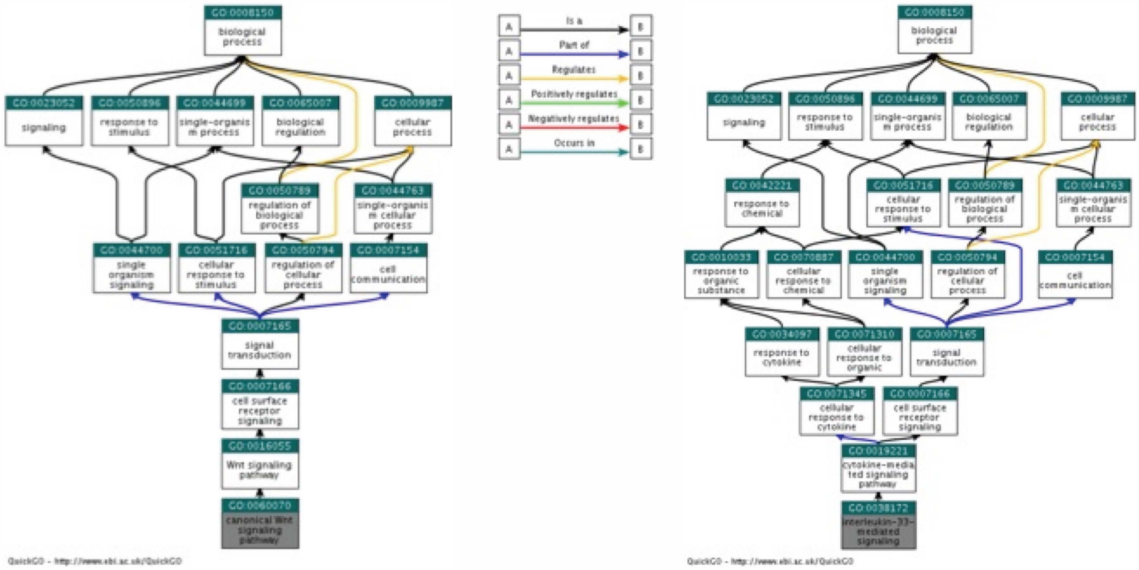

Figure 3. The Gene Ontology ancestor chart for both the canonical Wnt Signaling Pathway and interleukin-33 mediated signaling pathway identified as significant by the SPICi clustering approach generated by QuickGO (http://www.ebi.ac.uk/QuickGO).

Results between state and transition-based approaches in the generic algorithm are inherently *different* from one another – the same GO terms do not show up in the traditional and transition-based approaches. The structures with the *most* overlap between clusters were the nodes *gained* and *lost* from the young to middle and middle to aged networks in the generic *igraph clusters* algorithm analysis, and this trend was only present in 10 of the observed GO annotations. The (three large) clusters generated by the generic approach all had large size containing anywhere from 1,043 to 1,071 nodes – but very few of the GO terms found to be enriched in any of them overlapped directly with one another. Finally, there were *very few* nodes retained from the young to middle and middle to aged networks (young to middle – 97 and middle to aged – 2).

By contrast, results between state and transition-based approaches in the biologically motivated algorithm are more similar to one another – around 65% of the terms found in the state-based approach were found in the transition based approach. This implies that perhaps the biologically motivated algorithm SPICi is better able to find clusters that are robust in the network. Regarding our objective, it would appear that there is more consistency between biological relevance of the transition-based approach and the traditional state-based approach. To summarize, we denote the number of unique annotations identified in each approach by comparing both transition- and state-based approaches (Table 3). Here, we find that SPICi's biologically motivated clustering approach found 110 unique annotations for the top 5 clusters (out of hundreds) generated by the algorithm using a state-based approach and finds 89 unique annotations for the top 5 clusters generated by the algorithm using a transition-based approach. So, trends here appear to indicate that the transition-based approach finds fewer annotations in general and that biologically motivated algorithms find many annotations. Further investigation should be performed to identify which is a preferable outcome, and more importantly, which outcome is most meaningful to the biomedical researcher.

Finally, we examine some of the Gene Ontology annotations found in the SPICi results to identify if the terms had any biological relevance. Two GO terms in particular were found to have both consistency and relevance: *canonical Wnt signaling pathway* and *interleukin-33 mediated signaling pathway*. In the SPICi results, *canonical Wnt signaling pathway* was found in the state-based Aged approach and *interleukin-33 mediated signaling pathway* was found in the state-based Middle approach. Both of these terms have both depth and specificity in the Gene Ontology biological process tree (Figure 2). While it may be exciting to investigate the biological relationships of these terms to aging in the mouse hypothalamus, the sheer inconsistency with which these terms were found (spanning from Young to Middle to Aged to Y_M Lost approaches) is enough to draw concern. It appears that even within a biologically motivated network clustering algorithm, results can return extremely different and inconsistent results – not only with regard to cluster size and structure but to the "content" of clusters themselves.

TABLE 4. SUMMARY OF UNIQUE AND SIGNIFICANT GENE ONTOLOGY BIOLOGICAL PROCESS ANNOTATIONS IDENTIFIED BY EACH APPROACH.

Analytic Approach

State-based	Transition-based
110 unique annotations	89 unique annotations

IV. DISCUSSION

In this research, we identify a novel way to examine changes across states in correlation networks – a basic transition based approach to identifying changes in network structure. Our objective was to compare state-based network

analysis of correlation networks to transition-based network analysis and to determine if there is consistency of biological relevance between the two. Our results have found that the transition-based approach seems to identify more consistent results. These results have sparked interest in the future directions of performing a comprehensive assessment of biologically motivated clustering algorithms to determine how different or similar their results may be – based on these very preliminary results, we would caution the "black-box" type of approach to network analysis where structural identification approaches such as clustering algorithms are chosen by convenience or familiarity.

REFERENCES

[1] Brohee S, van Helden J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*. 2006;7:488. doi: 10.1186/1471-2105-7-488.

[2] Dempsey K, Ali H. Evaluation of essential genes in correlation networks using measures of centrality. . 2011:509-515. http://dx.doi.org/10.1109/BIBMW.2011.6112421. doi: 10.1109/BIBMW.2011.6112421.

[3] Newman ME. Assortative mixing in networks. *Phys Rev Lett*. 2002;89(20):208701.

[4] Soltis AR, Saucerman JJ. Robustness portraits of diverse biological networks conserved despite order-of-magnitude parameter uncertainty. *Bioinformatics*. 2011;27(20):2888-2894. doi: 10.1093/bioinformatics/btr496; 10.1093/bioinformatics/btr496.

[5] Dempsey K, Ali H. On the robustness of the biological correlation network model. *International Conference on Bioinformatics Models, Methods and Algorithms (BIOINFORMATICS 2014)*. 2014:186-195.

[6] Albert R. Scale-free networks in cell biology. *J Cell Sci*. 2005;118(Pt 21):4947-4957. doi: 118/21/4947 [pii].

[7] Altaf-Ul-Amin M, Afendi FM, Kiboi SK, Kanaya S. Systems biology in the context of big data and networks. *BioMed Research International*. 2014;2014.

[8] Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet*. 2004;5(2):101-113. doi: 10.1038/nrg1272.

[9] Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512. doi: 7898 [pii].

[10] Costanzo M, Baryshnikova A, Bellay J, et al. The genetic landscape of a cell. *Science*. 2010;327(5964):425-431. doi: 10.1126/science.1180823 [doi].

[11] Fujimoto RM, Perumalla K, Park A, Wu H, Ammar MH, Riley GF. Large-scale network simulation: How big? how fast? . 2003:116-123.

[12] Girvan M, Newman ME. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*. 2002;99(12):7821-7826.

[13] Oldham MC, Langfelder P, Horvath S. Network methods for describing sample relationships in genomic datasets: Application to huntington's disease. *BMC Syst Biol*. 2012;6:63-0509-6-63. doi: 10.1186/1752-0509-6-63; 10.1186/1752-0509-6-63.

[14] Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512. doi: 7898 [pii].

[15] Horvath S, Dong J. Geometric interpretation of gene coexpression network analysis. *PLoS Comput Biol*. 2008;4(8):e1000117. doi: 10.1371/journal.pcbi.1000117; 10.1371/journal.pcbi.1000117.

[16] Dempsey K, Ali H. On the discovery of cellular subsystems in gene correlation networks using measures of centrality. *Current Bioinformatics*. 2013;8(3):305-314.

[17] Pfitzner R, Scholtes I, Garas A, Tessone CJ, Schweitzer F. Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Phys Rev Lett*. 2013;110(19):198701.

[18] Zanetti MS, Scholtes I, Tessone CJ, Schweitzer F. Categorizing bugs with social networks: A case study on four open source software communities. . 2013:1032-1041.

[19] GNU. R. http//r-project.org

[20] Csardi G, Nepusz T. The igraph software package for complex network research. *InterJournal, Complex Systems*. 2006;1695(5):1-9.

[21] Jiang P, Singh M. SPICi: A fast clustering algorithm for large biological networks. *Bioinformatics*. 2010;26(8):1105-1111. doi: 10.1093/bioinformatics/btq078 [doi].