Landscape of Gene Regulatory Network Motifs

by

Shawn Striker

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2020 by the
Graduate Supervisory Committee:

Christopher Plaisier, Chair
David Brafman
Xiao Wang

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

The human transcriptional regulatory machine utilizes hundreds of transcription factors which bind to specific genic sites resulting in either activation or repression of targeted genes. Networks comprised of nodes and edges can be constructed to model the relationships of regulators and their targets. Within these biological networks small enriched structural patterns containing at least three nodes can be identified as potential building blocks from which a network is organized. A first iteration computational pipeline was designed to generate a disease specific gene regulatory network for motif detection using established computational tools. The first goal was to identify motifs that can express themselves in a state that results in differential patient survival in one of the 32 different cancer types studied. This study identified issues for detecting strongly correlated motifs that also effect patient survival, yielding preliminary results for possible driving cancer etiology. Second, a comparison was performed for the topology of network motifs across multiple different data types to identify possible divergence from a conserved enrichment pattern in network perturbing diseases. The topology of enriched motifs across all the datasets converged upon a single conserved pattern reported in a previous study which did not appear to diverge dependent upon the type of disease. This report highlights possible methods to improve detection of disease driving motifs that can aid in identifying possible treatment targets in cancer. Finally, networks where only minimally perturbed, suggesting that regulatory programs were run from evolved circuits into a cancer context.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

The result of a large amount of genetic information contained within the genome is the requirement to regulate and express certain genes in a context-dependent manner. One level of control is transcriptional regulation which contain molecules called transcription factors (TFs) that can either activate or repress the expression level of genes through binding to promoter regions. This biological control system also regulates itself, as one TF can control the expression of one or more TFs (including itself, autoregulation). One method to model the landscape of TF regulation is to use a network that links each regulator with their targets. These gene regulatory networks (GRNs) can include protein coding genes, TFs, and other regulatory control elements such as MicroRNA. The structure of these networks show that elements can have multiple inputs and outputs resulting in combinatorial interactions. The interactions result in biological control circuits which can have the complexity to produce engineering control structures such as AND/OR gates and higher order feedback loops. The larger network can be broken down into subcomponents or functional modules. Some identified functions include kernels which function in building the body during embryogenesis , plug-ins which also regulate developmental fate, and I/O switches which activate under specific conditions [1], [2].

Networks can differ between cell types due to the epigenetic regulation of the genome which blocks access to DNA. Cells naturally move from a point of high differential potential, such as stem cells, into more defined cell types like a fibroblast.

Changes in regulation occurs across time and space resulting in the GRN being spatially and temporally defined. This can be visualized by using a Waddington landscape, but in the context of network states [3]. There are various states of regulatory activity which can be more unstable than others. In the case of an unstable state that network would move down the landscape from a position of high potential to a valley where the network is more stable. A normal cell will travel along developmental trajectories that lead to a normal stable state. It has been hypothesized that cancer attractor states are potentials outside of this normal route of development. Mutations can lower epigenetic barriers and increase the possibility that a cell enters a cancer attractor.

Within a GRN recurring patterns of interactions between a small number of nodes can denote a motif of interest. These structured groups, termed motifs, can have varied structures which can perform unique and complex behaviors. As described in the work by Uri Alon, these motifs can generally be separated into two functional network types [4]. The first being sensory networks. Motifs such as feed forward loops (FFLs) can appear in eight varieties and which their function and include sign sensitive delays or pulse generators. The other type of network is developmental. Feedback loops are most functional in this category because they can act to hold memory in a system by either toggling or by maintaining a permanent change in a system. Identifying and understanding the functionality of enriched motifs in a network can give evidence to the functionality of the phenotype of the cell the network is modeled from. In this assessment we utilize enriched motif detection in cancer and other system states to identify possible drivers as well as compare topologies across in a multi-disease method.

This study can be split into two different sections with the first utilizing motifs and patient data to identify disease significant motifs and the other is a more holistic approach which includes non-cancer data. While separate, they are built upon the same computational foundation which will be explained prior to computational results. We began our investigation with the focus on attempting to utilize motif structures in GRNs and attach patient survival data at the back end to identify which motifs could be a factor in the driving of cancerous phenotypes. Alternate approaches have been made which utilize transcriptional regulators to identify downstream genes affected by mutation [5]. These methods utilized patient survival as well as immune infiltration data to identify disease relevant mutations, transcription factors, and genes which were all linked using causal and mechanistic data. Our approach is a simplification of this to highlight the usage of motifs and how applicable leveraging motif detection tools are in this context of disease relevant regulatory structures. The second section is a bifurcation from the foundational pipeline that does not utilize clinical information but rather is a comparative approach. We pulled a variety of datasets from GEO to fit categories of having a GRN which has been altered by disease. An enrichment vector was generated from the output of the motif detector that summarized which motif types were more enriched in the network. We compared how similar each dataset was to an identified normal enrichment vector from a previous study. Finally, we performed hierarchical clustering and identified which datasets were disrupting the network is similar ways.

CHAPTER 2

MOTIF DETECTION PIPELINE

The ability to generate context specific GRNs from RNA expression data will allow us to utilize current motif detection and enrichment tools. Survival associated enriched motifs are excellent potential therapeutic targets. The comparative level of enriched motifs types may discern between a wild type, evolved network, and a perturbed network broken by mutations.

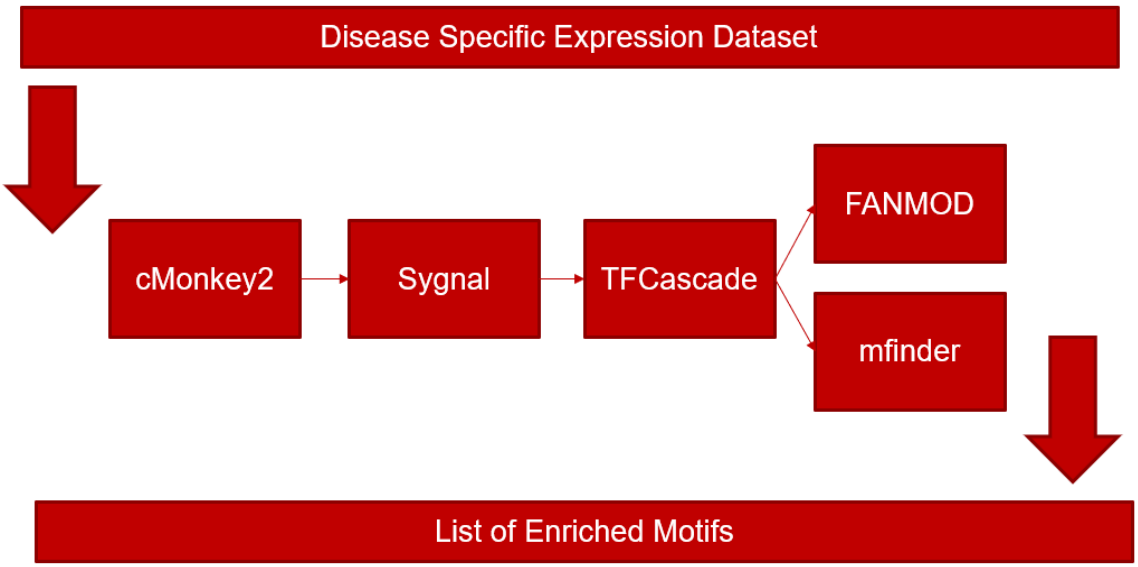


Figure 1: Foundational Computational Pipeline
Process begins with bulk RNA expression dataset. This can include RNA-seq or microarray. cMonkey$_2$ fills a set number of clusters with co-regulated genes called biclusters. These biclusters are filtered by SYGNAL and additional computational tools are run with it. TFCascade is the GRN constructor which pulls dataset relevant TFs from SYGNAL. Then it feeds the filtered and focused network into one of the two motif detection tools which identify enriched motifs in the network.

We used the SYGNAL pipeline developed by C. Plaisier et al as a method to filter the initial selection of TFs which to build each dataset's GRN model [5,6]. We applied this established pipeline to our initial gene expression datasets to identify TFs likely to be

important nodes in the constructed network models. SYGNAL begins by running a dimensionality reduction program called cMonkey$_2$ which mechanistically infers TFs for a specific dataset. These enriched TFs are organized into co-regulated gene clusters called biclusters. cMonkey$_2$ constructs these biclusters by referencing a TF to target gene database which was generated using experimental evidence of DNA-binding to a specific genomic site. Using mechanistic predictions, biclusters are sorted into a predetermined number of biclusters. For the datasets used in this study, excluding TCGA, 3,000 initial genes were chosen to cluster under cMonkey$_2$ using the mean absolute deviation function from the Python package Pandas. Previous usage of the cMonkey$_2$ program identified that there should be on average 30 genes per biclusters. With 3,000 genes we set 200 biclusters for the program to fill.

The next stage in SYGNAL was filtering biclusters. The co-expression biclusters generated by the previous program cMonkey$_{2,}$ were then filtered using quality control metrics and informative associations. The variance explained by the first principle component of biclusters was used as a QC metric for bicluster co-expression. Other computational tools are run within the pipeline which identify correlated TFs from the biclusters using transcription binding motif information. These TFs are the focus for which our GRNs were built. It is important to note that SYGNAL had additional functionalities which were not used for TF selection. The most important of which is disease relevance. The cancer TCGA datasets were the only datasets which included overall survival information for disease relevance.

TFCascade was a Python program built specifically for this study. It utilized the TFBS_DB (https://tfbsdb.systemsbiology.net) [6] a current database mechanistically inferred TF to target gene interactions. The TFBS_DB was used to construct the gbmSYGNAL [6] and panImmuneSYGNAL [5] networks. We used this database to generate a directed GNR for a specific dataset using a list of input TFs. A directed graph contains edges which are modified as arrows meaning the relationship passes one direction. Directed graphs can still have mutual relationships between nodes exemplified by a two headed arrow. TFs can serve as activators or repressors in a regulatory system. TFCascade accounts for this by performing a Pearson correlation test between the regulator and the target to determine if the interaction is significant ($R \geq$ cutoff). We also use the direction of the correlation coefficient to determine if the interaction is activating (positive R, colored green) or repressing (negative R, colored red). The final aspect of this computational step is to expand the possible TF included in the network by including TF families [6]. TFs within the same family will have similar DNA binding motifs and TF family expansion provides the means to infer edges between TFs and target genes for TFs that currently do not have an experimentally determined binding motif.

The final section in the computational pipeline used in this study was dedicated to motif detection. Two different tools were run for information about enriched motifs in the network. The first tool utilized on our data was FANMOD. Developed by S. Wernicke and F. Rasche as a fast motif detection tool when compared to the alternative tool also used in this study by N. Kashtan [7]. FANMOD was initially used during the cancer motif analysis subsection of this study. Mfinder was utilized in the motif landscape

analysis instead to facilitate comparison with a seminal paper by Neph et al which described the observation of a consistent motif enrichment pattern across 41 different cell types[8]. Rather than identifying the specific enriched motifs from the network the program was used to calculate the motif enrichment Z-scores across each of the 13 possible three node motif structures possible. This resulted in each dataset having a vector of enrichment and significance. Both detectors function on the basis of comparing random networks to the input network. One method of generating a random network is to switch edges between vertices, however some of these tools allow the user to specify specific properties of the graph to be retained during randomization. Mfinder was run using its default method to stay consistent with the usage from Neph's paper. FANMOD was run through its command line counterpart wrapped by T. Benyamini and Yoab Teboulle from Tel-Aviv University. The program was run using 10,000 random networks and to set to regard the color of the edges during randomization.

Thus we leverage established computational methods to analyze transcriptional expression data to construct networks comprised of regulatory interactions. We then extract the regulatory factors from the GRN and by integration with a database of TF targeting create a context specific regulatory factor only subnetwork. Finally, motif enrichment algorithms we used to calculate the enrichment values of motifs in the context specific regulatory factor subnetworks, and we compare these enrichment values across cancers, and other normal and disease types.

CHAPTER 3

CANCER MOTIFS

The focus for this portion of the study was dedicated to identifying motifs which could be activated in at least two different ways and provide support for which activation pattern would lead to a clinically worse patient outcome. To begin, the motifs needed to be analyzed dynamically. The basic FANMOD motifs were only static models, which lacks a dimension of pseudo-time. Different approaches can be taken to approximate and model the potential activity based on the interactions within the motifs. Boolean networks have been regularly used, and their dynamics have been significantly analyzed [9]. There are other methods to model the dynamics of networks. The use of differential equations or linear approximations can be used however, the simplicity of Boolean lends itself better to the large number of motifs that required screening by our approach.

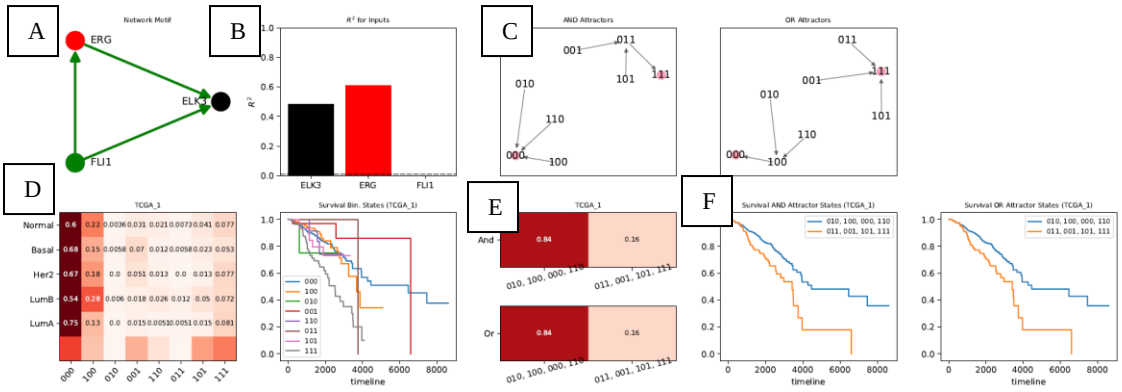


Figure 2: Attractor Motif
The output of a single motif picked to exemplify a feed forward loop in the context of the pipeline. This motif was pulled from BRCA cancer type and had a mild level of disease significance between the two attractor states.

To model our motifs and approximate their behavior across time we identified attractor states to categorize the possible patterns of expression for a motif. An attractor

state is a Boolean steady state where the node values fix in a specific state. If there are no outside factors manipulating the motif, then the motif will flow down into that expression pattern. The example motif shown in figure 2A displays a typical feed forward loop which functions as a sensor where the two input nodes ERG and FLI1 turn on ELK3. If we assume that the combinatorial interaction of both inputs acts as an AND gate, then both inputs would need to be active for ELK3 to begin activation. Figure 2C shows what changes will occur in node activity across a time step. If the starting node activity is represented as 001 then at that time only FLI1 is activated. Over the next time period FLI1 activates ERG leading the next activity representation to be 011. In the final time step the output node ELK3 is receiving the required two inputs, activating the node. The final activity representation is 111 and with every node activated there will be no change over future time steps until an outside force flips one of the nodes leading the motif down another attractor path. In this example we can consider that the motif nodes are either all activated or repressed where the expression states upstream of the attractor will eventually flow down into one of the two attractor states. For this reason, we can categorize this motif as two different expression patterns for any possible combination of expression.

We used gene expression data to highlight specific motif instances whose expression matched the network model and were associated with patient survival for disease relevance. First, we tested the fit for a linear regression model for each node in R using R^2 values. The identified motifs are not isolated in expression meaning there can be multiple other regulation mechanisms including other TFs from the network outside the

motif. The results showed that the median R^2 values for an edge in an enriched motif was below 0.4 with Mesothelioma being particularly bad (Appendix A). This shows that even if a motif was identified as enriched in the GRN that does not mean that motif will lead to a good model for expression. Some outliers can be simplified to exclude outside input factors due to their higher variation explained solely by the motif. However, is unclear how often it is expected that this will be the case. This issue combined with disease relevance is a current problem that has not yet been explored in these studies.

After analyzing the consistency of our motifs, we binarized the patient gene expression data and identified the motif expression state for each patient. Binarization of data was performed in Python by clustering gene expression values into two possible clusters with the higher expressing cluster representing one and the lower expression cluster set to zero. In Figure 2D the heatmap identifies the distribution of patients based on their motif expression pattern stratified by cancer subtype. This example shows that most of the motifs were in the 000 state. If there was differential motif expression preference between cancer subtypes the motif activity might be further investigated for therapeutic potential. The second heatmap shows the categorization of motif expression pattern based on the possible combinatorial gate interactions and the attractor group (Figure 2E). In this example the attractor groups are identical, and thus all AND/OR gate comparisons have the same result, but this is not always the case.

Due to the lower than expected Pearson correlation values for the edges in each motif we had to lower the R^2 cutoff value to 0.5 to begin our search for possible disease relevant motifs with variable effects on patient survival. Turning back to the example in

Figure 2, we generated a single Kaplan-Meier plot containing sample characterizations for every possible motif expression pattern (Figure 2F). We identified that between the two attractor states there is a noticeable difference in survival drop off as the active motif state maintained a more rapid decline in survival. The next two survival curves represent the two attractor groups rather than the individual motif expression pattern states. Again, in this example there is no difference between the AND/OR gate types resulting in the same figure. The results show evidence for a possible divergence in survival between the attractor groups.

This prototype pipeline for detecting cancer relevant motifs searched only three node motifs, and future work could explore motifs with 4 or 5 nodes. In addition, we could incorporate the activity (activator or repressor) in the motif enrichment analysis. The discovered motifs could be explained partially by linear models, and it was shown to be possible to have the motif state be a significant predictor of patient survival. These survival associated motifs are disease specific and would be of great interest in follow up experimental studies. In future work an important area of interests should be the replacement of the linear models used to determine consistency of each node's inputs with a better modeling approach. Another step which identifies to identify they likely downstream hallmark of cancer could be added to facilitate the generation of hypotheses for in vitro studies.

CHAPTER 4

LANDSCAPE OF MOTIFS

Taking a more holistic approach we assessed whether the network topology was altered in disease states. As reviewed before, the connected structure of network motifs can generate different functional abilities included in the variety of feed forward loops and feedback loops. We aimed at encapsulating that structure information by generating a single vector for the frequency of each possible motif type. To simplify the number of possible three node motifs we ran mfinder while ignoring the color of the edges. This resulted in 13 different motif structure possibilities displayed in Figure 3. This method of GRN analysis was developed in the paper by Neph et al which compared 41 different cell type specific transcriptional networks [8]. The results of which identified a pattern of network motifs conserved across each of the cell types and sharing that same enrichment vector with the C. elegans neuronal connectivity network. Utilizing the networks generated from TCGA with other added wild type and non-cancerous disease states we were able to make a comparison of network construction between these different context-specific system states.

Disease/Type	Network Type	RNA Source	Source
C. Elegans	Wild Type	Neural Network	Neph et al.
Env. Exposure (Air Pollution)	Wild Type	Whole Blood	GSE83864
Wound Healing	Wild Type	Skin Grafts	GSE28914
5 Stem Cell Differentiations	Wild Type/Evolved	Floor Plate, Nocireceptor, Dopaminergic Neuron, Melanocyte, Placode	GSE20573, GSE26867, GSE32658, GSE45223, GSE51533
AD	Normal Evolved Response	Postmortem Brain	GSE84422
Staphylococcus Aureus	Normal Evolved Response	Whole Blood	GSE33341
Irritable Bowel Syndrome	Normal Evolved Response	Rectal Biopsies	GSE36701
Listeria	Non-Evolved Network	Whole Blood	GSE67983
Dengue Virus	Non-Evolved Network	Whole Blood	GSE51808
3 HIV	Non-Evolved Network	PBMC, PBMC, CD4+ and CD8+ T Cells	GSE68563, GSE2171, GSE6740
32 Cancer Types	Non-Evolved Network	TCGA	TCGA

Table 1: Datasets Included in Study
The formatted table of the datasets used in the second section of this study.

We hypothesized that there would be a difference in network construction for systems with states of disease that alter the expression of a cell, either through mutation or viral infection. The TCGA data allowed the possibility of having a pan-cancer study, however including these other systems into the analysis would yield a better picture about the change in network structure in the diseased samples. We categorized these network altering diseases as network perturbing, which generate a non-evolved cell phenotype and regulatory structure. Within this category we placed cancer as a definite disease, due to high frequency of mutations, and viral infections such as Dengue virus and HIV as probable diseases. Listeria was included tentatively because of its intracellular infectious

nature, which may manipulate the cellular expression levels and perturb the regulatory network. For a category to include disease states which elicit an evolved expression response to stress or disease we grouped acute, degenerative, and chronic diseases. Infections such as Staph and degeneration of Alzheimer's Disease were estimated to not impact the expressive nature of the cell. The final category called wild type was included to serve as a control for the other datasets. We determined that wound healing, pollution exposure, and the C. Elegans systems were good comparators for normal cell expression data. Stem cell differentiation data were on hand at the time and so were included in this WT category as well. While we did place stem cell data in the WT category, we estimated possible divergence from the other WT datasets due to the nature of expression and functional changes during differentiation.

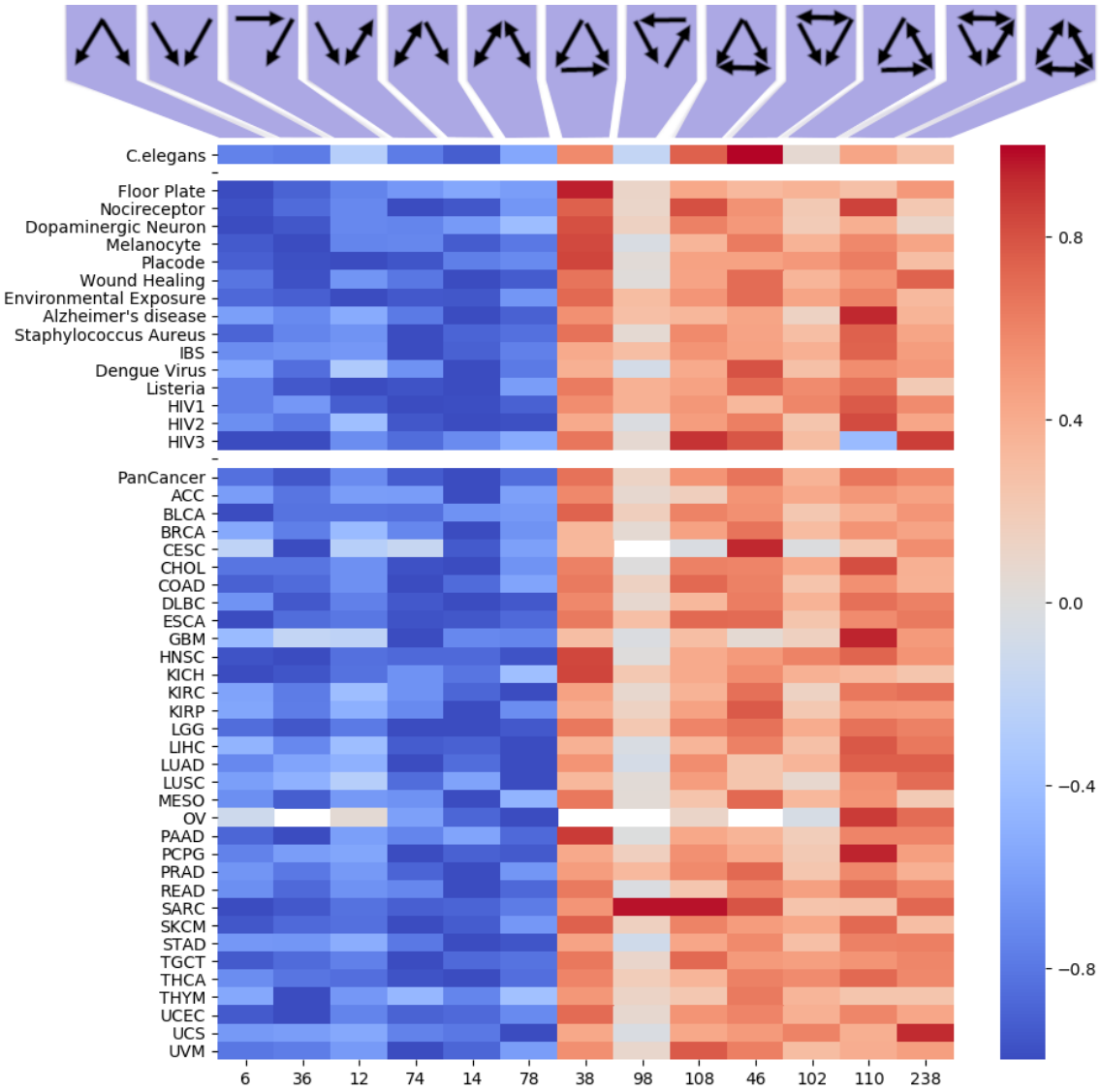


Figure 3A: Topology of Motifs
Shows the 13 different motifs at the level of enrichment for each dataset. C. Elegans (top) represents the conserved enrichment pattern identified in a previous study. Each dataset has been normalized for comparative purposes.

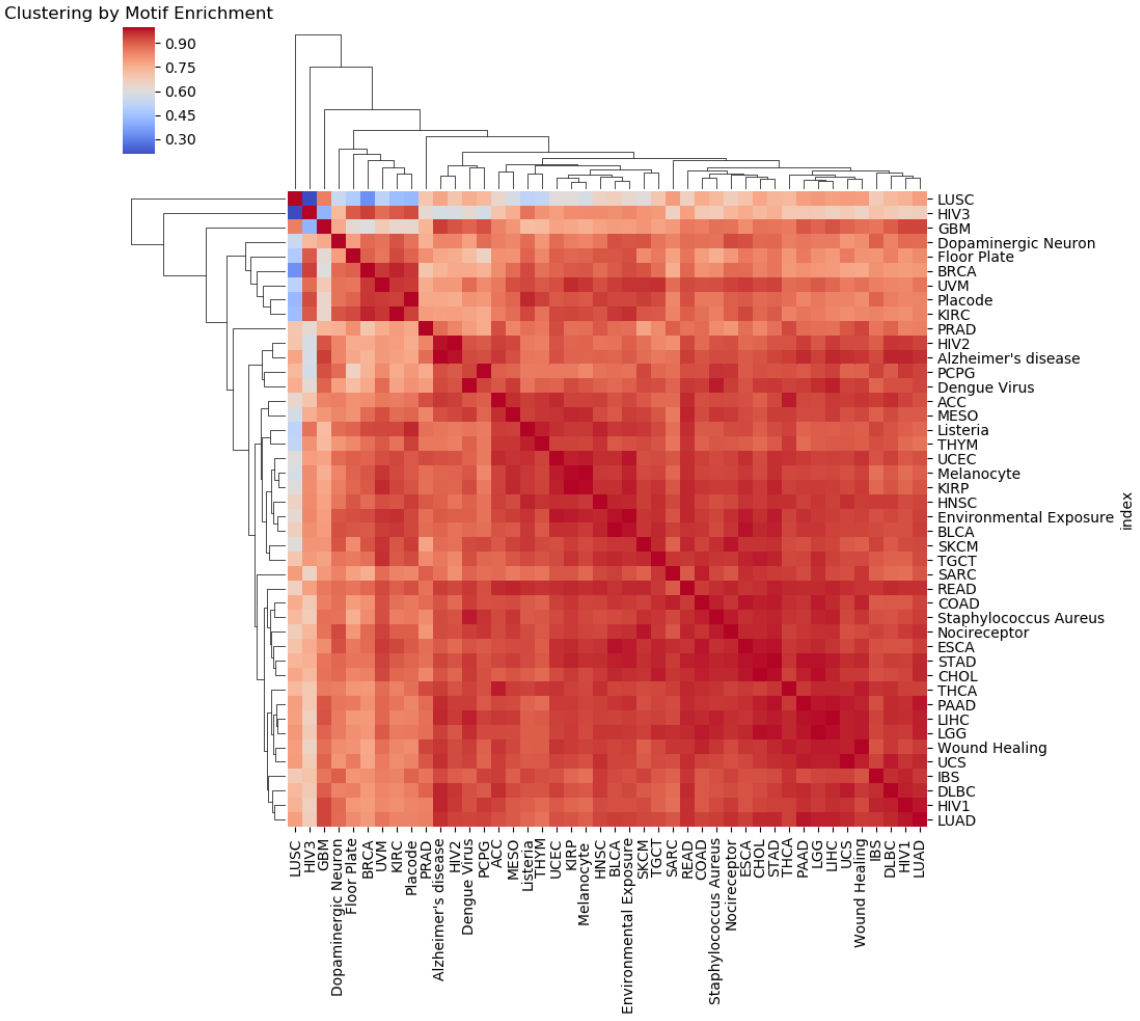


Figure 3B:
Clustering output identifies high similarity between almost every datatype. There is no clustering between cancerous and non-cancerous datasets as estimated.

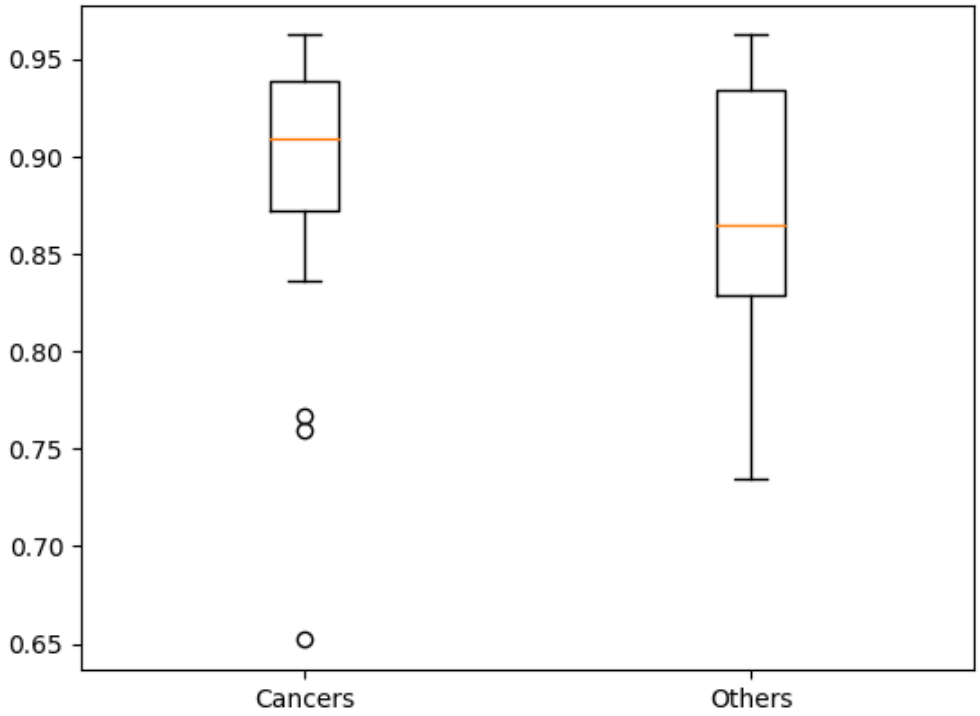


Figure 3C:
Box plot categorizing the correlation values between each dataset and the C. Elegans normal enrichment vector. Categories are split between cancer datasets and non-cancer datasets.

GRN construction was modified for better comparison between the different datasets. One of the differences between the constructed networks was their size. Networks with many edges and nodes will have a larger sample space to identify enriched structures. The motif detector would be able to identify more motifs. If we estimated that some of these perturbed networks would have a portion of their motifs altered, then by including a larger number of common and normal motifs we hypothesize that the disease signal would be drown out. We attempted to limit this problem and the variability of network size by increasing the correlation tolerance to 0.5 then taking only

the top 300 correlating edges. After this small modification we ran the rest of the programs down the pipeline and obtained our motif enrichment data from mfinder. Two specific datasets had trouble with this running setup as OV, CESC, and KICH cancer types both contained one or more “nan” Z enrichment value for their motifs.

The initial results from this multi-disease comparative assessment showed a remarkable resemblance to previous findings by Neph et al [8]. This pattern which was previously observed over multiple different healthy cell types was again identified here. The previous expectation to find deviation between perturbed networks and normal evolved regulatory networks did not appear is any obvious manner. We described a “normal” or WT pattern of enrichment using the C. Elegans enrichment vector then compared it to every other dataset’s enriched motifs. This identified how correlative the two enrichment vectors would be (Figure 3c). This was performed in Python using Pearson’s R correlation value. The results highlighted that both the cancer data sets, and the other data sets had normal motif topologies. Again, OV, CESC, and KICH were omitted from this figure due to poor quality motif results. Outliers in the cancer included LUSC, GBM, and PRAD which were all considered outliers.

The Alzheimer’s Disease datasets appears to be slightly deviated from the normal enrichment vector with a correlation coefficient of 0.83 (p-value = 0.00038). One explanation for this is the quality of the brain samples taken. The known effect of post-mortem samples on RNA quality is well known and its degradation is used in forensic pathology. However, for uses in RNA expression studies post-mortem tissue samples will have lower quality leading to unreliable results [10].

The final analysis of this data was to perform hierarchical correlation using these enrichment vectors. We used the Seaborn package in Python and identified some outliers which were also found to be deviating from the normal motif topology. Our results again identified the fact that these enrichment vectors were all very similar with the few exceptions. LUSC and HIV3 were both the lowest scoring in terms of correlation with a normal motif topology and again we had identified them as deviating in this cluster analysis.

The hypothesis of cancer and other possible expression altering diseases modifying the motif enrichment topology was not supported in this study. Instead this evidence shows that the previously identified topology which is conserved across cell types and organisms is also conserved in various kinds of diseases, including cancers with high mutational rates. It may be possible to expand upon this pan-disease type study of GRNs by comparing other network modules or performing alternate network metrics across each GRN.

CHAPTER 5

DISCUSSION

We utilized two approaches to characterize motifs in the context of systems with disease. In our first section we focused on identifying clinical and experimental applications of motif detection by applying disease relevance to network isolated motifs. We identified three issues which hampered our progress. First, many of the enriched motifs detected contained edges which did not correlate with the original expression data. Motifs with this issue are filtered out with a set threshold and were in the majority. The second issue is disease relevance. Our approach utilized patient survival data to identify differentially expressing motifs with variable survival time. Again, most of the motifs had states that did not have a difference in survival. The final issue encountered was simply about structurally functional motifs. Many of the motifs identified did not belong is the category of a feed forward loop or a feedback loop. Some had only one attractor state and others were constructed with every node positively regulating every other node. Our investigation did not find a perfect motif which avoided all these issues; however, we laid the groundwork for the first iteration of a pipeline to detect relevant motifs. Future efforts should include an experimental validation component. To validate a motif as a factor in cancer growth and progression the correct phenotypic assay would need to be selected. While overall survival helps apply disease relevance it does not narrow what factors of the hallmarks of cancer the motif may modulate. Efforts should be made to integrate motif states with cellular or tissue measurements of immune response, invasiveness, or proliferation. A simple and cancer relevant assay would be of proliferation. Utilizing a

gene editing method such as CRISPR to knock-down or knock-out one of the key regulators in a specific cancer motif we can generate a modified cancer cell line without a functioning oncomotif. A combination of fluorescence-activated cell sorting (FACS) and live cell imaging of motif node abundances can allow the real-time study of modified and un-modified cell populations.

While we did not identify a significant difference between the topology of enriched motifs of cancerous and non-cancerous regulatory networks, we were able to validate the previous finding of a conserved motif topology. This suggests that the perturbations modulate cellular networks, but they effect cellular phenotypic changes through use of evolved network topology. In Chapter four we identified little or no difference between enriched motifs from cancer regulatory networks versus other regulatory network types, then these cancerous networks are being constrained by other factors than typical multicellular evolutionary pressures as they ignore and break multicellularity rules. Possible improvements and additional experimentation of this method of motif analysis could include a rework of network building. To generate a disease-oriented network, utilization of additional data could be helpful. For example, time series data at different levels of infection or inclusion of survival analysis into the SYGNAL pipeline could help secure disease relevance. This might solve the issue of identifying just the common parts and missing the alterations made by the disease in the network.

Overall the use of motif analysis of GRNs is not yet fully actualized. There are many issues associated with attempting to find a solid driver within a single cancer type

or examining motifs in a holistic approach across many disease types. We have produced a first iteration pipeline for disease relevant motifs and validated previous findings across the landscape of motifs. With possible improvements in the study of motifs and with future validating experiments in vitro, more rigorous assessments can be made on the utilization of motifs in research.

REFERENCES

[1] E. H. Davidson and D. H. Erwin, “Gene Regulatory Networks and the Evolution of Animal Body Plans,” *Science*, vol. 311, no. 5762, pp. 796–800, Feb. 2006, doi: 10.1126/science.1113832.

[2] D. H. Erwin and E. H. Davidson, “The evolution of hierarchical gene regulatory networks,” *Nat. Rev. Genet.*, vol. 10, no. 2, pp. 141–148, Feb. 2009, doi: 10.1038/nrg2499.

[3] S. Huang, I. Ernberg, and S. Kauffman, “Cancer attractors: A systems view of tumors from a gene network dynamics and developmental perspective,” *Semin. Cell Dev. Biol.*, vol. 20, no. 7, pp. 869–876, Sep. 2009, doi: 10.1016/j.semcdb.2009.07.003.

[4] U. Alon, “Alon, U.: Network motifs: theory and experimental approaches. Nat. Rev. Genet. 8, 450,” *Nat. Rev. Genet.*, vol. 8, pp. 450–61, Jul. 2007, doi: 10.1038/nrg2102.

[5] V. Thorsson *et al.*, “The Immune Landscape of Cancer,” *Immunity*, vol. 48, no. 4, pp. 812-830.e14, Apr. 2018, doi: 10.1016/j.immuni.2018.03.023.

[6] C. L. Plaisier *et al.*, “Causal Mechanistic Regulatory Network for Glioblastoma Deciphered Using Systems Genetics Network Analysis,” *Cell Syst.*, vol. 3, no. 2, pp. 172–186, Aug. 2016, doi: 10.1016/j.cels.2016.06.006.

[7] S. Wernicke and F. Rasche, “FANMOD: a tool for fast network motif detection,” *Bioinformatics*, vol. 22, no. 9, pp. 1152–1153, May 2006, doi: 10.1093/bioinformatics/btl038.

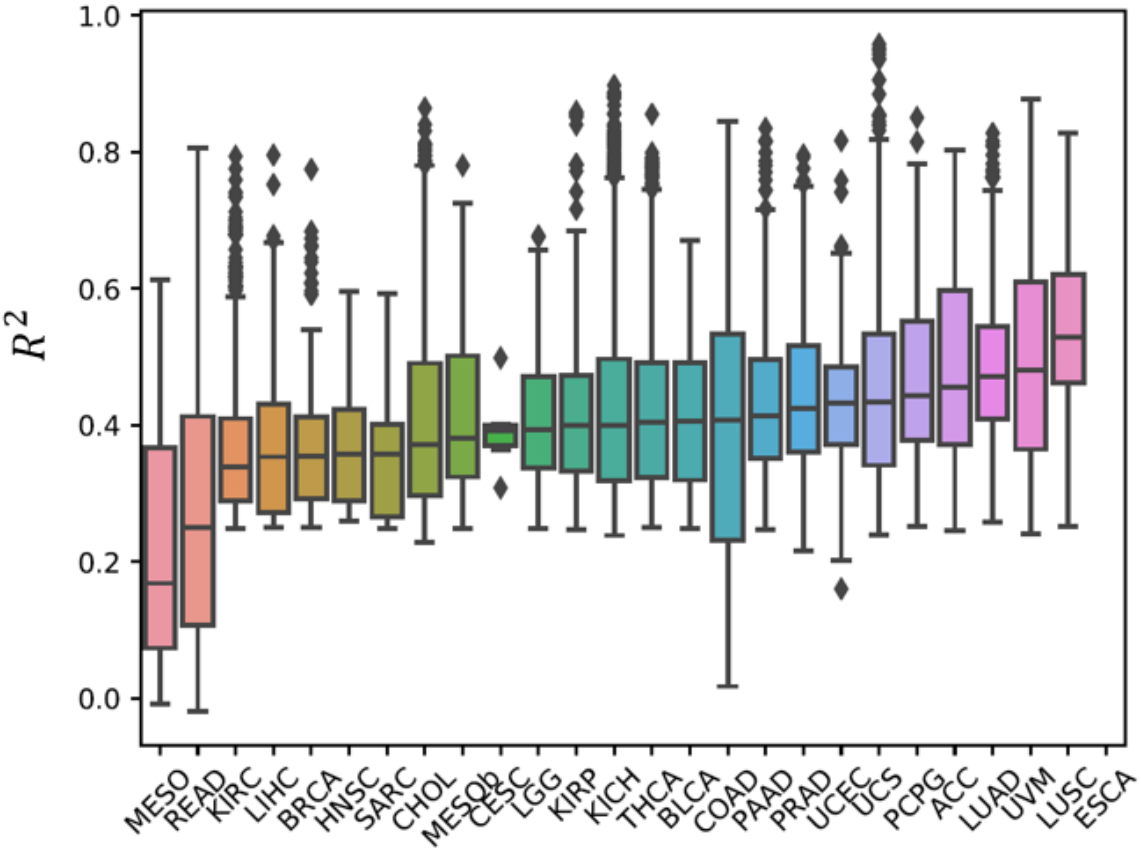
[8] S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos, “Circuitry and dynamics of human transcription factor regulatory networks,” *Cell*, vol. 150, no. 6, pp. 1274–1286, Sep. 2012, doi: 10.1016/j.cell.2012.04.040.

[9] J. X. Zhou, A. Samal, A. F. d’Hérouël, N. D. Price, and S. Huang, “Relative stability of network states in Boolean network models of gene regulation in development,” *Biosystems*, vol. 142–143, pp. 15–24, Apr. 2016, doi: 10.1016/j.biosystems.2016.03.002.

[10] M. Sidova, S. Tomankova, P. Abaffy, M. Kubista, and R. Sindelka, “Effects of post-mortem and physical degradation on RNA integrity and quality,” *Biomol. Detect. Quantif.*, vol. 5, pp. 3–9, Sep. 2015, doi: 10.1016/j.bdq.2015.08.002.

APPENDIX A

MOTIF CORRELATION WITH EXPRESSION



The correlation of edges and between the motif and the expression data is variable and depends on the dataset. This is a result of data quality as well as correct representation of the network with real expression data. The difference between a good quality dataset and a lower one can be seen between MESO, which comes from TCGA, and MESOb which comes from a mutually exclusive study using RNA-seq. What this figure represents for this study is the issue of validating detected enriched motifs. While it can depend on quality, most edges in motifs have a low correlation value between the regulator and target node.