Community Detection in Transcription Factor Networks For S. Cerevisiae

Lisa Liao, Liza Sivriver

Washington University in St. Louis, St. Louis, 63130

December 11, 2019

Abstract

A central goal of postgenomic biology is to explore the regulatory relationships that governs cell activities, and Transcription Factor (TF) networks have long been a focus for computational biologists [1]. Statistical models are typically used for network mapping, analysis of TF-Gene interaction, and for the inference of functional similarities. As opposed to attempting network mapping or network construction from the statistical side, we make use of graph theory. We use naive graph community detection algorithms on a mapped network of the *S.cerevisae* to uncover functional similarities of TF and genes that participate in the same biological pathway, using modularity to analyze the quality of these communities. Furthermore, we apply shared Gene Ontology (GO) term for functional analysis to determine biological significance.

***Keywords*— transcription factor network, analysis of graph topology, bioinformatics**

1 Introduction

In biology, a central goal is turning sequences of DNA into proteins, which are responsible for various tasks within a cell. In order to extract information and create different proteins, there are 2 crucial steps: transcription and translation. Transcription produces an intermediate product called mRNA, which is a different format of the original DNA sequence. mRNAs undergo the process of translation, turning them into proteins. This two step process illustrates the central dogma of biology [2].

Transcription factors (TF) are proteins that regulate the expression of a gene, essentially controlling the rate a specific gene is transcribed. One or more TFs can work together to regulate expression of the same gene. Thus, genes can have one or more regulators. When a TF increases the level of expression of a gene, it is an activator. On the other hand, TFs that decrease expression of a gene are repressors. Each TF can directly regulate different target genes, or they can regulate other TFs which regulates other target genes [2]. These TF-Gene and TF-TF interactions make up the TF network, which is the network of interest in this project. The constructed network and its composition are discussed later in the report.

The understanding of the complex relationships and interactions behind these components is crucial in deciphering and engineering the mechanisms that govern the myriad of biological processes that are vital to living organisms. A deeper dive into the substructure of these networks allows prediction of potential effects when TF activity is perturbed, where perturbation is the act of disrupting the system by removing or adding a TF.

2 Motivation

We are interested in how graph analysis techniques perform on TF networks, which is a less common practice in TF network analysis. We chose to study Saccharomyces cerevisiae (**S. cerevisiae**), a species of yeast, because of the focus on the species as a model organism. Not only is there an industrial incentive to study yeast, but *S. cerevisiae* research is an economic driver with applications to medicine, food science, and more [1]. We aim to further the investigation of this model organism, and apply graph theory to fundamental biological study.

Applying community detection algorithms to find clusters in *S. cerevisiae* has an incredible impact on the intersection of random graph theory and the study of TF networks. One of the goals in the field of TF network inference is using statistical model to predict how activities of TFs fluctuate with perturbation. This is a primary concern because physically measuring TF activities can become expensive when the network size increases. With respect to statistical models, an accurate starting point is crucial in finding optimal model parameters that yield statistically and biologically significant predictions. However, the accumulation of prior knowledge of the network is needed to build a model that possesses high prediction accuracy [3]. By using naive community detection

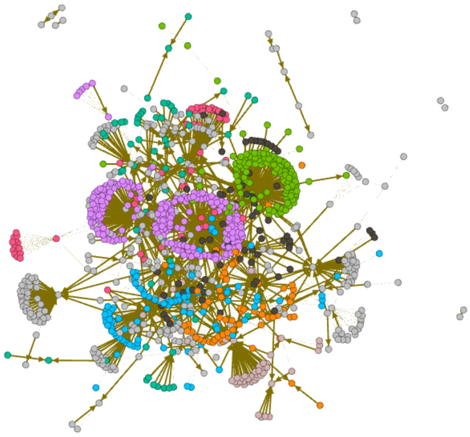

Figure 1: Directed network in Gephi with 7 connected components

algorithms, we hope to find TF-gene clusters that possess graphical and intracluster functional similarities. If we can find such clusters without prior biological information, this can be a novel tool to initiate statistical models and improve performance in predicting TF activity. The outcome can be applied to real-world problems such as finding effective drugs, developing novel medicines, and conducting gene editing processes like CRISPR.

3 Data & Graph Construction

Our data source is a regulatory network of *S. cerevisiae* compiled by Tchourine et al., where the network is rendered as a binary adjacency matrix representing a graph of TF-gene interactions [4]. The rows of this matrix represent genes, while the columns represent TFs. From this data, we construct a weighted, directed graph representing a regulation map that indicates which TF regulates which gene. This source represents 993 genes, 98 transcription factors, and 1,403 interactions. The nodes of graph illustrate TFs and genes, while the edges of graph represent interactions between TFs and genes, which is illustrated in Figure 1, visualized by Gephi [5].

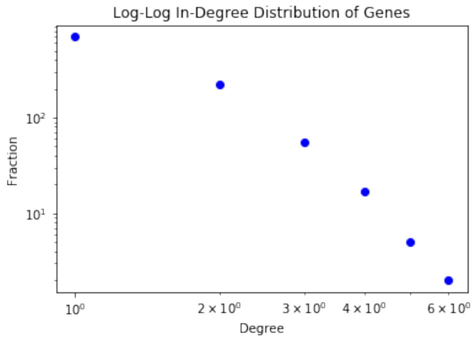

Figure 2: Degree distribution of in degrees of genes, which is exponentially distributed

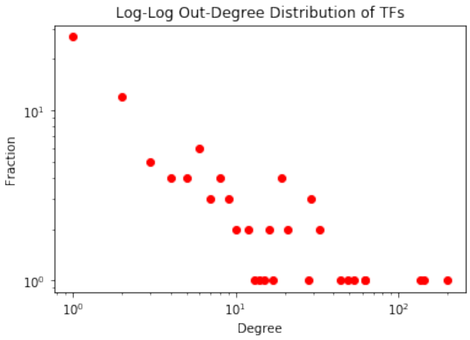

Figure 3: Degree distribution of out degrees of TFs, which exhibits a power law distribution

Degree distributions of TFs and genes are depicted in Figures 2 and 3. Because the network is directed mostly from TF to genes, the out degree of a TF node shows how many target genes it has, while the in degree of a gene node represents the number of regulators [6]. We identify hubs and see the degree distribution of the out degrees is scale-free [7]. The in degrees of gene nodes are exponentially distributed while the out degrees for TF nodes behave like a power-law distribution. This means that yeast TF network belongs to a mixed class of networks, between exponential and power-law [8].

After analyzing this network, we created an undirected graph by retaining all edges from the largest connected component ($n = 1055$) of the directed graph. By making the network undirected, we turn the regulation map into a interaction map. We can apply community detection algorithms on this undirected graph to further identify clusters. When there are disconnected components, the community detection algorithm techniques are not fully accurate, so we removed all nodes not included in the largest connected component.

4 Algorithms & Methods

4.1 Louvain Algorithm

The Louvain method for community detection is greedy optimization method that attempts to optimize the "modularity" of a partition of the network, based on the commonly used modularity maximization algorithm. First, the method searches for small communities by optimizing local modularity. Afterwards, the Louvain method aggregates nodes belonging to the same community, constructing a new network whose nodes are the found communities. The method repeats this process iteratively until it attains maximum of modularity and produces a hierarchy of communities. With wideranging applications, this detection algorithm has been used to identify communities in Twitter, mobile phones, and other very large networks, making it relevant for our particular network.

4.2 Spectral Clustering

Spectral Clustering is a community detection algorithm that takes advantage of the eigenvectors of the graph Laplacian L, where L is given as

$$L = D - A \tag{1}$$

where D is the degree matrix

$$D_{ii} = deg(v_i) = \sum_{j=1} A_{ij} \tag{2}$$

and A refers to the adjacency matrix. We performed spectral clustering on the adjacency matrix of the undirected graph, which determined clusters based on the eigenvectors of L [9].

4.3 Modularity

Modularity, a mesoscopic network measure, is used to determine the strength of found clusters in a given graph. Biological networks usually exhibit high measures of modularity. Modularity Q is given by the following equation:

$$Q = \frac{1}{2m} \sum_{ij} [A_{ij} - \frac{k_i k_j}{2m}] \delta(c_i, c_j) \tag{3}$$

4.4 Shared Gene Ontology (GO) Terms

We used the Gene Ontology Resource database [10], which is the largest source of information on the function of genes. This database houses the most comprehensive documentation of biological annotations [11]. In Tables 2 and 3, the column "Number Mapped" suggests the number of genes or TFs the database found a biological annotation for. The "% Shared" columns represent the proportion of genes sharing the same biological process or molecular function with other nodes in the cluster. The data in these columns is calculated by dividing the number of genes sharing the same biological process or molecular function by the number mapped. The highlighted rows are clusters with the same annotation, both biological processes and molecular functions, across results of community detection algorithms.

5 Results

The Louvain algorithm found 20 distinct clusters with an average modularity score of 0.7356. In terms of shared GO terms, the average percentage of TFs and genes are components of the same biological process within a cluster is 44.90%. The highest percentage for an individual cluster reaches up to 76.47% in Cluster 16. On the other hand, the average percentage of TFs and genes sharing the same molecular function is 36.83%, with the highest individual percentage being 77.42% for Cluster 8.

The Spectral Clustering algorithm found 20 clusters with an average modularity score of 0.7007. With regards to shared GO terms, the average percentage of TFs and genes that are components of the same biological process within a cluster is 50.26%, reaching up to 100% in Cluster 2. When analyzing the proportion of shared TFs with respect to molecular functions, the average percentage is 39.52%.

Algorithm	Modularity Score	% Shared Biological Process	% Shared Molecular Function
Louvain	0.7356	44.90%	36.83%
Spectral Clustering	0.7007	50.26%	39.52%

Table 1: Modularity and Shared GO Terms for each community detection algorithm

5.1 Analysis

With regards to modularity scores, Louvain algorithm outperforms Spectral Clustering algorithm. This suggests clusters are better defined when detecting communities with the Louvain algorithm. Even though there is a potentially negligible difference between these two numbers, we conclude that the Louvain algorithm performs better on average than Spectral Clustering.

However, because we are investigating a biological dataset, the biological significance of the clustering needs to be more heavily considered than the results based on the modularity scores. Quantitatively, spectral clustering outperforms Louvain algorithm as the intracluster percent of shared GO terms for both biological process and molecular function is higher. Quantitatively, both algorithms capture the umbrella functional modules such as various biosynthesis processes that produce biomolecules and metabolic processes harvesting biomolecules for energy. In Table 2, 12 clusters are annotated with similar biological processes and molecular functions between the 2 algorithms.

We find a significant pattern in the results, where Spectral Clustering finds more specialized subclusters than Louvain algorithm. When examining the node composition of the remaining 8 clusters that do not have matching processes or molecular functions, Cluster 8 in Spectral Clustering contains nodes that are subsets of Cluster 9 in Louvain algorithm. These two clusters share the same biological processes but possess different molecular functions. However, the molecular function for Cluster 9 in Louvain algorithm (catalytic activity) encompasses the molecular function for cluster 8 in Spectral Clustering (DNA-binding transcription factor activity, RNA polymerase II specific). This phenomenon is observed in other clusters, where Cluster 18 in Spectral Clustering is a subset of Cluster 7 in Louvain algorithm. Similarly, Cluster 13 in Spectral Clustering is a subset of Cluster 5 in Louvain algorithm. We see this behavior in two other clusters, where Cluster 2 in Spectral Clustering is a subset of Cluster 2 in Louvain algorithm, and Cluster 13 in Spectral Clustering is a subset of Cluster 13 in Louvain algorithm. All of the clusters in Spectral Clustering have molecular functions encompassed by their corresponding superset clusters in Louvain algorithm. This significant trend indicates functional submodules within our current clusters that can be divided into subclusters with more specific processes. Spectral clustering is able to capture these submodules better than Louvain algorithm because 5 out of 8 non-highlighted clusters displays the phenomenon described above.

When analyzing the percentage of shared GO terms within clusters, Spectral Clustering finds a higher average percentage in both biological processes and molecular functions. This phenomenon is attributed to the generally smaller and more function-specific clusters found by Spectral Clustering. Although Cluster 2 only had four nodes, the shared percentage of biological process reached 100%. The same 4-node cluster also reaches 100% similarity in molecular function. Because of the relatively small size of the cluster, there must be high similarity between nodes for the algorithm to collect these nodes into a cluster, so it follows that 100% of this cluster would share both biological process and molecular function. Furthermore, these nodes perform Pentose Transmembrane Transport, which is a highly specialized pathway.

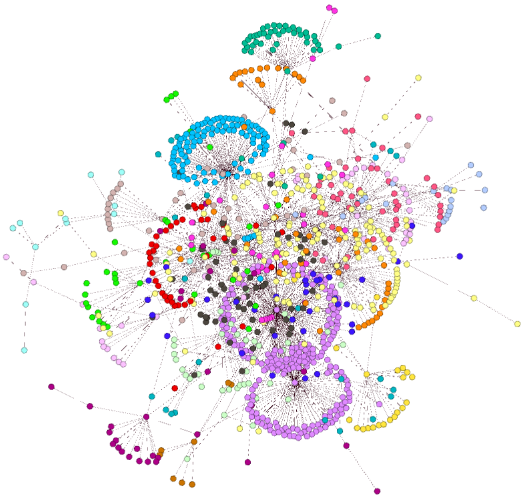

Figure 4: Largest connected component of the TF network

5.2 Network Exploration

It is crucial to analyze the largest connected component of our network since the community detection algorithms do not perform optimally with not fully connected graphs. We applied the Louvain and Spectral Clustering algorithms on this largest connected component. With 1,055 nodes, this graph had 1,391 interactions, both between TFs and genes as well as amongst TFs. The largest community, depicted in bright magenta, helps frame our understanding of the theoretical underpinnings supporting our network.

By studying the largest communities in the network, we can apply both biological and graphical interpretations to the most significant communities. In the Louvain algorithm results, Cluster 7 was the largest community with 230 nodes, as visualized in Figure 4. Figure 5 shows the largest community found with Spectral Clustering, which was also Cluster 7, with 260 nodes. The biological function for a majority of the nodes is cellular protein complex disassembly, while the dominant molecular function is structural molecule activity. This makes sense that the function is related to structural activity since many genes are needed to digest and break down biomolecules. Because disassembling a protein complex means breaking down its structure, it is natural that these functions and processes would be associated with the largest communities.

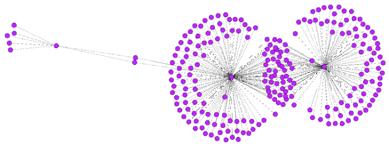

Figure 5: Cluster 7, the largest community found with the Louvain algorithm

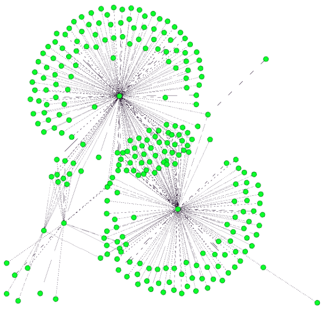

Figure 6: Cluster 7, the largest community found with Special Clustering

6 Discussion

6.1 Limitations

One limitation in our project was that we did not take advantage of graph metrics specifically considering the biological significance of nodes and clusters. Jalili et. al [12] described 4 different centrality measures tailored to biological networks that include integration of omics data and a combination of classical centrality measures.

Since we were only able to apply our community detection algorithms to undirected graphs, we see potential to utilize other algorithms that take in directed graph data. Applying directed graph algorithms could have made our results more nuanced.

6.2 Conclusion

Overall, the main goal of this project is to create a starting point to study TF activity. The clusters that we have found can be used to build statistical models that analyze perturbation. We were able to find functional clusters with high intracluster percent shared GO annotations. At the same time, this is done without prior biological knowledge using naive community detection algorithms, which is novel in the intersection of graph theory and computational biology. Our research will be a useful tool in guiding future research applying graphical analysis to TF networks. Code for this project can be found at `https://github.com/lislia423/transcriptionFactor_communityDetection`

6.3 Future Work

Based on the results of this project, there is great potential for future research and exploration. One direction would be to find and analyze biological significance of overrepresented motifs in the network. Motifs are crucial to interpret because motifs enhance the versatility of information processing in a TR network [7]. Furthermore, an abundance of motifs suggests the overall functional robustness that motifs provide during evolutionary adaptation to changing environmental conditions [7]. Researchers have already detected Feed Forward Loops (FFL), Single regulatory interaction with mutual regulation(SMR), as well as convergence with mutual regulation (CMR) motifs in *S. Cerevisiae*, yet this is only a starting point in comprehending the variety of motifs in the organism. Currently, larger motifs have not been detected. Many motif detection algorithms are unfeasible when the possible number of sub-graphs is large [13], so it is crucial to develop algorithms to capture larger motifs in this species of yeast.

An additional path for exploration is analyzing other types of protein interaction, since TF interaction is only one subcategory of protein interaction. Protein-protein and protein-metabolyte are interesting avenues for network analysis. This analysis would contribute to a deeper understanding of *S. Cerevisiae*'s structure. Because

S. Cerevisiae is a model organism, we can apply information about this species to develop our understanding of more complex species.

References

[1] I. Farkas, H. Jeong, T. Vicsek, A.-L. Barabási, and Z. N. Oltvai, "The topology of the transcription regulatory network in the yeast, saccharomyces cerevisiae," *Physica A: Statistical Mechanics and its Applications*, vol. 318, no. 3-4, pp. 601–612, 2003.

[2] K. R. Miller, *Miller & Levine Biology*. Pearson, 2010.

[3] R. Bonneau, D. J. Reiss, P. Shannon, M. Facciotti, L. Hood, N. S. Baliga, and V. Thorsson, "The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome biology*, vol. 7, no. 5, p. R36, 2006.

[4] K. Tchourine, C. Vogel, and R. Bonneau, "Condition-specific modeling of biophysical parameters advances inference of regulatory networks," *Cell reports*, vol. 23, no. 2, pp. 376–388, 2018.

[5] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," in *Third international AAAI conference on weblogs and social media*, 2009.

[6] W. Z. Ouma, K. Pogacar, and E. Grotewold, "Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties," *PLoS computational biology*, vol. 14, no. 4, p. e1006098, 2018.

[7] I. J. Farkas, C. Wu, C. Chennubhotla, I. Bahar, and Z. N. Oltvai, "Topological basis of signal integration in the transcriptional-regulatory network of the yeast, saccharomyces cerevisiae," *BMC bioinformatics*, vol. 7, no. 1, p. 478, 2006.

[8] C. Rodriguez-Caso, M. A. Medina, and R. V. Sole, "Topology, tinkering and evolution of the human transcription factor network," *The FEBS journal*, vol. 272, no. 24, pp. 6423–6434, 2005.

[9] C. Bauckhage, "Numpy/scipy/networkx recipes for data science: spectral clustering," 2016.

[10] M. L. Acencio, A. Lægreid, M. Kuiper, *et al.*, "The gene ontology resource: 20 years and still going strong.," 2019.

[11] H. Mi, X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas, "Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements," *Nucleic acids research*, vol. 45, no. D1, pp. D183–D189, 2016.

[12] M. Jalili, A. Salehzadeh-Yazdi, S. Gupta, O. Wolkenhauer, M. Yaghmaie, O. Resendis-Antonio, and K. Alimoghaddam, "Evolution of centrality measurements for the detection of essential proteins in biological networks," *Frontiers in physiology*, vol. 7, p. 375, 2016.

[13] A. Masoudi-Nejad, F. Schreiber, and Z. R. M. Kashani, "Building blocks of biological networks: a review on major network motif discovery algorithms," *IET systems biology*, vol. 6, no. 5, pp. 164–174, 2012.

Appendix

Table 2: Louvain Algorithm Results

Cluster	Size	Number Mapped	Biological Process	% Shared	Molecular Function	% Shared
0	38	35	DNA duplex unwinding	11.43%	DNA helicase activity	11.43%
1	45	42	Cellular amino acid metabolic process	50.00%	Small molecule binding	50.00%
2	32	29	ATP metabolic process	44.83%	Structural constituent of ribosome	31.03%
3	98	88	Oxidation-reduction process	28.41%	Oxireductase activity	28.41%
4	14	13	Zinc ion transmembrane transport	23.08%	Zinc ion transmembrane transporter activity	23.08%
5	51	37	Small molecule metabolic process	48.65%	Oxidoreductase activity	24.32%
6	22	18	Ascospre formation	38.89%	Transmembrane transporter activity	22.22%
7	230	225	Cellular protein complex disassembly	53.78%	Structural molecule activity	56.44%
8	33	31	Organonitrogen compound metabolic process	64.52%	Catalytic activity	77.42%
9	126	116	Oxoacid metabolic process	68.10%	Catalytic activity	66.38%
10	65	59	Carbohydrate metabolic process	15.25%	DNA-binding transcription factor activity	15.25%
11	40	34	Oxidation-reduction process	52.94%	Oxireductase activity	47.06%
12	40	33	Protein folding	57.58%	Protein binding	63.64%
13	64	48	Mitotic cell cycle	33.33%	Sequence specific DNA binding	22.92%
14	45	33	Response to chemical stimulus	30.30%	DNA-binding transcription factor activity	12.12%
15	12	10	Lipid biosynthetic process	70.00%	Fatty-acyl-CoA synthase activity	20.00%
16	20	17	Transport	76.47%	Transmembrane transporter activity	58.82%
17	16	15	Cellular transition metal ion homeostasis	33.33%	Metal ion binding	33.33%
18	14	13	Xenobiotic detoxification by transmembrane export across the plasma membrane	21.43%	ATPase activity	21.43%
19	45	41	Protein metabolic process	75.61%	Hydrolase activity	51.22%

Table 3: Spectral Clustering Algorithm Results

Cluster	Size	Number Mapped	Biological Process	% Shared	Molecular Function	% Shared
0	10	10	DNA duplex unwinding	40.00%	DNA helicase activity	40.00%
1	73	67	Cellular amino acid metabolic process	31.34%	Small molecule binding	31.34%
2	4	3	Pentose transmembrane transport	100%	Pentose transmembrane transport activity	100%
3	26	24	Oxidation-reduction process	62.50%	Oxireductase activity	54.17%
4	23	21	Zinc ion transmembrane transport	14.29%	Zinc ion transmembrane transporter activity	14.29%
5	42	33	Small molecule metabolic process	54.55%	Oxidoreductase activity	27.27%
6	22	18	Ascospre formation	38.89%	Transmembrane transporter activity	22.22%
7	260	239	Cellular protein complex disassembly	48.12%	Structural molecule activity	56.07%
8	10	10	Oxoacid metabolic process	70.00%	Catalytic activity	30.00%
9	153	139	Oxoacid metabolic process	51.08%	Catalytic activity	71.22%
10	214	179	Response to chemical	24.58%	DNA-binding transcription factor activity	12.29%
11	16	12	Negative regulation of transcription by RNA polymerase II	33.33%	DNA-binding transcription factor activity, RNA polymerase II-specific	33.33%
12	44	39	Protein folding	43.59%	Protein binding	41.03%
13	20	12	Meiotic cell cycle	33.33%	Heterocyclic compound binding	25.00%
14	38	31	Response to chemical	35.48%	Oxidoreductase activity	29.03%
15	12	10	Lipid biosynthetic process	70.00%	Fatty-acyl-CoA synthase activity	20.00%
16	21	18	Transport	77.78%	Transmembrane transporter activity	55.56%
17	13	10	DNA replication	40.00%	Sequence-specific DNA binding	40.00%
18	8	8	Carbohydrate derivative metabolic process	62.50%	Ribonucleoside-diphosphate reductase activity	37.50%
19	46	42	Protein metabolic process	73.81%	Hydrolase activity	50.00%