2011

A noise reducing sampling approach for uncovering critical properties in large scale biological networks

Karthik Duraisamy
University of Nebraska at Omaha

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

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

Sanjukta Bhowmick
University of Nebraska at Omaha, sbhowmick@unomaha.edu

A Noise Reducing Sampling Approach for Uncovering Critical Properties in Large Scale Biological Networks

K. Duraisamy, K. Dempsey, H. Ali and S. Bhowmick
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE 68182, USA
{hali,sbhowmick}@unomaha.edu

ABSTRACT

A correlation network is a graph-based representation of relationships among genes or gene products, such as proteins. The advent of high-throughput bioinformatics has resulted in the generation of volumes of data that require sophisticated in silico models, such as the correlation network, for in-depth analysis. Each element in our network represents expression levels of multiple samples of one gene and an edge connecting two nodes reflects the correlation level between the two corresponding genes in the network according to the Pearson correlation coefficient. Biological networks made in this manner are generally found to adhere to a scale-free structural nature, that is, it is modular and adheres to a power-law degree distribution. Filtering these structures to remove noise and coincidental edges in the network is a necessity for network theorists because unfortunately, when examining entire genomes at once, network size and complexity can act as a bottleneck for network manageability.

Our previous work demonstrated that chordal graph based sampling of network results in viable models. In this paper, we extend our research to investigate how different orderings affect the results of our sampling, and maintain the viability of resulting network structures. Our results show that chordal graph based sampling not only conserves clusters that are present within the original networks, but by reducing noise can also help uncover additional functional clusters that were previously not obtainable from the original network.

KEYWORDS: bioinformatics, biological properties, chordal graphs, correlation networks, graph theory, noise reduction, parallel algorithms.

1. INTRODUCTION

The recent explosion of data in the biomedical research has provided us the opportunity to discover new mechanisms behind aging and disease. A popular method for analyzing this data is based on modeling information as networks. One particular set of networks, the correlation network, represents a set of genes and gene products whose expression is measured at certain time environments, such as diseased and normal states.

The focus of current analyses in correlation networks is based on discovering certain structural properties. For example, high-degree nodes generally represent genes that are key to network robustness and thus are essential for organism survival. Clusters of genes in a protein interaction network have been known correspond to key components of protein complexes. Therefore, combinatorial algorithms are extensively used to discover how structural properties of the network affect the well-being of the organism *in vivo*.

Correlation networks are generated from probes spanning the coding sequence of an entire genome and are therefore very large and complex. For instance, a complete network made from 40k inputs will produce a model with over 800 million edges. Analyzing these networks is a computationally expensive task. In order to efficiently explore this deluge of data, it is imperative to sample the network and obtain a reduced data intensive representation while maintaining key network structures.

Most sampling methods focus only on retaining important combinatorial structures of the network, such as the high-degree nodes or hubs, the clustering coefficients, or the number of cliques. However, not all edges in the network accurately represent genuine gene correlation. For example, two nodes may have a high correlation not because they are co-expressed, but because they have a

common neighbor. Both nodes may be co-expressed with that common neighbor but under different environments or controls, thus an edge is falsely drawn between them. In order to avoid this and obtain accurate analysis, we require a sampling method that not only retains important functional relationships that form key structures in the network in form of clusters, but can improve the quality of clusters as compared to the original network by reducing the noise in the network.

In our earlier work [14], we demonstrated that chordal graph based sampling conserves the important clusters in correlation networks. We also observed that in some cases, functionalities that were not found in the original network were identified in the sampled graph. This phenomenon motivated us to investigate and compare the effectiveness of different ordering strategies of chordal subgraphs for removing noise. In this paper, we apply an extensive search of the Gene Ontology database on larger networks (over 40,000 vertices and 200,000 edges) and compare chordal graph sampling based on Breadth First Search (BFS) and Reverse Cuthill Mckee (RCM) [2] ordering. Our experiments show that the sampled graphs, in particular those with RCM ordering, preserves important functionalities and removes some large clusters which occur in the original network but do not have any cohesive location in the ontology tree. These clusters, though combinatorially valid, do not map to any specific functionality. Thus our sampling technique can provide greater insight to data interpretation beyond combinatorial clustering techniques.

This paper is arranged as follows. In the background section we provide an introduction on how correlation networks are generated. We also briefly describe some important properties of chordal graphs and why they are suitable for sampling. In Section 3, we present our main contribution, a noise reducing sampling scheme for large complex networks. In Section 4, we describe our experiment design and provide results on networks based on mouse hypothalamus. We conclude with a discussion of our future research plans.

2. BACKGROUND

A correlation network is a graph model, where nodes represent genes and a set of sample expression levels for that gene, and an edge represents the level of correlation between two genes. Different measurements of correlation have been used to build these networks, such as the partial correlation coefficient [3], the Spearman correlation coefficient [4], or more commonly, the Pearson correlation coefficient [ref]. The network built from a dataset where all nodes (genes) are connected to each other is called the complete network, K, where n = the number of nodes/genes in the network. In K_n network, the number of edges is equal to $n*(n-1)/2$; this implies that in the case of datasets with a large number of genes, analysis of the K_n network is computationally and algorithmically taxing; thus, thresholding is a common method used for network reduction.

There are many methods for thresholding the correlation network. The most straightforward involves removing edges with a low correlation. In a network created using the Pearson correlation coefficient, this would mean removing edges at and around 0.00. In the larger of these networks, this threshold will need to become more and more stringent as the number of edges gets larger in order to maintain a size of network that can be quickly and properly analyzed. We use a threshold of ±0.70 to ±1.00 based on the fact that the coefficient of determination for these correlations will be at least 0.49. This determination threshold is chosen because of the indication that correlations remaining within the network will represent genes whose expression levels can be described as approximately 50% dependent on each other's expression. Carter *et al*. 2004 [5] used this method of "hard" thresholding by correlation level and used a p-value ≤ 0.0001 threshold to ensure that only significant correlations had been retained.

When one examines the log/log representation of the node degree distribution in a filtered correlation network, it follows a linear pattern associated with the power-law distribution that indicates a scale-free network structure [6]. Adherence to this distribution indicates that there are many nodes in the network that are poorly connected and a few nodes that are very well connected; these nodes are known as "hubs". Hubs have been found in multiple biological networks to correspond to essential genes [7] as well as being a characteristic structure of this particular type of network. Other properties have been found to be important within the scale-free network structure, such as a low clustering coefficient [6] indicating that the network has the tendency to form modules. These structures can be found by applying graph theoretic algorithms on the network and more importantly, can be found without the help of extra data such as the inclusion of biological attributes within the network. Thus, the method that finds structures within the network and later sorts noise from causative structures with true cellular function lends itself toward a higher impact result. However, this is only plausible for smaller networks. The issue remains that networks built from microarray data are too large for current structure finding algorithms to find clusters and modules in reasonable time (even with parallel computing resources at one's disposal), creating the need for more powerful analysis tools, and/or the ability to filter the network further.

Recently, several research papers [8,9] have explored the use of machine learning techniques to reduce noise in biological data. This work focuses on using supervised learning techniques to create a decision model based on prior information. Graph sampling methods [10,11,12], however, are generally used to obtain representative samples of the original network, rather than to remove noise. Our sampling technique focuses on identifying densely connected portions of the network, by extracting the maximal chordal subgraphs. Chordal graphs are graphs where the length of a cycle is not more than three [2]. Chordal subgraphs therefore include the highly connected portions of the network, such as cliques. Finding the maximum chordal subgraph is a NP-hard problem. In our implementation, we use a polynomial heuristic to find maximal chordal subgraphs from [13].

3. CHORDAL GRAPH BASED SAMPLING

Advances in high-throughput assays within the bioinformatics domain have resulted in high yield of massive datasets. Experimental technologies can measure gene expression of multiple gene products and isoforms across an entire genome. Analyses of these datasets are typically based on statistical analysis and comparison of each gene as an individual element with techniques such as Gene Set Enrichment Analysis (GSEA). Correlation network models allow for the modeling of *relationships* rather than individual elements, resulting in the desired ability to identify and isolate sets of genes responsible for observed functions. The size and complexity of these networks, however, remains a problem because current network analyses typically cannot handle large networks.

In our earlier work [14], we introduced a sampling technique based on extracting nearly chordal (quasi-chordal) subgraphs of the network in parallel. We demonstrated that the maximal chordal subgraphs contain almost all the high density subgraphs in the original graph. Since chordal graphs do not include large chordless cycles, maximal chordal subgraphs of a given graph are likely to conserve dense neighborhood relationships and filter out relationships in sparse areas of the graph. Hence, chordal subgraphs of a given network are likely to include highly connected regions of the network such as cliques and therefore in most cases can preserve the important structural properties. This can potentially reduce the impact of having false edges added to the network due to noise and highlight the presence of important clusters that could not be detected in the original network.

We therefore anticipate that a proper sampling technique can both limit the size as well as reduce the noise of the network. As a result, we propose a new data analysis technique by obtaining a carefully selected maximal subgraph of the original graph that represents the correlation network. We propose a new approach for sampling networks by removing noise contributes to the field of network analysis in two key ways: first, by creating speedup by parallel computation and filtering, and second, by preservation and improvement of biological functions of network structures. This method of sampling graphs for noise reduction conserves the properties from the original network while removing noise that is notorious in the typical correlation network. This leads to our key hypothesis below:

Hypothesis H_0*:* Given a graph G representing a correlation network obtained from bioinformatics expression data, a maximal chordal subgraph G1 of G will maintain most of the highly dense subgraphs of G while excluding edges representing noise-related relationships in the network. In particular, G1 will have the following properties:

H_{0a} - Key functional properties found in the clusters of unfiltered networks G are maintained in the sampled networks G1; and

H_{0b} - New clusters with biological function are uncovered. The identification of this novel function is revealed because functional attributes previously lost in noise can now be identified.

Our experimental results in the following section prove that both these hypothesis are true, and are particularly effective in the case sampling using RCM ordering. The basic algorithm [13] for identifying the maximal chordal subgraph is based on growing the graph from a starting vertex and adding edges so long as they maintain the chordal characteristics. Therefore ordering of the vertices, as well as the number of partitions in the parallel algorithm, play a significant role in determining the size and quality of the maximal quasi-chordal graph. In this paper we present a comparison between quasi-chordal graphs with vertices ordered using breadth first search (BFS) and reverse Cuthill Mckee (RCM) [2].

BFS ordering is based on a level by level traversal of the graph, where the level of a vertex is its shortest distance from the starting vertex. The gene correlations networks are often formed of disconnected components, and accessing the vertices using BFS assures that the vertices in the same connected graph component will be processed together.RCM ordering, in addition to accessing connected components, ensures that closely connected groups of vertices are placed together. That is the temporal access pattern of the vertices is based not only on whether they are in the same component but is also

proportional to how closely they are connected to each other. RCM ordering is implemented by reversing the vertex order obtained from a BFS search, with the constraint that the starting vertex is a peripheral vertex [2]. We believe that RCM ordering will be particularly suitable for obtaining maximal chordal subgraphs because (i) ordering closely connected vertex groups together, therefore will result in a greater probability that more modular portions of the networks will be included within the quasi-chordal subgraph and (ii) in case of parallel implementation of our sampling method, RCM ordering reduces the number of edges across partitions. This helps in lowering the communication and also in improving the result of the quasi-chordal graph detection by reducing the number of larger cycles (length>3).

4. EMPIRICAL RESULTS

4.1 Test Suites

Datasets were downloaded in January 2011 from NCBI's Gene Expression Omnibus, a publicly-available repository of high-throughput gene expression data. Datasets used were derived from series GSE8150, which was originally devised to identify the impact of anti-inflammatory elements on the young and mouse whole brain. The subsets of these series used included 5 samples of young mouse whole brain, and 5 samples of old mouse whole brain. Two subsets, which were mice treated with anti-inflammatories, were not used. Networks were created for the young dataset and the old dataset using the methods described previously, with a correlation of 0.96 to 1.00, and p-value ≤ 0.005.

4.2 Experimental Design

Our objective was to obtain strongly connected portions on the network through identifying the maximal chordal subgraphs. Given the large size of the network, we implemented a parallel algorithm as follows: We divided the network across P processors. Within each partition, we obtained the local maximal chordal subgraph formed only of the edges whose endpoints lie completely within the processor. The edges that lie across processors were included only if two border edges with a common vertex, combined with a previously marked chordal edge to form a triangle. This implementation generated quasi-chordal subgraphs, since we did not check whether the inclusion of border edges increase the length of any cycle by more than three. A detailed description of the method can be found in [14].

The scalability of our algorithm was limited by the size of the networks. The networks exhibited good speedup from 2-8 processors, but as we moved to 16 and 32 processors, the speedup deteriorated due to increase in edges that fell across processors, which in turn lead to increase in communication. As expected, implementations RCM ordering executed faster that those using BFS, which indicates that a more appropriate graph partitioning strategy would help improve the scalability. However, the focus of this paper is more on improving sampling than on improving the performance. Therefore due to the space constraints, we will address the aspects of scalability in a future work.

We had observed in our earlier work, that chordal subgraphs conserve the common functionalities of the original network. Our goal in this paper was to observe how many **new** functionalities, beyond those found in the original network, can be discovered through sampling. For each network we obtained 12 different samples based on two different orderings (BFS and RCM) and 6 different partitions (on 1, 2, 4, 8, 16 and 32 processors).

We then clustered the data using AllegroMCODE which uses a node weighting algorithm to identify tightly connected groups of genes within the network and ranks them according to their novel clustering score. We ran the AllegroMCODE algorithm on each network under default settings and took the top 5 clusters from each network. After performing Gene Ontology Enrichment, we present the results from each cluster in Figures 1-4.

4.3 Analysis of Results

Here we provide the results of our analysis using clusters from the original networks and the sampled networks using BFS and RCM ordering. We highlight that original cluster key functions were in some cases maintained, or new cluster function was uncovered with noise removal using our graph sampling technique.

Each column in the Table denotes clusters and corresponding enrichment score found in the original network and through sampled networks on 1, 2, 4, 8, 16 and 32 processors respectively. Enrichment for a Gene Ontology term can be described as the ratio of the number of genes in the cluster with the specified term (c) to the number of genes in the cluster (n), divided by the ratio of the number of genes in the entire genome with the specified term (C) to the total number of genes in the tested genome (N). The formal equation to identify enrichment, then, is $E = (c/n)/(C/N)$. The higher the enrichment score, the better. Using this Gene Ontology enrichment, most of the genes in the same cluster can be identified as having the same functionality. We verified

the Gene Ontology classification of original clusters by filtering results of analysis to p-val < 0.005 and compared them to the GO classifications of the top clusters found in each 6 networks per dataset. A detailed description of the results is given as follows.

The young mouse data, sampled using RCM ordering (Figure 1) preserved the Metabolism enriched cluster (cluster 6) from the original network (for sampling in one processor and two processors). New clusters identified were enriched in transport (cluster 2), metabolism (cluster 1 and 3), and development (cluster 4). Compared to the BFS results (Figure 2), these results were more functionally specific, suggesting that RCM may retain knowledge better than BFS.

In the young mouse dataset, the original network had 2 of the top clusters enriched in with GO terms associated with Development and Transport. Clusters matching to these functionalities were also found in the sampling method using BFS ordering (Figure 2). The BFS results identified the Development cluster (cluster 3) for each number of partitions (1, 2, 4, 8, 16, and 32) whereas the Transport cluster (cluster 5) was only identified on the sample using one processor. The BFS method also helped in discovery of new clusters which were enriched in metabolism (cluster 1), development (clusters 2 and 3), and transport (cluster 4).

For the middle aged mouse network, sampling using the BFS ordering (Figure 3) identified only clusters enriched in transport and localization (clusters 2 and 6) in the original network; however the sampling results identified other new clusters rich in Immune Defense (clusters 3, 4 and 5), Cell Communication (cluster 7), and Cell Division (cluster 1). The uniformity of novel clusters was not as conserved as novel clusters identified in the young method, which may be expected from a more aged network.

In the case of the RCM ordering (Figure 4), clusters enriched in Development (clusters 1 and 5), Cell Cycle Metabolism (cluster 2), and Homeostasis (cluster 3) were conserved from the original network. These clusters were conserved for the majority by sampling using 1 processor; novel clusters identified included those enriched in Defense and Immune Response (clusters 4, 6, and 7).
Our results indicate that RCM had more matches to original GO clusters identified, indicating that lowering the bandwidth of the corresponding matrix can help in obtaining more clustered regions. Additionally, both methods performed exceptionally at identifying novel clusters within networks, which indicates that sampling based on identifying quasi chordal subgraphs can indeed eliminate poorly connected edges, which form noise in the network. The two methods together identified methods identified around 20 novel clusters, but the RCM method had higher conservation of novel cluster identification than BFS across number of partitions, suggesting it may be more stable than the BFS method.

5. CONCLUSIONS

Our analysis has shown that a correlation network obtained from bioinformatics expression data, a maximal chordal subgraph will maintain or improve upon the biological information contained within the highly dense subgraphs. By excluding edges representing noise-related relationships in the network, we identify a sampled network that has fewer edges and where functional properties found in the clusters of unfiltered networks G are maintained in the sampled networks or new clusters with biological function are uncovered. The identification of this novel function is achieved through noise removal.

These results indicate that while both methods of ordering in our parallel graph sampling method are useful in removing noise, the RCM method retains better cluster functionality from the original data and also finds a number of novel clusters. This is important in the continuing search for a method to reduce network size and noise while retaining important structural information, thus maintaining functional properties of each individual network. In the future we plan to investigate the impact of implementing other methods for reducing noise in the correlation network, such as identifying Steiner trees or hypergraphs.

ACKNOWLEDGEMENT

This publication was made possible by Grant Number P20 RR16469 from the NCRR, a component of the NIH. The authors would also like to thank the Nebraska EPSCoR First Award and the College of IS&T, University of Nebraska at Omaha for their funding and support.

REFERENCES

[1] T. Barrett, D. B. Troup, S. E. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. F. Kim, A. Soboleva, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, R. N. Muertter, R. Edgar, "Ncbi geo: archive for high-throughput functional genomic data," Nucleic Acids Res.(Database issue) Jan;37.

[2] J. L. Gross, J. Yellen, HANDBOOK OF GRAPH THEORY AND APPLICATIONS, CRC Press, 2004.

[3] N. S. Watson-Haigh, H. N. Kadarmideen, A. Reverter, Pcit: "An R package for weighted gene co-expression networks based on partial correlation and information theory

approaches," *Bioinformatics*(Oxford, England) 26 (3) (2010) 411–413.

[4] W. J. Ewens, G. R. Grant, STATISTICAL METHODS IN BIOINFORMATICS (Second Edition ed.), New York, NY: Springer, 2005.

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

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

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

[8] A. Miranda, . Garcia, A. Carvalho, A. Lorena, "Use of classification algorithms in noise detection and elimination," in: E. Corchado, X. Wu, E. Oja, A. Herrero, B. Baruque (Eds.), HYBRID ARTIFICIAL INTELLIGENCE SYSTEM, Vol. 5572 of Lecture Notes in Computer Scie nce, 2009, pp. 417–424.

[9] G. L. Libralon, A. C. P. d. L. F. d. Carvalho, A. C. Lorena, "Preprocessing for noise detection in gene expression classification data," *Journal of the Brazilian Computer Society* 15 (2009) 3 – 11.

[10] J. Leskovec, J. Kleinberg, C. Faloutsos, "Graphs over time: densification laws, shrinking diameters and possible explanations" in: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005, pp. 177–187.

[11] J. Leskovec, C. Faloutsos, "Sampling from large graphs," in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '06, 2006, pp. 631–636.

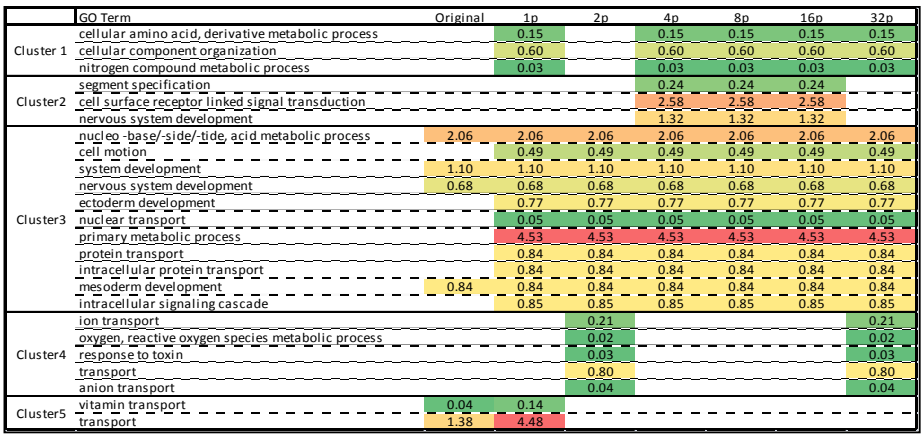
[12] A. Rasti, M. Torkjazi, R. Rejaie, N. Duffield, W. Willinger, D. Stutzbach, in: Respondent-Driven Sampling for Characterizing Unstructured Overlays, 2009, pp. 2701–2705.

[13] P. M. Dearing, D. R. Shier, D. D. Warner, "Maximal chordal subgraphs," *Discrete Applied Mathematics* 20 (3) (1988) 181 – 190.

[14] K. Dempsey, K. Duraisamy, H. Ali, S. Bhowmick, "A parallel graph sampling algorithm for analyzing gene correlation networks" in: ICCS 2011, 2010.

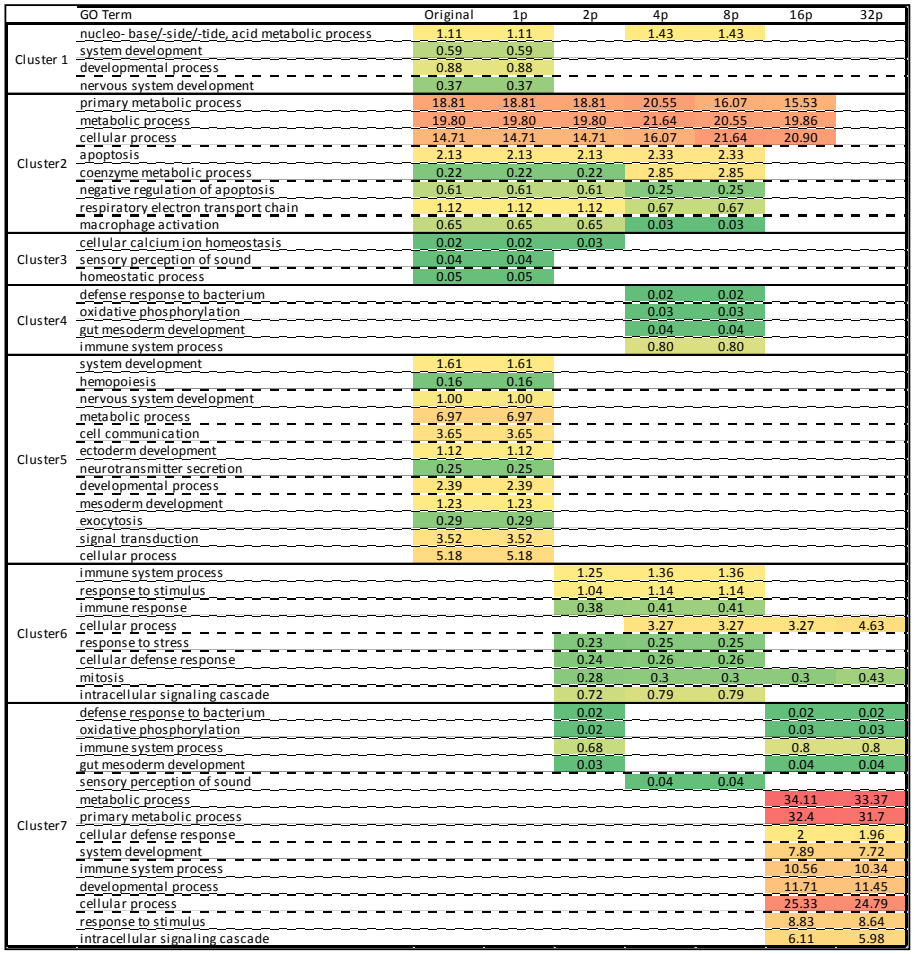
	GO Term	Original	1p	2p	4p	8p	16p	32p
Cluster 1	signal transduction		1.86	2.04	1.86	1.86	1.86	
	cell communication		1.92	2.11	1.92	1.92	1.92	
	cell surface receptor linked signal transduction		1.03	1.14	1.03	1.03	1.03	
	cellular process		2.72			2.72	2.72	
	sulfur metabolic process		0.04	0.05	0.04	0.04	0.04	
	phosphate metabolic process					0.09	0.09	
	sensory perception					0.44	0.44	
	immune system process					1.14	1.14	
	phosphate transport					0.03	0.03	
Cluster2	nuclear transport		0.02	0.02	0.02			
	response to interferon-gamma		0.03	0.03	0.03			
Cluster3	cellular amino acid,derivative metabolic process					0.12	0.15	0.15
	cellular component organization					0.48	0.60	0.60
	nitrogen compound metabolic process					0.02	0.03	0.03
	cellular process					2.18		
	cell communication					1.54		
	complement activation					0.05		
Cluster4	regulation of phosphate metabolic process		0.00	0.00	0.00			
	skeletal system development		0.10	0.10	0.10			
	muscle organ development		0.10	0.10	0.10			
	tricarboxylic acid cycle		0.01	0.01	0.01			
	cellular calcium ion homeostasis		0.01	0.01	0.01			
	homeostatic process		0.03	0.03	0.03			
	mesoderm development		0.32	0.32	0.32			
	phosphate metabolic process		0.04	0.04	0.04			
Cluster5	mammary gland development					0.02	0.04	
	dorsal/ventral axis specification					0.03	0.05	
Cluster6	protein metabolic process	3.34	2.13	0.76				
	metabolic process	8.07	5.13	1.83				
	primary metabolic process	7.66		1.74				

Figure 1: The gene functionality of clusters for the young mouse network with RCM ordering. Enrichment scores are colored from low (green) to high (red). Spaces with no enrichment mean that for that number of partitions, there was no cluster found for that partition. Number of conserved clusters: 1. Number of new clusters in sampled networks: 6.



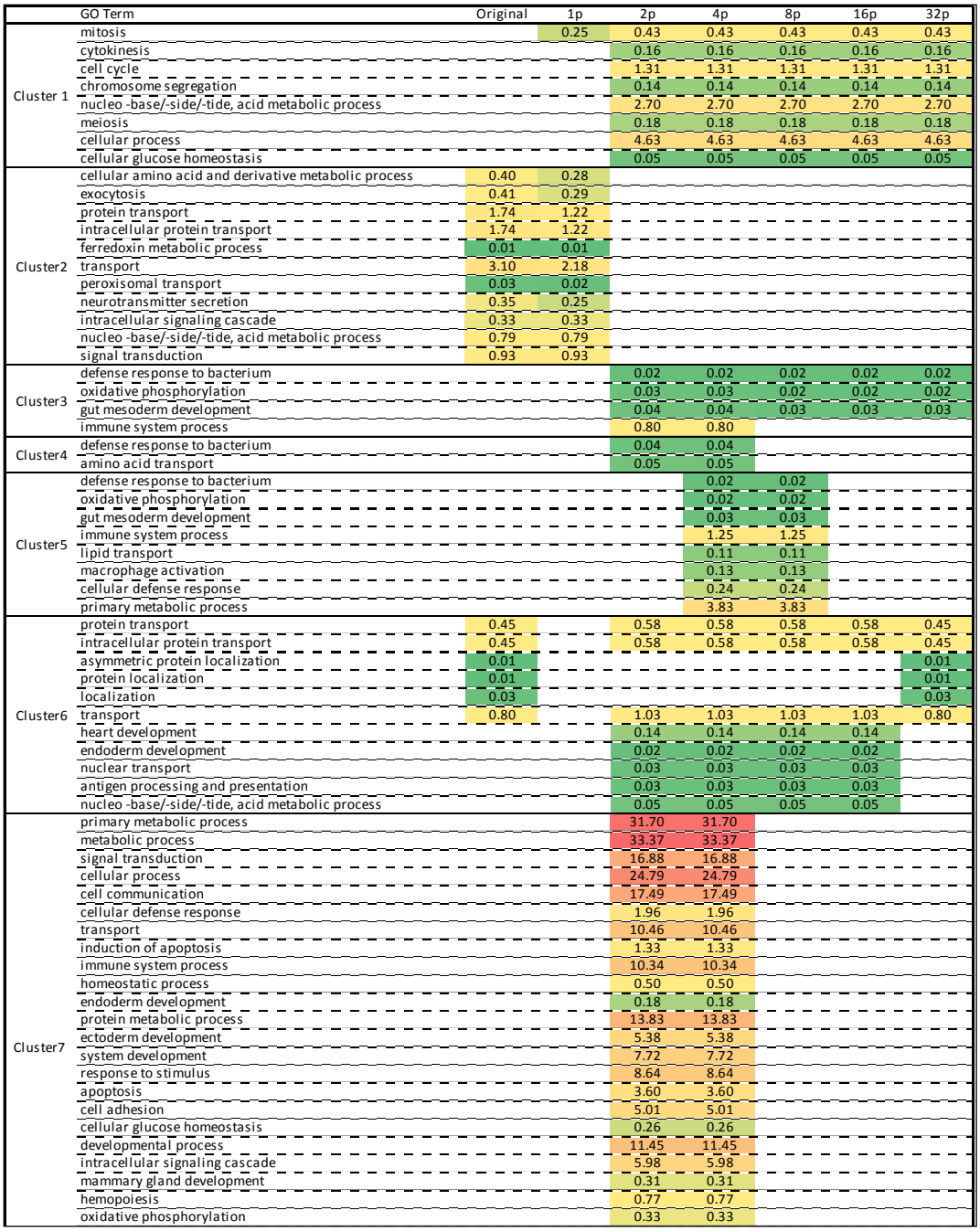
	GO Term	Original	1p	2p	4p	8p	16p	32p
Cluster 1	cellular amino acid, derivative metabolic process		0.15		0.15	0.15	0.15	0.15
	cellular component organization		0.60		0.60	0.60	0.60	0.60
	nitrogen compound metabolic process		0.03		0.03	0.03	0.03	0.03
Cluster2	segment specification				0.24	0.24	0.24	
	cell surface receptor linked signal transduction				2.58	2.58	2.58	
	nervous system development				1.32	1.32	1.32	
Cluster3	nucleo -base/-side/-tide, acid metabolic process	2.06	2.06	2.06	2.06	2.06	2.06	2.06
	cell motion		0.49	0.49	0.49	0.49	0.49	0.49
	system development	1.10	1.10	1.10	1.10	1.10	1.10	1.10
	nervous system development	0.68	0.68	0.68	0.68	0.68	0.68	0.68
	ectoderm development		0.77	0.77	0.77	0.77	0.77	0.77
	nuclear transport		0.05	0.05	0.05	0.05	0.05	0.05
	primary metabolic process		4.53	4.53	4.53	4.53	4.53	4.53
	protein transport		0.84	0.84	0.84	0.84	0.84	0.84
	intracellular protein transport		0.84	0.84	0.84	0.84	0.84	0.84
	mesoderm development	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	intracellular signaling cascade		0.85	0.85	0.85	0.85	0.85	0.85
Cluster4	ion transport			0.21				0.21
	oxygen, reactive oxygen species metabolic process			0.02				0.02
	response to toxin			0.03				0.03
	transport			0.80				0.80
	anion transport			0.04				0.04
Cluster5	vitamin transport	0.04	0.14					
	transport	1.38	4.48					

Figure 2: The gene functionality of clusters for the young mouse network with BFS ordering. Enrichment scores are colored from low (green) to high (red). Spaces with no enrichment mean that for that number of partitions, there was no cluster found for that partition. Number of conserved clusters: 1. Number of clusters with additional genes: 1. Number of new clusters in sampled networks: 3.



	GO Term	Original	1p	2p	4p	8p	16p	32p
Cluster 1	nucleo- base/-side/-tide, acid metabolic process	1.11	1.11		1.43	1.43		
	system development	0.59	0.59					
	developmental process	0.88	0.88					
	nervous system development	0.37	0.37					
Cluster2	primary metabolic process	18.81	18.81	18.81	20.55	16.07	15.53	
	metabolic process	19.80	19.80	19.80	21.64	20.55	19.86	
	cellular process	14.71	14.71	14.71	16.07	21.64	20.90	
	apoptosis	2.13	2.13	2.13	2.33	2.33		
	coenzyme metabolic process	0.22	0.22	0.22	2.85	2.85		
	negative regulation of apoptosis	0.61	0.61	0.61	0.25	0.25		
	respiratory electron transport chain	1.12	1.12	1.12	0.67	0.67		
	macrophage activation	0.65	0.65	0.65	0.03	0.03		
Cluster3	cellular calcium ion homeostasis	0.02	0.02	0.03				
	sensory perception of sound	0.04	0.04					
	homeostatic process	0.05	0.05					
Cluster4	defense response to bacterium				0.02	0.02		
	oxidative phosphorylation				0.03	0.03		
	gut mesoderm development				0.04	0.04		
	immune system process				0.80	0.80		
Cluster5	system development	1.61	1.61					
	hemopoiesis	0.16	0.16					
	nervous system development	1.00	1.00					
	metabolic process	6.97	6.97					
	cell communication	3.65	3.65					
	ectoderm development	1.12	1.12					
	neurotransmitter secretion	0.25	0.25					
	developmental process	2.39	2.39					
	mesoderm development	1.23	1.23					
	exocytosis	0.29	0.29					
	signal transduction	3.52	3.52					
	cellular process	5.18	5.18					
Cluster6	immune system process			1.25	1.36	1.36		
	response to stimulus			1.04	1.14	1.14		
	immune response			0.38	0.41	0.41		
	cellular process				3.27	3.27	3.27	4.63
	response to stress			0.23	0.25	0.25		
	cellular defense response			0.24	0.26	0.26		
	mitosis			0.28	0.3	0.3	0.3	0.43
	intracellular signaling cascade			0.72	0.79	0.79		
Cluster7	defense response to bacterium			0.02			0.02	0.02
	oxidative phosphorylation			0.02			0.03	0.03
	immune system process			0.68			0.8	0.8
	gut mesoderm development			0.03			0.04	0.04
	sensory perception of sound				0.04	0.04		
	metabolic process						34.11	33.37
	primary metabolic process						32.4	31.7
	cellular defense response						2	1.96
	system development						7.89	7.72
	immune system process						10.56	10.34
	developmental process						11.71	11.45
	cellular process						25.33	24.79
	response to stimulus						8.83	8.64
	intracellular signaling cascade						6.11	5.98

Figure 3: The gene functionality clusters for the middle aged mouse network with BFS ordering. Enrichment scores are colored from low (green) to high (red). Spaces with no enrichment mean that for that number of partitions, there was no cluster found for that partition. Number of conserved clusters: 4 Number of new clusters in sampled networks: 7.



	GO Term	Original	1p	2p	4p	8p	16p	32p
Cluster 1	mitosis		0.25	0.43	0.43	0.43	0.43	0.43
	cytokinesis			0.16	0.16	0.16	0.16	0.16
	cell cycle			1.31	1.31	1.31	1.31	1.31
	chromosome segregation			0.14	0.14	0.14	0.14	0.14
	nucleo -base/-side/-tide, acid metabolic process			2.70	2.70	2.70	2.70	2.70
	meiosis			0.18	0.18	0.18	0.18	0.18
	cellular process			4.63	4.63	4.63	4.63	4.63
	cellular glucose homeostasis			0.05	0.05	0.05	0.05	0.05
Cluster2	cellular amino acid and derivative metabolic process	0.40	0.28					
	exocytosis	0.41	0.29					
	protein transport	1.74	1.22					
	intracellular protein transport	1.74	1.22					
	ferredoxin metabolic process	0.01	0.01					
	transport	3.10	2.18					
	peroxisomal transport	0.03	0.02					
	neurotransmitter secretion	0.35	0.25					
	intracellular signaling cascade	0.33	0.33					
	nucleo -base/-side/-tide, acid metabolic process	0.79	0.79					
	signal transduction	0.93	0.93					
Cluster3	defense response to bacterium			0.02	0.02	0.02	0.02	0.02
	oxidative phosphorylation			0.03	0.03	0.02	0.02	0.02
	gut mesoderm development			0.04	0.04	0.03	0.03	0.03
	immune system process			0.80	0.80			
Cluster4	defense response to bacterium			0.04	0.04			
	amino acid transport			0.05	0.05			
Cluster5	defense response to bacterium				0.02	0.02		
	oxidative phosphorylation				0.02	0.02		
	gut mesoderm development				0.03	0.03		
	immune system process				1.25	1.25		
	lipid transport				0.11	0.11		
	macrophage activation				0.13	0.13		
	cellular defense response				0.24	0.24		
	primary metabolic process				3.83	3.83		
Cluster6	protein transport	0.45		0.58	0.58	0.58	0.58	0.45
	intracellular protein transport	0.45		0.58	0.58	0.58	0.58	0.45
	asymmetric protein localization	0.01						0.01
	protein localization	0.01						0.01
	localization	0.03						0.03
	transport	0.80		1.03	1.03	1.03	1.03	0.80
	heart development			0.14	0.14	0.14	0.14	
	endoderm development			0.02	0.02	0.02	0.02	
	nuclear transport			0.03	0.03	0.03	0.03	
	antigen processing and presentation			0.03	0.03	0.03	0.03	
	nucleo -base/-side/-tide, acid metabolic process			0.05	0.05	0.05	0.05	
Cluster7	primary metabolic process			31.70	31.70			
	metabolic process			33.37	33.37			
	signal transduction			16.88	16.88			
	cellular process			24.79	24.79			
	cell communication			17.49	17.49			
	cellular defense response			1.96	1.96			
	transport			10.46	10.46			
	induction of apoptosis			1.33	1.33			
	immune system process			10.34	10.34			
	homeostatic process			0.50	0.50			
	endoderm development			0.18	0.18			
	protein metabolic process			13.83	13.83			
	ectoderm development			5.38	5.38			
	system development			7.72	7.72			
	response to stimulus			8.64	8.64			
	apoptosis			3.60	3.60			
	cell adhesion			5.01	5.01			
	cellular glucose homeostasis			0.26	0.26			
	developmental process			11.45	11.45			
	intracellular signaling cascade			5.98	5.98			
	mammary gland development			0.31	0.31			
	hemopoiesis			0.77	0.77			
	oxidative phosphorylation			0.33	0.33			

Figure 4: The gene functionality clusters for the middle aged mouse network with RCM ordering. Enrichment scores are colored from low (green) to high (red). Spaces with no enrichment mean that for that number of partitions, there was no cluster found for that partition. Number of conserved (or partially conserved) clusters: 4. Number of new clusters in sampled networks: 4