2013

On Mining Biological Signals Using Correlation Networks

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

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

Claudia Cortes
University of Nebraska at Omaha

Zack Eriksen
University of Nebraska at Omaha

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

On Mining Biological Signals using Correlation Networks

Kathryn Dempsey, Ishwor Thapa, Claudia Cortes, Zach Eriksen, Dhundy K. Bastola, and Hesham Ali*
College of Information Science and Technology, University of Nebraska at Omaha, Omaha, NE
*Contact email: hali@unomaha.edu

***Abstract*—Correlation networks have been used in biological networks to analyze and model high-throughput biological data, such as gene expression from microarray or RNA-seq assays. Typically in biological network modeling, structures can be mined from these networks that represent biological functions; for example, a cluster of proteins in an interactome can represent a protein complex. In correlation networks built from high-throughput gene expression data, it has often been speculated or even assumed that clusters represent sets of genes that are co-regulated. This research aims to validate this concept using network systems biology and data mining by identification of correlation network clusters via multiple clustering approaches and cross-validation of regulatory elements in these clusters via motif finding software. The results show that the majority (81-100%) of genes in any given cluster will share at least one predicted transcription factor binding site. With this in mind, new regulatory relationships can be proposed using known transcription factors and their binding sites by integrating regulatory information and the network model itself.**

Keywords—correlation networks, motif finding, transcription factor binding sites, clustering, mining biological signals

I. INTRODUCTION

Correlation networks provide a powerful tool for modeling the massive amounts of data available via high-throughput experimental assays. These models represent gene probes as nodes and the correlation of their expression patterns as edges with weights corresponding to strength. Biological networks have become a critical tool for the representation of increasing amounts of data since as early as 1999[1], when Barabasi *et al.* first linked structure to signal in the network model in various types of networks. Since then, many types of networks have arisen to model "big data," or data that is (1) massive in size, (2) spans multiple time points, (3) heterogeneous, and contains noise (4). For example, in protein-protein interaction networks, nodes with high degree ("hubs") have a 60% likelihood of corresponding to essential genes compared to a 20% likelihood of randomly chosen non-hub nodes[2]. Further, in these same models, proteins that display complete in-network connection (i.e. for any set of nodes, all connections possible between those nodes exist) are typically found in the cell as protein complexes that physically interact[3]. Similar studies that link gene essentiality to centrality[4,5] and discrete cellular function to clustering[6-9] have been performed in the correlation network.

Despite these extensive studies, it remains unclear what the function of clusters in a correlation network represent. If the network can theoretically be free of noise, a cluster of genes would be expected to represent a set of genes whose expression follows a similar general pattern, which is an inherent manifestation of co-regulation. Studies comparing networks from common origins with different experimental conditions (drug treatments or time points) have found that clusters do not typically show an overlap in gene expression patterns but for a few select genes. As such, it can be speculated that these networks capture but a snapshot of the cellular environment at a given time, and as transcription is inherently a transient and potentially quickly changing process, it stands to reason that co-regulation could be the cause of these dramatic changes in co-expression of genes in the correlation network. While this link between clusters and co-regulation has long been speculated, this link between structure and function has not yet been confirmed.

A. Hypothesis

As the correlation network is built from gene expression data, one would expect that adjacent nodes in the resulting network would share some correlation of expression, and therefore could possibly be co-regulated. The goal of this work is to validate or disprove this speculation. It has been found in previous work related to correlation networks that gene clusters or modules in correlation networks have common functions based on Gene Ontology (GO) information[6,10,11] suggesting that there is a common function and possibly a common regulator of these genes. As such we present the following hypothesis, *H_0: Given a cluster* C *from a correlation network* G*, the genes that form that cluster will be co-regulated by one or more transcription factors, and as such, we can extract novel transcription factor binding sites from network clusters with unknown regulatory mechanisms.* This hypothesis, if found to be true, should be *robust* to dataset type, clustering method, and transcription factor binding site software.

The overall approach of this study is highlighted in Figure 1. A correlation network is created from expression data, then clusters are extracted. Next, the upstream gene regions of the genes found in each cluster are extracted for use in motif finding. Then, these sets of upstream region sequences are run through pattern finding (transcription factor binding site finding) softwares to identify common motifs per cluster. Results are then assessed to determine (1) if there is a common

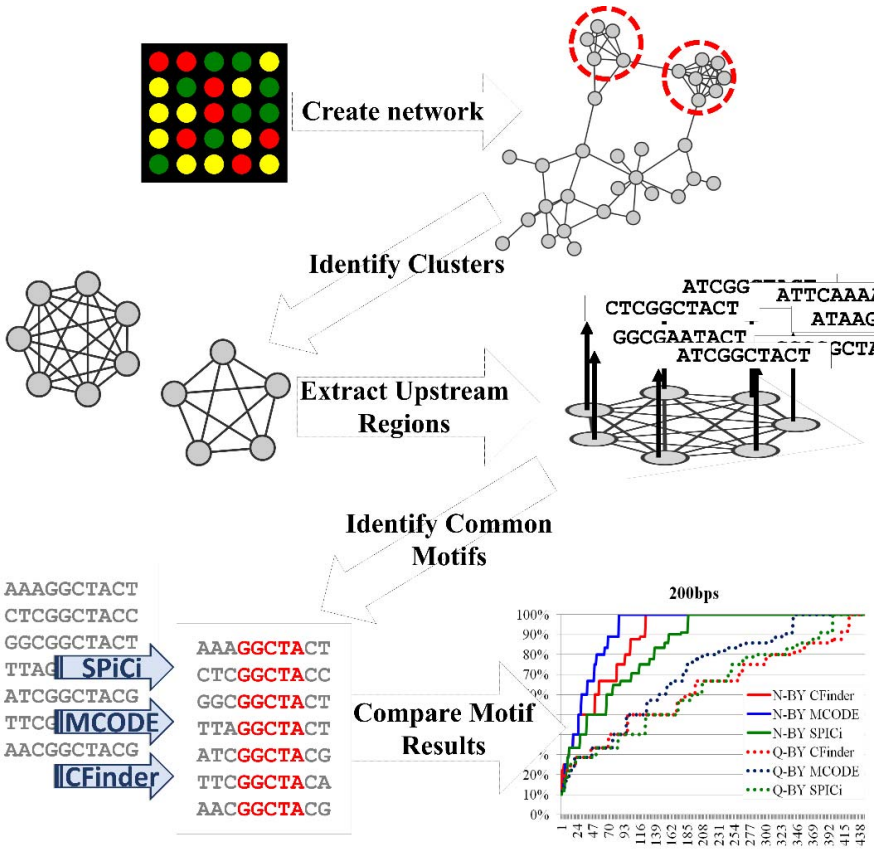

Figure 1. A flowchart of the process described in this research. First, networks are created by comparing gene expression values. Clusters are extracted and the upstream regions of the genes in those clusters are extracted for use in motif finding/pattern finding software. The results of these pattern finding methods are compared to determine if a pattern is common to the clusters, and if so, how many of the upstream regions contain those patterns.

motif per cluster and (2) how many of the genes in that cluster contain that motif.

II. METHODS

A. Data acquisition and network creation

All data for this work was downloaded from the National Center for Biotechnology Information's (NCBI) website using the Gene Expression Omnibus (GEO)[12]. As it is perhaps one of the most well-studied and understood model organisms, data from *Saccharomyces cerevisiae* was used in this study. A summary of the three datasets used in this study follows in Table 1, where column 1 represents the GEO Series number, column 2 represents what was compared in the assay, column 3 represents the yeast cell line used, column 4 represents the number of samples for the given experimental condition, column 5 represents the manufacturer, and column 6 represents the network ID that will be used throughout this work.

To create correlation networks, pair-wise Pearson Correlation Coefficient (PCC) was computed for each probe per each of the six experimental sets (Q-BY, N-BY, Q-S, N-S, I-0, and I-40). In the final correlation network, nodes represent genes/probes and edges connect two nodes if the PCC of their expression vectors fall within the $0.70 \leq \rho \leq 1.00$ threshold. Edges not meeting a statistical threshold of $P<0.0005$ using the student's T-test were thrown out. Network sizes are described in Table 2, where column 1 represents the network ID, column 2 represents the number of probes in the original experiment, column 3 represents the number of nodes in the filtered network, column 4 represents the number of edges in the filtered network, and column 5 represents the edge density of the filtered network, where *density* = (*number of edges*) / [(*n*(n-1))/2*] and *n* = the number of nodes. The I-0 and I-40 networks, despite having the largest amounts of nodes and edges, are the sparsest of the six.

TABLE I. EXPRESSION DATASET DESCRIPTIONS.

Series #	Experimental Description	Cell line	*N*	Man.	Net. ID
GSE8542	Quiescence	BY4742	10	Qiagen	Q-BY
	Non-quiescence	BY4742	10		N-BY
GSE8559	Quiescence	S288c	10	Qiagen	Q-S
	Non-quiescence	S288c	10		N-S
GSE46384	IPA at 0 time	BY4743	4	Agilent Tech.	I-0
	IPA at 40 time	BY4743	4		I-40

TABLE II. CORRELATION NETWORK SIZES

Network	Probes	Nodes	Edges	Edge Density
QUI_BY	6,979	2,543	5,363	0.166%
NON_BY	6,979	1,541	2,515	0.212%
QUI_S	6,967	3,434	9,483	0.161%
NON_S	6,967	1,945	2,186	0.116%
IPA_0	6,317	5,969	11,425	0.064%
IPA_40	6,317	5,903	16,640	0.096%

B. Network clustering

There are a number of different network clustering algorithms available in the biological realm; outside of the biological focus even more clustering approaches and variations are available. Among the most popular for systems biologist are MCODE[13], MCL[14], and CFinder[15]. MCODE[13] is a clustering approach that is lauded as a discovery tool, and was designed for identifying tightly connected groups of nodes in a protein-protein interaction network (i.e. those proteins likely involved in a complex). MCODE is available as a Cytoscape plug-in. The clusters it identifies are disjointed and are ranked according to cluster size and density. SPICi is presented as a fast clustering algorithm, also motivated by complex finding in protein-protein interaction networks (PPI's), that uses a greedy approach to finding clusters while maintaining density[16]. CFinder allows for cluster overlap and is based on clique percolation and has been found applicable not only in biological networks, but in social, metabolic, and similarity networks as well[15]. Comparison of CFinder and MCODE by Li *et al.* 2010 found that they were comparable in precision, accuracy, and identification of relevant complexes in multiple datasets[17]; SPICi is a self-proclaimed "fast" algorithm and was included as an algorithm that used a local greedy search. Each network was clustered using the following parameters (chosen by default, which is typically recommendation of the software provider):

- MCODE v.1.2[13]: Find clusters in *whole network*, loops not included, Degree cutoff (the minimum number of node connections) of 5, Haircut included (singly connected nodes removed), Node score (proprietary density and connection score) of 0.2, K-Core (size of the minimum clique) of 4, and Max. Depth (how many hops into the network to check) of 50.

- SPICi[16]: Find clusters with minimum density threshold of 0.5, minimum cluster size of 4 nodes, minimum support threshold of 0.5, and assuming a sparse network.

- CFinder[15]: Find clusters in an undirected network with a minimum k-clique size of 4.

After clustering, individual cluster files were parsed and gene ID's converted to yeast Open Reading Frame (ORF) name for promoter sequence preprocessing. Generally, CFinder and SPICi found the most clusters per network, MCODE found the least, and quiescent networks (those that represent a halted state of growth) tended to contain more clusters than non-quiescent networks (as shown in Figure 2).

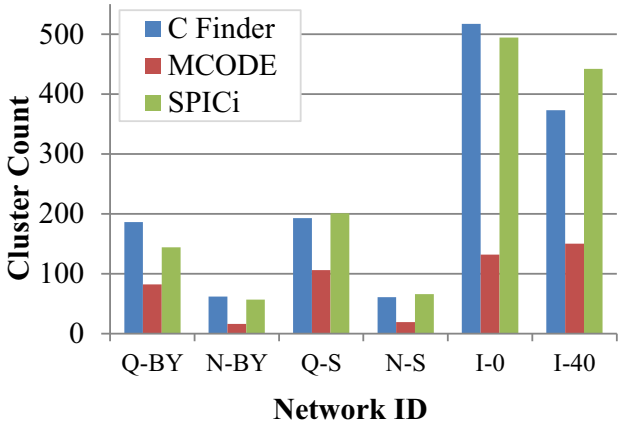

Fig. 2. Cluster sizes for each method by network. X-axis references the network ID described in Table 1; y-axis references the number of clusters found by the method (CFinder in blue, MCODE in red, and SPICi in green).

C. Sequence acquisition

The promoter or regulatory region of a gene is typically located upstream of the coding region of the gene that becomes the pre-mRNA. An example of the promoter region is shown in Figure 3. For each gene in each cluster, the promoter region was extracted at 50 base pairs (bp) and 200bp upstream of a given gene. Quest *et al.* 2008 performed an assessment of multiple transcription factor binding site tools and suggested that the smaller the window size upstream of the coding sequence, the better the motif or binding site prediction[18]. Promoter sequence extraction was performed using R scripting via Biomart (www.biomart.org)[19]. If no ORF name was available for the gene, that gene was not included in the promoter sequence. After extraction, files with less than 3 genes were excluded because 2 gene annotation pairs are common. These preprocessing steps resulted in a total of 1,917 files with sequences for SPICi, 732 files with sequences for MCODE, and 1,935 files with promoter sequences for CFinder.

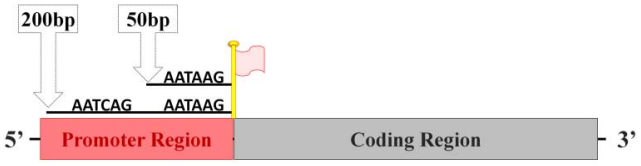

Fig. 3. An example of the layout of the promoter region of the gene. Reading the gene forward (5' to 3'), the promoter typically resides upstream (5') of the coding region (which becomes expressed as mRNA). The promoter sequences 50bp upstream of the gene coding region begins at the gene start site and runs 50 bases in the 5' direction. The same goes for 200bp upstream.

D. Motif finding

Five motif finding algorithms were used to identify potential transcription factor binding sites in the promoter regions of genes that shared common clusters. ELPH (v.1.0.1) uses a Gibbs sampling method designed for identifying patterns specifically in gene flanking regions, upstream or

downstream[20]. GLAM uses local alignment without gaps that models itself from Gibbs Sampling and uses BLAST scoring to enhance patterns found in sequences without having to input the motif length[21]. Our third chosen method, the MEME (v.4.9.0) algorithm allows for *de novo* pattern finding via expectation maximization[22]. The final program is Weeder (v.1.4.2), which uses background based on the organisms sequence and statistical analysis to identify and then hone patterns of interest[23]. To determine the presence of binding sites, upstream promoter regions were fed to each program under default parameters other than those described here: Not all sequences were required to contain the motif; the forward strand only was searched, no expected number of motif occurrences per sequence was given, and the length of the motif searched was equal to 10. If the program gave variable results (for example, Weeder selects motifs of lengths 6, 8, and 10), only motifs of length 10 were chosen. Results were parsed and standardized for comparison.

E. Method scoring

There is a limited amount of information that can be easily inferred from each file as each output is proprietary; this includes determining how many sequences from the original file contained motifs, how many sequences total were input, the number of times a motif was found per sequence, and the motif itself. Only one motif was reported for each program; if more were found, the top result of length 10 was used. For each motif found in each output, the following were exported: the transcription factor binding site (TFBS) program type, the percent of genes per cluster with the given shared motif versus total genes (% Shared Motif), the identifier for which the sequence was found, the frequency of the motif, and the sequence of the motif itself. To clarify, the % shared motif is measured on a *per cluster* basis, so for example: If a cluster has 10 nodes and 10 of them are found to have a motif in common, the % Shared Motif for that cluster will be 100%, and every gene in that cluster will have an annotation of 100% for Shared Motif. By contract, a cluster with 10 genes and only 4 of them having a shared motif will have a % Shared Motif of 40%.

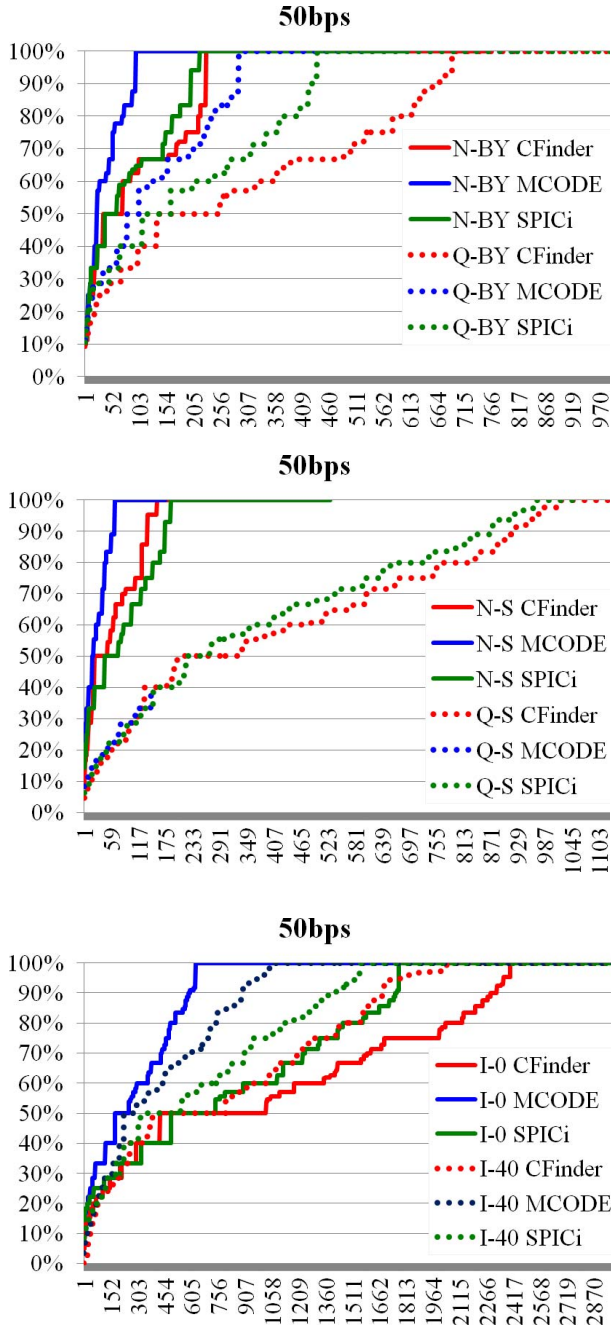

Fig. 4. The networks at 50bps upstream. Top: GSE8542. Middle: GSE8559. Bottom: GSE46384. The horizontal axis represents the rank of the Shared Motif percentage (%Shared Motif) and the vertical axis represents the percentage of genes among a cluster with a common predicted transcription factor binding site.

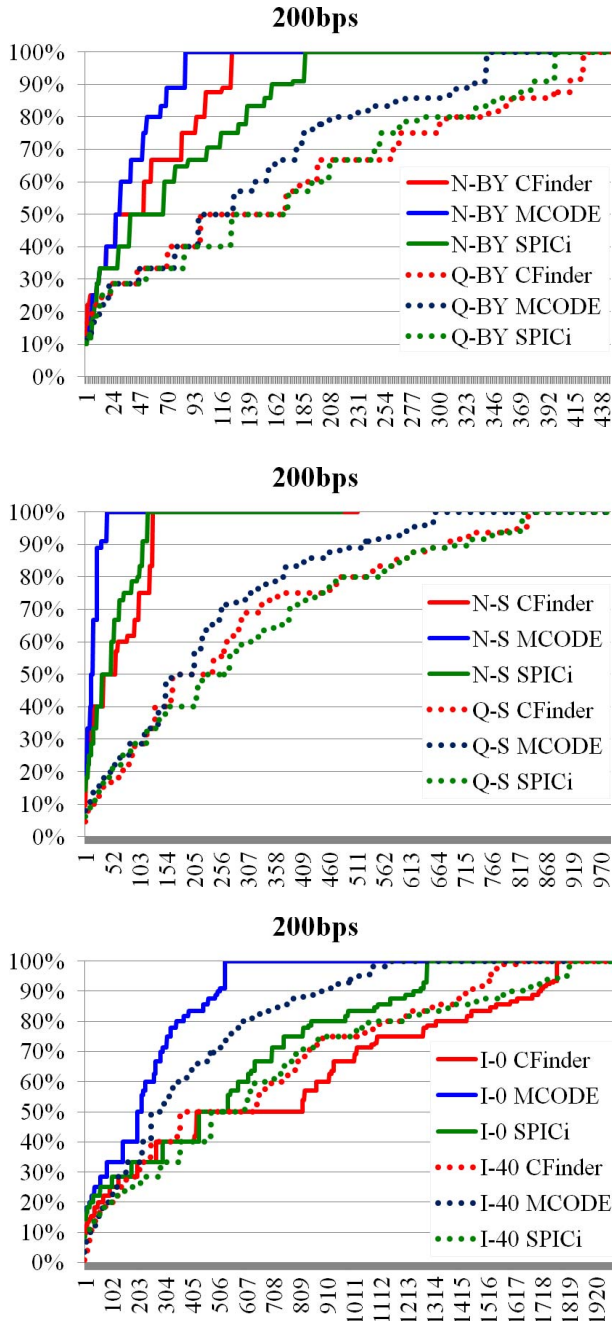

Fig. 5. The networks at 200bps upstream. Top: GSE8542. Middle: GSE8559. Bottom: GSE46384. The horizontal axis represents the rank of the Shared Motif percentage (%Shared Motif) and the vertical axis represents the percentage of genes among a cluster with a common predicted transcription factor binding site. Up to 1000 datapoints making up the tails of these figures (100% shared motif scores) have been removed for easier viewing.

III. RESULTS

Revisiting our original hypothesis, the goal of this work is to investigate the common patterns between the regulatory regions of genes found in common clusters in correlation networks. In an ideal setting, the network would be free of noise from a biological viewpoint and from noise artifacts due to model building. However, previous studies[7,24-26] and the nature of correlation both confirm that noise will be present in the network, regardless of thresholding and statistical analysis. While there are methods to counteract and remove this noise[7,24,27-29], it is thought that clusters are typically the most biologically reliable structures to examine due to the fact that "where there is smoke, there is fire." A cluster of size 10 that has 80% density is unlikely to be composed of entirely noisy relationships. As such, some clusters that did not have common transcription factors were expected to be found. Despite this, over half the clusters found and examined by any TFBS method had 100% of the genes in the given cluster share a motif.

A. The upstream region window size of 200bps is more robust than the window size of 50bps

Figures 4 and 5 examine the three networks (I-0 v. I-40, Q-BY v. N-BY, and Q-S v. N-S), identify the % Shared Motif for each cluster found by each motif finding program and rank them in increasing order. These plots include all 4 motif finding programs but the figures themselves do not discriminate between them. Examining the plots for each, the following can be seen:

- The plots comparing CFinder and MCODE for each network for 50bps and 200bps tend to mirror each other in %Shared clusters found and plot pattern, suggesting that the method is robust to clustering method.
- As the upstream region gets larger, the difference in plot patterns in each of the figures (comparing I-0 and I-40 at 50bps and I-0 and I-40 at 200bps) becomes smaller in all cases. This suggests that as the size of the upstream region becomes larger, the more robust to clustering the method becomes.

B. All networks show the majority of Shared Motif scores in the 81-100% range

The hypothesis, if proven true, indicates that the *majority of genes in a cluster in a correlation network are co-regulated by a common transcription factor*. Comparing the frequency of Shared Motif percentage counts supports this hypothesis. Figure 6 takes all % Shared Motif scores for each network+clustering+motif finding tool and determines whether those scores fall within five ranges: 0-20%, 21-40%, 41-60%, 61-80%, or 81-100%. In each and every case, the 81-100% range has a much higher frequency than any other range, and significantly so. This means that in the vast majority of clusters, at least 80% of the genes within the cluster contain a common binding motif, and this is consistent across clustering and motif finding approaches, indicating this is robust and not an artifact of any method used.

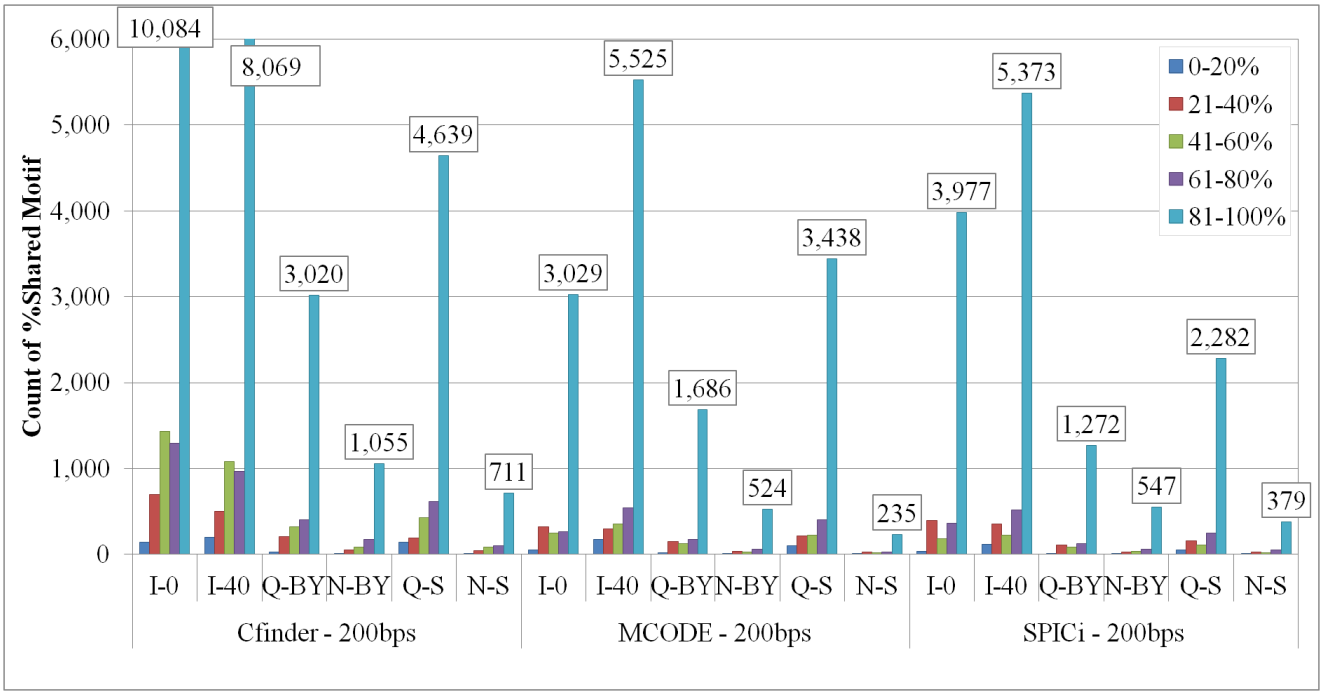

Fig. 6. Frequency of shared motifs. X-axis represents the clustering method and the network at 200bps upstream; the y-axis represents the count of genes with a given % Shared Motif. There are 5 ranges that a shared motif percentage can fall into (0-20%, 21-40%, 41-60%, 61-80%, 81-100%).

Figure 7 shows three examples of motifs derived from the clustering process; the first comes from cluster 1 of the I-40 network with SPICi clustering and ELPH TFBS motif finding; the second, from the first cluster of the I-40 network with MCODE clustering and ELPH TFBS motif finding, and the third from I-40 cluster 1 using CFinder and GLAM motif finding. Each is a sequence logo created using Berkeley's WebLogo program (http://weblogo.berkeley.edu/)[30]. Note: While these motifs all come from the first cluster per method, the first cluster from CFinder will not be the same as the first cluster from MCODE. These motifs, therefore, are not related.

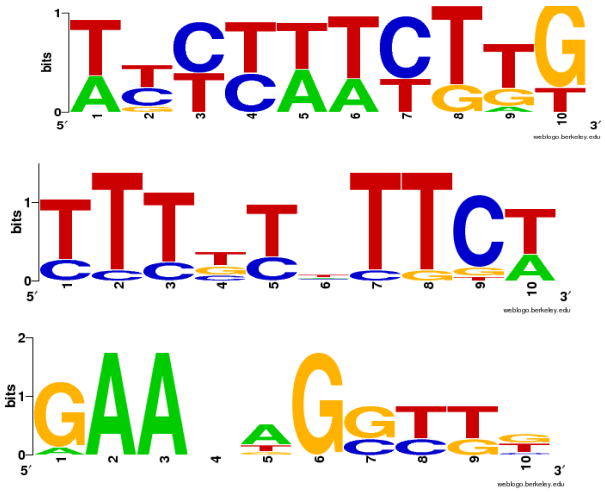

Fig. 7: Top: the motif from cluster 1 of the I-40 network with SPICi clustering and ELPH TFBS motif finding; Middle: motif from the first cluster of the I-40 network with MCODE clustering and ELPH TFBS motif finding; Bottom: motif from I-40 cluster 1 using CFinder and GLAM motif findinge networks at 200bps upstream. All were taken from 200bps upstream of the coding start and all motifs were found in 100% of genes (100% Shared Motif). The vertical axis represents how much of the content is represented as that base.

C. Novel transcription factor binding site discovery

The second motif was chosen randomly to investigate its biological reference; to validate the co-regulation of the genes in the cluster that produced this motif, the 19 genes found in this cluster were extracted and ran through YEASTRACT's database of documented and potential transcription factors and binding sites. YEASTRACT provides two types of transcription factor-to-site annotation, *Documented* and *Potential*. Documented indicates that the regulatory relationship specified has been found in literature in *Saccharomyces cerevisiae* and is available online at their website. Potential indicates that the regulatory region of the gene in question has a binding site sequence match with the known binding site(s) of the transcription factor. Finally, if a gene has a "predicted" site, within the context of this paper, this indicates that we are predicting the TF-binding site relationship based on "guilt-by-association;" that the genes in the cluster share a motif binding site and a known transcription factor, and as such, we expect the others with a found motif will also have potential binding sites for this TF. Of these nineteen genes found in the cluster, eleven (*Ssa2, Ufd1, Pmc1, Vtc2, YOR389W, BNI5, HSP104,YTP1, RPL5, MMR1,* and *RPC82*) were found to be commonly regulated by transcription factor *Hsf1p* by documented or potential evidence. *Hsf1p* (YHR073W) is involved in regulating response to stress, particularly stress related to temperature changes. The other eight genes (*RPB5, CUE2, DBR1, DBF4, MPS3, GDS1, SRP72,* and *RPL43A*) found then become genes

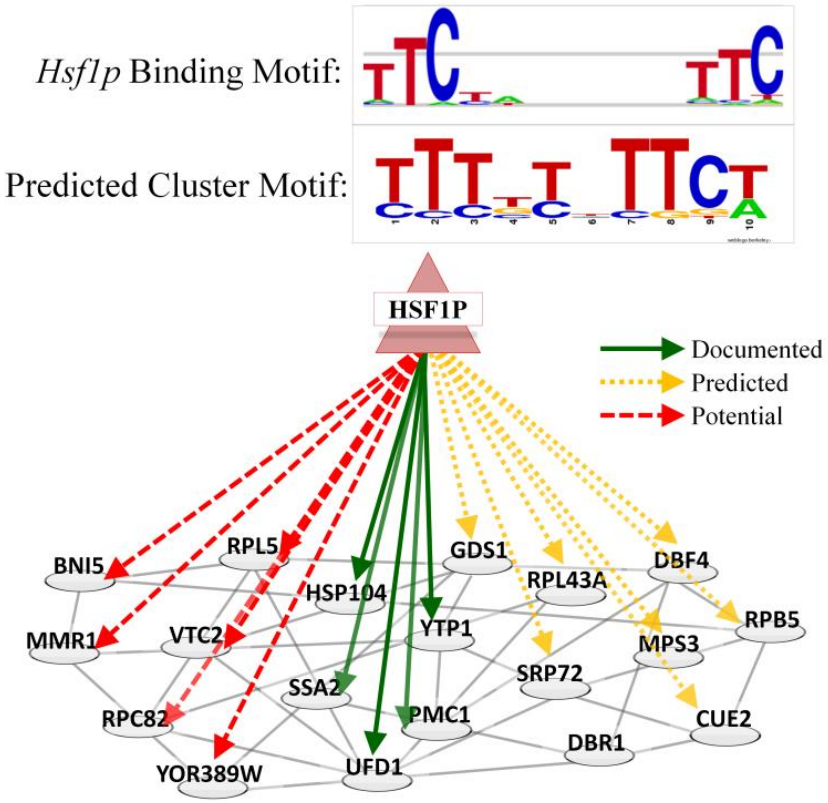

Fig. 8. An example of motif prediction using clustering and shared motifs for transcription factor *Hsf1p*.

that we are able to predict may be regulated by *Hsf1p*; this is further confirmed by examining the known binding site(s) of *Hsf1p*, which include TTCYNNNNNNTTC and cTTCtaGAAgcTTCtaGAAg, the first of which best fits our found motif. An example of this is shown in Figure 8.

D. Functional Analysis

To determine whether or not these genes had a common function, an assessment of Gene Ontology biological process annotation was performed on the genes from the I-40 cluster. While the set was not large enough to produce significant enrichment results, it was found that there were certain functions shared by the genes in the cluster: Documented, Predicted, and Potential. These results show that these genes largely play roles in various types of binding and localization, among others not noted here, as shown in Table 3.

IV. Discussion

This research aims to confirm the long-speculated thought that clusters in gene correlation networks (built from expression data) represent co-regulated, and as such, co-expressed, genes. The research here employs three different input sets, three different clustering methods, two different window sizes, and four different motif finding methods to identify if this hypothesis is (1) true and (2) robust to changes in approach. This work is a first step in investigating the usefulness of a cluster as a network feature that can be used for prediction of novel transcription factor binding sites, and eventually, co-regulation.

The results of this work found that a larger window size is preferential for method robustness (50bps is less stable than 200bps) although more rigorous testing of this in a larger benchmarking study would be recommended. However, the small changes between the two window sizes might not be significant enough to warrant these steps. Further, we found that the majority of genes within the network clusters largely shared a motif; i.e., each gene that was part of a cluster was very likely to be a member of a cluster where the motif found was shared by 81-100% of that group.

Finally, we shared an example of how this approach can lead to novel transcription factor binding site discovery, and/or discovery of novel co-regulated genes. This approach can harness the wealth of data available publicly to create models, reducing time and money spent and aiding in laboratory decision support by providing better targets for transcription factor binding in different cellular environments, and can also potentially lead to the discovery of new transcriptional binding sequences.

TABLE III. GENE ONTOLOGY ANNOTATIONS FOR CLUSTER GENES

ID	Name	Type	Localization	Binding
YNL166C	BNI5	Potential	X	
YKL090W	CUE2	Predicted	X	
YDR052C	DBF4	Predicted	X	X
YLL026W	HSP104	Documented	X	
YLR190W	MMR1	Potential		
YJL019W	MPS3	Predicted	X	
YGL006W	PMC1	Documented		
YBR154C	RBP5	Predicted		
YPR190C	RPC82	Potential		
YPR043W	RPL43A	Predicted	X	
YPL131W	RPL5	Potential		X
YPL210C	SRP72	Predicted	X	
YLL024C	SSA2	Documented		X
YGR048W	UFD1	Documented		X
YFL004W	VTC2	Potential		
YOR389W	YOR389W	Potential	X	
YNL237W	YTP1	Documented		
YKL149C		Predicted		X
YOR355W		Predicted		X

ACKNOWLEDGMENT

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 it's contents are the sole responsibility of the authors and do not necessarily represent the official views of NCRR or NIH.

REFERENCES

1. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science*. 1999;286(5439):509-512.

2. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. *Nature*. 2001;411(6833):41-42.

3. Barabasi AL, Oltvai ZN. Network biology: Understanding the cell's functional organization. *Nat Rev Genet*. 2004;5(2):101-113.

4. Dempsey K, Ali H. Evaluation of essential genes in correlation networks using measures of centrality. . 2011:509-515.

5. Dempsey K, Ali H. On the discovery of cellular subsystems in correlation networks using centrality measures . *Current Bioinformatics*. 2012 (Pending).

6. Dempsey K, Thapa I, Bastola D, Ali H. Identifying modular function via edge annotation in gene correlation networks using gene ontology search. *2012 IEEE International Conference on Bioinformatics and Biomedicine Workshops*. 2011;0:255-261.

7. Dempsey K, Duraisamy K, Bhowmick S, Ali H. The development of parallel adaptive sampling algorithms for analyzing biological networks. . 2012:725-734.

8. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005;4:Article17.

9. Dong J, Horvath S. Understanding network concepts in modules. *BMC Syst Biol*. 2007;1:24.

10. Alexeyenko A, Lee W, Pernemalm M, et al. Network enrichment analysis: Extension of gene-set enrichment analysis to gene networks. *BMC Bioinformatics*. 2012;13:226-2105-13-226.

11. Song L, Langfelder P, Horvath S. Comparison of co-expression measures: Mutual information, correlation, and model based indices. *BMC Bioinformatics*. 2012;13(1):328.

12. Barrett T, Wilhite SE, Ledoux P, et al. NCBI GEO: Archive for functional genomics data sets--update. *Nucleic Acids Res*. 2013;41(Database issue):D991-5.

13. Bader GD, Hogue CW. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*. 2003;4:2.

14. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*. 2002;30(7):1575-1584.

15. Adamcsek B, Palla G, Farkas IJ, Derenyi I, Vicsek T. CFinder: Locating cliques and overlapping modules in biological networks. *Bioinformatics*. 2006;22(8):1021-1023.

16. Jiang P, Singh M. SPICi: A fast clustering algorithm for large biological networks. *Bioinformatics*. 2010;26(8):1105-1111.

17. Li X, Wu M, Kwoh CK, Ng SK. Computational approaches for detecting protein complexes from protein interaction networks: A survey. *BMC Genomics*. 2010;11 Suppl 1:S3-2164-11-S1-S3.

18. Quest D, Dempsey K, Shafiullah M, Bastola D, Ali H. MTAP: The motif tool assessment platform. *BMC Bioinformatics*. 2008;9 Suppl 9:S6.

19. Kasprzyk A. BioMart: Driving a paradigm change in biological data management. *Database (Oxford)*. 2011;2011:bar049.

20. The ELPH home page. http://cbcb.umd.edu/software/ELPH/. Updated 2006. Accessed 08/09, 2013.

21. Frith MC, Hansen U, Spouge JL, Weng Z. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res*. 2004;32(1):189-200.

22. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res*. 2009;37(Web Server issue):W202-8.

23. Pavesi G, Mereghetti P, Zambelli F, Stefani M, Mauri G, Pesole G. MoD tools: Regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res*. 2006;34(Web Server issue):W566-70.

24. Duraisamy K, Dempsey K, Ali H, Bhowmick S. . 2011:721-728.

25. Halappanavar M, Feo J, Dempsey K, Ali H, Bhowmick S. . 2012:58-67.

26. Dempsey K, Bhowmick S, Ali H. Function-preserving filters for sampling in biological networks. *Procedia Computer Science*. 2012;9(0):587.

27. Dempsey K, Duraisamy K, Ali H, Bhowmick S. A parallel graph sampling algorithm for analyzing gene correlation networks. . 2011.

28. Dempsey K, Bonasera S, Bastola D, Ali HH. A novel correlation networks approach for the identification of gene targets. . 2011:1-8.

29. Dempsey K, Thapa I, Bastola D, Ali H. Functional identification in correlation networks using gene ontology edge annotation. *Int J Comput Biol Drug Des*. 2012;5(3-4):222-244.

30. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. *Genome Res*. 2004;14(6):1188-1198.