Regulatory Network Models in Biology

Manasi Vartak

Major Qualifying Project in the Mathematical Sciences

Worcester Polytechnic Institute

April 29, 2010

Advisors:

Prof. Brigitte Servatius (W.P.I.)

Prof. Marian Walhout (U.M.M.S.)

Acknowledgements:

I would like to express my deep gratitude towards Prof. Marian Walhout at UMMS for giving me this fantastic MQP opportunity. Her encouragement, enthusiasm, guidance and mentoring have been instrumental in making this project successful. I would also like to thank Prof. Brigitte Servatius at WPI for her guidance, supervision and mentoring. This project would not have been possible without her support. I further thank Prof. Zhiping Weng and Alos Diallo at UMMS for their help with the project. Finally, I'd like to thank the WPI Mathematics Department for their support in this collaborative MQP.

Regulatory Network Models in Biology

Manasi Vartak
Worcester Polytechnic Institute
mvartak@wpi.edu

April 29, 2010

Abstract

In this project, we studied transcriptional regulation in C. elegans through a network approach. We used techniques analyzing degree distribution, motifs, gene regulation subgraphs etc. to investigate various properties of the network. Our motif analysis discovered previously unknown motifs that are likely to have biological significance. We introduced a new technique for quantifying amount of gene regulation and formulated a new hypothesis to predict autoregulation. Our results will serve as a basis for future biological experiments.

1 Introduction

The life of an organism is best described by one word - "connectedness." Starting at the highest levels of complexity, this connectedness is seen within organisms in an ecosystem, within biological systems making up an organism, within cells making up biological systems and finally, within molecules making up cells. Life processes of organisms are not the product of a single biological unit working in isolation; instead, they result from multiple entities working as one.

Traditionally, biologists have studied organisms by separating biological systems into their building blocks and investigating these basic units. However, a reductionist approach provides limited insight into how these building blocks fit together and work as a unit [1]. Therefore, in this project, we seek to understand the working of cells - gene regulation in particular - through a *network approach*. By modeling the biological system as a graph, we seek to analyze the interconnections between biological units and understand their functioning at the system level [2]. Through the analysis of this network using techniques such as degree distribution, motifs and gene regulation subgraphs, we aim to extract biologically meaningful information. Results from such an analysis can provide hypotheses and theories that can then be verified through biological experiments. We now describe our study of the transcription regulatory network of the worm *Caenorhabditis elegans*.

2 Background

Cells in our body contain DNA, the code of life. This code of life stores all the information a cell needs to function, and is made up of thousands of protein-coding genes. Proteins in turn are responsible for utilizing DNA information to carry out cellular activities. They are involved in every cellular function one can imagine - digestion, carrying oxygen in blood, cell death, reproduction etc.

The production of proteins from genes is called *gene expression* and the *central dogma of biology* describes how information stored in DNA gets encoded as proteins [3]. The first step in producing proteins from genes is called *transcription*. Transcription involves the production of an mRNA molecule from DNA. This mRNA molecule is a mirror image of the DNA strand and encodes the information to make a protein. Once mRNA is made, it goes to the ribosome where the mRNA is read and the protein produced. The process of converting mRNA to protein is called *translation*. Thus, transcription and translation together are responsible for gene expression.

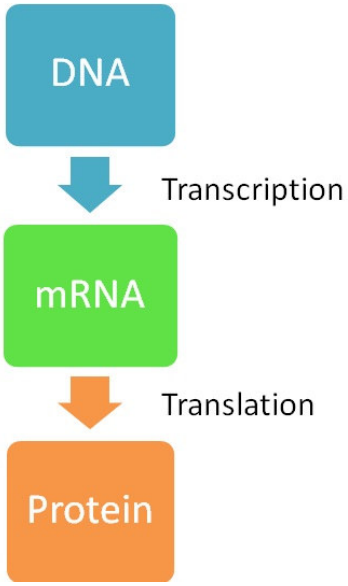


Figure 1: Producing proteins from genes

2.1 Differential Gene Expression

Although all cells in an organism contain the same DNA, and hence identical sets of genes, not all genes are expressed in all cells at all times. For instance, the protein haemoglobin is mainly produced in red blood cells, insulin in pancreas and serotonin receptors in the central nervous system. Furthermore, even if a protein is produced in certain types of cells, its amount does not remain constant over time. Although insulin is produced in the pancreas, its amount varies with glucose intake. Similarly, the amount of heat shock proteins increases as an organism is placed in a high temperature or stress environment. This difference in gene expression based on cell type, environment and external stimuli is termed *differential gene expression*.

Complex mechanisms in the cell control differential gene expression, and this controlling process is termed *gene regulation*. In the cell, gene regulation takes place at various levels such as: (1) Transcriptional regulation: Controls which genes produce mRNA, (2) Translational regulation: Controls which mRNAs are successfully converted into proteins, (3) Post-translational regulation: Controls which proteins are activated (e.g. through phosphorylation) or degraded etc. In this project, we focus on the transcriptional control of gene regulation.

2.2 Transcriptional Regulation

Since transcription is the first step in making a protein, transcription presents the first major opportunity to control gene expression. By controlling whether or not an mRNA molecule is made, the cell can avoid spending energy in the remaining protein production process. Transcriptional control is achieved through the action of special proteins called *transcription factors* (TFs). As shown in Figure 2, transcription factors bind to the *promoter* region upstream of the gene and influence the binding of RNA polymerase to DNA, thus controlling the production of mRNA. The binding of a transcription factor to a promoter can either aid or inhibit mRNA production.

Figure 2: Transcriptional regulation through transcription factors

3 Problem Description

Today, one of the important open problems in biology is understanding the exact mechanism of transcriptional regulation. We seek to answer questions such as: (1) Which TFs bind to which genes? (2) How are certain genes expressed in certain cells at certain times? (3) How does gene expression vary with external stimuli? etc. Various experimental and computational approaches have been proposed to tackle these challenging questions. In this work, we approach this problem from a network perspective.

3.1 Gene Regulation as Networks

Traditionally, networks (i.e. graphs) have been used to represented entities and their connections. For instance, a road network is used to represent cities and various routes connecting these cities. Similarly, a network can be used to model transcriptional regulation if we consider genes and TFs as entities and their interactions as connections [4]. Using this model, we can employ a rich set of graph theoretic methods to understand the working of the network. Moreover, while biological experiments tend to apply a divide and conquer approach, networks can put pieces together and enable us to analyze the whole system. Networks can thus provide system-level information that biological experiments cannot.

However, we must interpret results from network analysis with care because networks may not capture the complete information about a system. To illustrate, a network cannot model the expression of a gene over time; a technique like differential equations is better suited for such an analysis. Further, not all information gained from network analysis may be relevant or correct. Finally, whenever we analyze networks, we must be careful to not violate biological principles e.g. while randomizing the regulation network graph, we must not convert TF nodes into gene nodes.

4 Our Approach

To study the transcriptional regulation in *C. elegans*, we constructed a network model for the transcriptional system as follows.

4.1 Data

The data used in this project comes from biological experiments done at the Walhout Lab at University of Massachusetts Medical School. In these experiments, yeast-one hybrid assays were used to study interactions between TFs and genes in the worm *C. elegans*. These experiments adopt a gene-centric approach where each gene promoter is tested against a predetermined set of TFs. The presence or absence of regulation is inferred based on the binding interactions between TF-gene pairs. The data provides the set of likely regulatory interactions between TFs and genes along with metadata like strength of interaction, background noise etc.

The limitations of the current data set are: (1) Experimental techniques introduce errors in observed interactions. (2) Since data generation is ongoing, the data-set is incomplete and may suffer from a sampling bias. (3) The data does not encode information about whether the TF-gene interaction aids or inhibits gene expression.

4.2 Network Model

To convert the interaction data into a network, we use the following model. TFs and genes are considered nodes in the graph while regulatory interactions between TFs and nodes form the edges. The edges are directed from TFs to genes to signify that the TF binds to and regulates the gene. Thus, an incoming edge indicates that a node is regulating the current node, while an outgoing edges indicates that the current node is regulating another node. The nodes are divided into three sub-types to capture the differences between TFs and genes. The first class of nodes corresponds to genes and contains nodes with no outgoing edges. Thus, these nodes are regulated by other nodes, but they do not regulate any node. The second class of nodes corresponds to TFs and contains nodes with no incoming edge, i.e., these nodes regulate other nodes but are not regulated themselves. The third class of nodes contains hybrid nodes that have both incoming and outgoing edges, i.e., they regulate other nodes and are in turn regulated themselves. Figure 3 shows the network with each class of nodes condensed into a single node.

The network constructed via this model contains 1115 nodes and 8812 edges. Of the 1115 nodes, 115 nodes are hybrid, 790 nodes are genes and the remaining 210 nodes are TFs. Figure 4 shows a artificial network with properties similar to our network. We will illustrate various analysis techniques on this artificial network instead of the complete network.

5 Analysis

We analyzed the *C. elegans* transcription network at various levels of granularity beginning with the entire network and going to the individual node and edge.

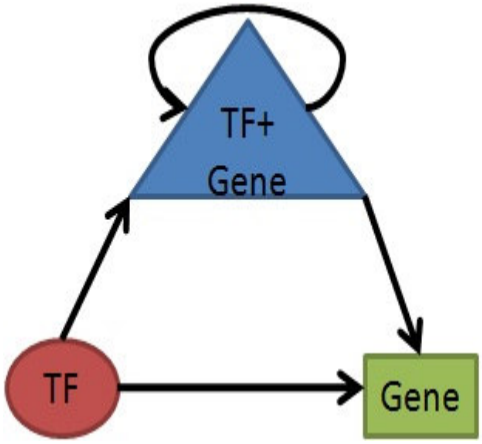


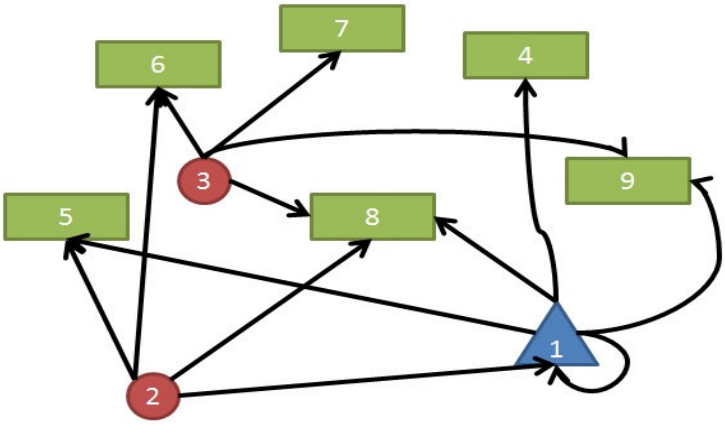
Figure 3: Prototype of Network

5.1 Degree Analysis

We first studied the overall architecture of the network through degree analysis. Figures 5 and 6 show the frequency distributions for in-degree and out-degree. From these figures, we see that the in-degree distribution follows an exponential distribution while the out-degree follows the power law or scale-free distribution. This result is in agreement with previously reported degree distributions for e-coli networks [5].

To explain the evolution of networks with these degree distributions, we studied the following models of network growth.

Preferential Attachment Model: The preferential attachment model proposed by Barabasi explains the evolution of scale-free networks through a "rich get richer" paradigm [6]. In this model, whenever a new node is added to the network, it connects to another node with probability proportional to the degree of the other node. Thus, a new node preferentially attaches to a node with high degree, and in-



Node	In	Out
1	2	5
2	0	4
3	0	4
4	1	0
5	2	0
6	2	0
7	1	0
8	3	0
9	2	0

Figure 4: Artificial Network

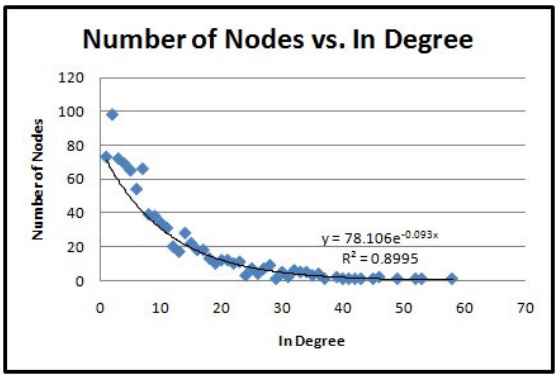


Figure 5: Distribution of In Degree

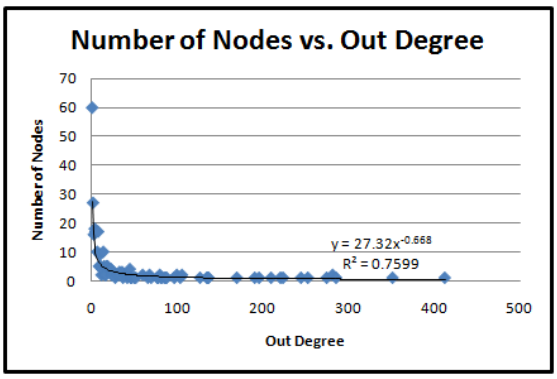


Figure 6: Distribution of Out Degree

creases the already high degree of the node. When applied to directed networks, however, this model gives rise to in and out-degree distributions that are both scale-free.

Callaway Model: The Callaway model explains the evolution of an exponential degree distribution [6]. In this model, when a node is added into the system, edges are added between a randomly picked pair of nodes. Thus, an older node is more likely to have a large degree. However, like the preferential attachment model, this model gives rise to an exponential distribution both in the in-degree and out-degree.

Vertex Copying Model: The vertex copying model explains the scale-free degree distribution in graphs by assuming that a graph grows by duplicating vertices [6]. Unlike the two models described above, the only way to add nodes in this model is to randomly pick a node and duplicate it while preserving all its connections. Edges can later be added or

removed at random to simulate the evolution process.

Our simulations show that the existing models do not fully explain the degree distribution observed in our network, and we predict that evolution of the gene network occurs through not one, but a combination of the above growth models.

5.2 Motifs

The second technique we used to analyze the network consisted of finding subgraphs that are over-represented in the original biological network as compared to randomized networks. These over-represented subgraphs are called *motifs* [7] and are important since they have been postulated to have biological significance.

To find motifs, we use the following procedure. For a given number of nodes n, all subgraphs with n nodes are enumerated and the original graph is analyzed to find the number of occurrences of each such subgraph. Following this, the original graph is randomized a large number of times, and each random graph is analyzed for the previously defined subgraphs. Once the number of occurrences of each subgraph has been measured in the original and randomized networks, we calculate the significance of each subgraph using Equation 1.

$$p - value(G_{sub}, N) = \frac{Random\ graphs\ with\ > N\ instances\ of\ G_{sub}}{Total\ number\ of\ Random\ graphs} \quad (1)$$

Thus, the p-value of a given subgraph is calculated as the number of random graphs that contain the given subgraph more often than the original network divided by the total number of random graphs. If this value is lesser than a threshold (0.05), then the subgraph is said to be a motif in the original network.

5.2.1 Randomization Techniques

Before we discuss the various motifs found in our network, we describe the types of graph randomization techniques used in this work [8]. The goal of our randomization techniques was to preserve certain graph characteristics while randomizing others. However, for all techniques, the classification of a node as a TF, gene or hybrid node is not changed through randomization.

Edge Replacement: This randomization is the strictest form of randomization that seeks to preserve the in and out-degree of each node while only randomizing its connections. For this purpose, the technique randomly picks two edges and swaps the endpoints of the respective edges. As shown in Figure 7, if the two edges chosen are A-B and C-D, edge replacement changes these edges to be A-D and B-C. This preserves the in and out degree of the four nodes while changing the connections.

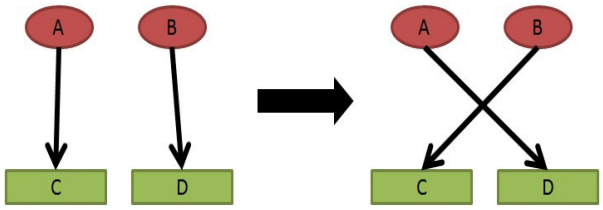


Figure 7: Edge Replacement

Node Replacement I: This randomization preserves the overall degree distribution of the network but changes the individual distribution of the nodes. For instance, if in the original network, node A had in-degree k and out-degree m, then in the randomized network, another node A' will have in-degree k and out-degree m. The connections for each node are also randomized. This method of randomization is shown in Figure 8. It is important to note that node replacement I gives the same motifs as edge

replacement, and hence it cannot be used to study the presence of motifs. This technique can however give insight into the biological properties of nodes in different motifs.

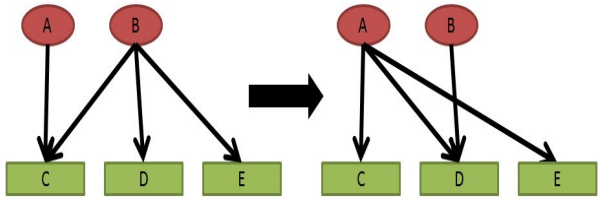


Figure 8: Node Replacement I

Node Replacement II: This randomization technique is the most lenient form of randomization. It only preserves the number of nodes and edges present in the graph and fully randomizes its connections. Figure 9 shows an example of this randomization technique.

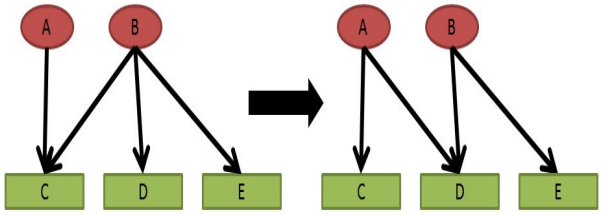
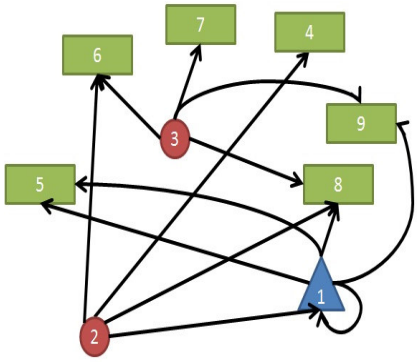


Figure 9: Node Replacement II

Figures 10, 11 and 12 show the random networks produced when the above methods are applied to the artificial network from Figure 4.

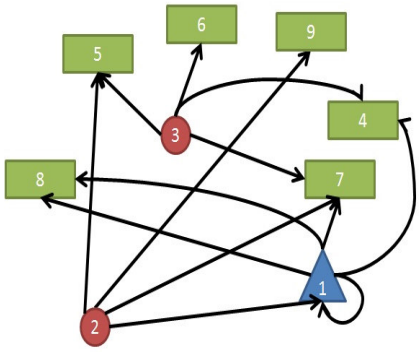
5.2.2 Observed Motifs

Based on the randomizations defined above, we used the motif-finding software packages Mfinder [9] and FANMOD [10] to study the motifs present in our network. We performed motif analysis for



Node	In	Out
1	2	5
2	0	4
3	0	4
4	1	0
5	2	0
6	2	0
7	1	0
8	3	0
9	2	0

Figure 10: Random Graph via Edge Switching

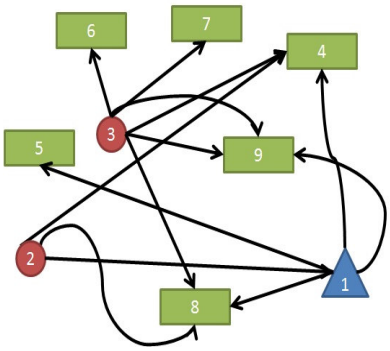


Node	In	Out
1	2	5
2	0	4
3	0	4
4	2	0
5	2	0
6	1	0
7	3	0
8	2	0
9	1	0

Figure 11: Random Graph via Node Replacement I

subgraphs of size 1, 2, 3 and 4 nodes. Figure 13 shows the motifs respectively found for these four classes of subgraphs. While we observe several motifs previously seen in yeast and *E. coli*, we also observe many new motifs [11]. This is biologically expected since we expect the regulatory networks to become more complex in more evolved organisms.

1-node Motif: The simplest motif possible is the 1-node self-loop, i.e., a node that is connected to itself. In terms of regulation, this implies that the entity is regulated by itself and hence must belong to



Node	In	Out
1	1	4
2	0	3
3	0	6
4	3	0
5	1	0
6	1	0
7	1	0
8	3	0
9	3	0

Figure 12: Random Graph via Node Replacement II

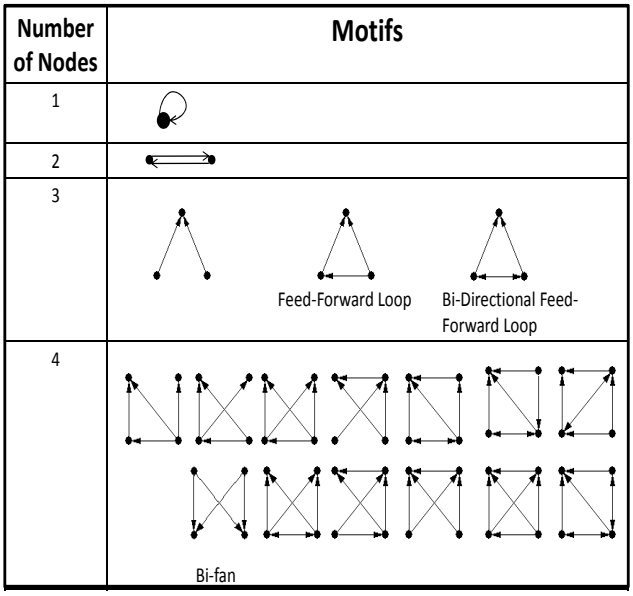


Figure 13: Motifs in C. elegans network

the hybrid class of nodes. Notice that in the artificial network of Figure 4, only node 1 has a self-loop. When the number of self-loops was counted in the original network, we found 15 nodes with autoregulation. However, there were very few self-loops in the randomized networks, giving a p-value of 0.003 for edge replacement and 0.001 for node replacement. This indicates that the self-loop is a significant motif in our network.

To characterize the nodes having self loops, we observed the presence of self-loops as a function of a node's in and out-degree. Figure 14 shows the in vs. out-degree distribution of all nodes in the network with an emphasis on nodes with autoregulation. As we can see from the chart, autoregulation is observed in nodes with both, a high in-degree and a high out-degree ($in-degree > 6\ AND\ out-degree > 10$). Based on this observation, we hypothesize that autoregulation exists in those nodes that are regulated by a large number of nodes and, in turn, regulate many nodes.

2-node Motif: The only 2-node motif we examined was the 2-cycle. While this motif was significant for random graphs generated by Node Replacement II, it was not significant for Edge Switching. Therefore, we did not further investigate this motif.

3-node Motifs: As shown in Figure 13, three 3-node motifs were found to be significant in our network. Of these, the feed forward loop has been studied extensively and is postulated to help eliminate cell response to transient stimuli. The bi-directional feedforward loop, on the other hand, has been postulated to play an important role in development.

4-node Motifs: Previous motif detection studies on *E. coli* and yeast have detected only one 4-node motif, namely the bifan motif shown in Figure 13. However, as seen in this figure, we find 12 other undiscovered motifs in our network. Some of these motifs are likely to be biologically significant and their absence in lower organisms may imply that they play a special role in the *C. elegans* network.

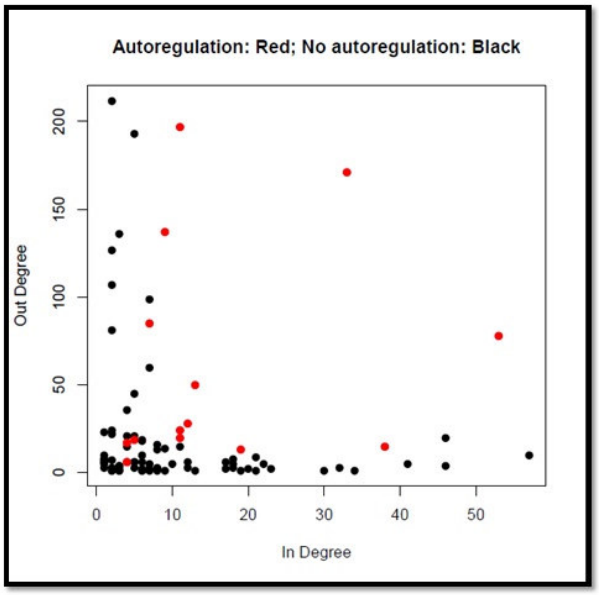


Figure 14: Autoregulation as a function of in and out degree

5.3 Gene Regulation as Directed Acyclic Graphs

To quantify the amount of regulation for a gene, we introduced a new technique involving the construction of directed acyclic graphs (DAGs) with the given gene as the root. The intuition for this technique lies in the fact that a gene can be regulated by any TF that is directly or indirectly connected to it. Therefore, we first find the TFs directly regulating the given gene, then find TFs directly regulating these TFs and repeat this process until we cover all TFs that regulate the original gene. We refer to the generated DAG as the *gene regulation DAG*.

The gene regulation DAGs can be analyzed by various methods and we focus here on measuring the number of edges in the DAGs. Again, the intuition behind this measure is that the removal of any one of the DAG edges can have an adverse effect on gene regulation. Figure 15 shows the frequency distribution of the number of edges in the gene regulation DAGs. This figure shows a striking pattern. There are two clear sets of genes: one set with low overall regulation and one with high overall regulation. This is a surprising result given that the in-degree distribution follows an exponential distribution and does not have a clear demarcation of high vs. low regulation. What is more interesting is that the genes that are highly regulated, e.g., Apoptosis inhibitor proteins, Cyclin T-dependent kinase CDK9, Casein kinase etc, appear to be genes that require fine control, i.e., they are present in specific types of cells, they control very important cellular processes or that their expression changes with time. In contrast, the genes having low overall regulation, e.g. 7-transmembrane receptor, Zn-finger, Fatty acid desaturase, appear to be those that are common across all cells and don't require delicate control.

While the edge distribution in DAGs shows a striking pattern, further analysis is required to determine if this observation is a true property of the network or an effect of the data bias (e.g., few hybrid nodes) or degree distribution.

Table 1: Results of Network Analysis

Property	Value
Clustering Coefficient	0.051
Diameter	11
Characteristic Path Length	3.484
Number of components	2

5.4 Other Techniques

Using the NetworkAnalyzer plug-in for Cytoscape [12,13] we analyzed the network for clustering coefficients, shortest distances between nodes etc. Table 5.4 shows the results from this analysis.

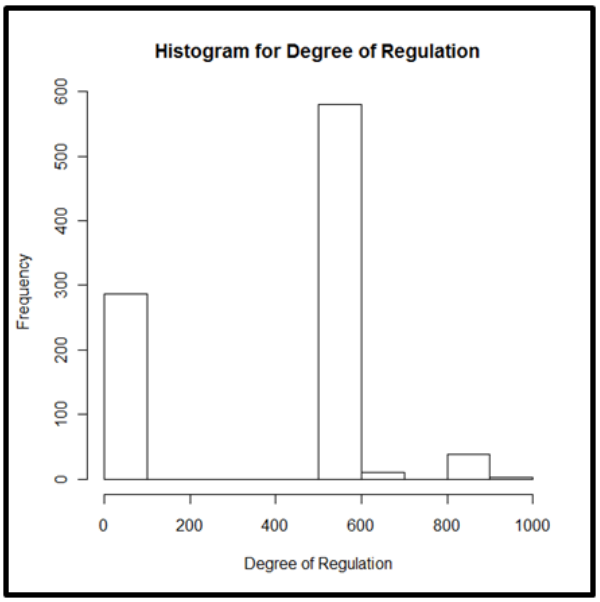


Figure 15: Frequency distribution of DAG edges showing demarcation between high and low regulation genes

Since the network data doesn't currently encode the effect that a TF has on a gene, i.e., whether a TF aids or inhibits a given gene's expression, we built a model to predict this interaction. This model was built on the yeast data [9] because this data includes the information about the type of regulation. Using the Weka tool [14], we built a J48 decision tree to predict the type of regulation based on the degrees of an edge's start and end nodes. This J48 tree had 88% accuracy on the yeast data set. However, adapting this model to our *C. elegans* data was difficult given that the size of both networks and degree distribution of nodes was very different.

6 Conclusion

In this project, we studied transcription regulatory mechanisms in the worm *C. elegans* using a network approach. We modeled transcriptional regulation as a directed graph and used techniques analyzing its degree distribution, motifs, and gene regulation subgraphs to investigate various properties. Our analysis obtained results in agreement with existing literature and additionally found several previously undiscovered network motifs. We also formulated a new hypothesis explaining autoregulation of genes. Our results will serve as the basis for future biological experiments to study gene regulation.

References

[1] A.-L. Barabasi, *Linked: How Everything Is Connected to Everything Else and What It Means*. Plume, 2003.

[2] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature Reviews Genetics*, vol. 5, pp. 101–113, 2004.

[3] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*, 4th ed. Freeman & Co., 2000.

[4] J. A. Bondy and U. S. R. Murty, *Graph Theory with Applications*. Elsevier Science Ltd., 1967.

[5] R. Dobrin, Q. Beg, A.-L. Barabasi, and Z. Oltvai, "Aggregation of topological motifs in the e. coli transcriptional regulatory network," *BMC Bioinformatics*, vol. 5, p. 10, 2004.

[6] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[7] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, pp. 450–461, 2007.

[8] N. J. Martinez, M. C. Ow, M. I. Barrasa, M. Hammell, R. Sequerra, L. Doucette-Stamm, F. P. Roth, V. R. Ambros, and A. J. Walhout, "A c. elegans genome-scale microrna network contains composite feedback motifs with high flux capacity," *Genes and Development*, vol. 22, pp. 2535–2549, 2008.

[9] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: Simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[10] S. Wernicke and F. Rasche, "Fanmod: a tool for fast network motif detection," *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, 2006.

[11] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genetics*, vol. 31, pp. 64–68, 2002.

[12] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. A. B. Schwikowski, and T. Ideker, "Cytoscape: A software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, November 2003.

[13] Y. Assenov, F. Ramirez, S.-E. Schelhorn, T. Lengauer, and M. Albrecht, "Computing topological parameters of biological networks," *Bioinformatics*, vol. 24, no. 2, pp. 282–284, 2008.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10–18, 2009.