Xander Krohannon

B528 – Computational Methods for Analysing High-throughput Data in Biology

Final Project Report

**A Comparison of GRN Models Incorporating Proteomic Data in *Escherichia* *coli***

**Abstract:**

In the modern world, networks are being utilized in a broadening number of fields, especially biologic disciplines. A common use of networks in a biological context include Gene Regulatory Networks (GRNs). GRNs come in many forms, modeling interactions between genes, RNA molecules, proteins, complexes, or other biological processes. While it is known that these networks are incomplete, due to inherent incomplete knowledge of biologic systems, an open question is if interactions at a one-level is sufficient to disturb the underlying structure at another level. This paper aims to answer that very question by comparing the results of an analysis of GRN from *Escherichia coli,* with the result of the same underlying network with the addition of known interactions between *E. coli* transcription factors.

**Background:**

GRNs are vital tools in allowing scientists to visualize and elucidate understanding from the complex series of interactions that dictate behavior in cells. However, tools and models are only as good as their ability to accurately predict the behavior of their ascribed system. Small scale GRNs have illustrated a robust ability to accurately predict the behavior of many biological systems under various perturbed conditions [1]; but the accuracy of these networks as they are scaled up to model whole-cell systems are yet to be determined. To that end, in their 2019 paper [2] Larsen et al performed a comprehensive comparison of *in silico* gene expression models of *E. coli* with wild type gene expression data from *E. coli*. The results clearly illustrated a large discrepancy between them, namely that current *in silico* models dramatically over estimate the effect of repressors and activators on biologic systems, despite those parameters being determined through experimental procedures. The authors conclude that *in silico* models could be improved by implementing data from multiple levels of omics.

This study aims to serve as a proof of concept to Larsen’s conclusion by comparing simulations using networks of all known interactions between transcription factors and genes in *Escherichia coli* with simulations of the same network with the addition of transcription factor and transcription factor interactions. Those simulations were performed using RACIPE, a C based simulation modeling software presented in Huang et al’s 2018 paper [3]. This program works by taking in a network topography file, including activator and repressor information, and creates a series of ordinary differential equations (ODEs) with corresponding hill functions. The tool then performs a large number of simulations by randomly assigning initial concentrations and kinetic parameters to each network interaction and solving the series of ODEs for a set of steady state solutions. The initial concentrations and kinetic parameters are assigned based upon statistical analysis of biologic values. Every 100 simulations, the initial concentrations and kinetic parameters are changed, until the number of simulations are reached. The program then returns a .dat file that contains the log-2 normalized value for each gene in the network, in the order that they appeared in the network topography file.

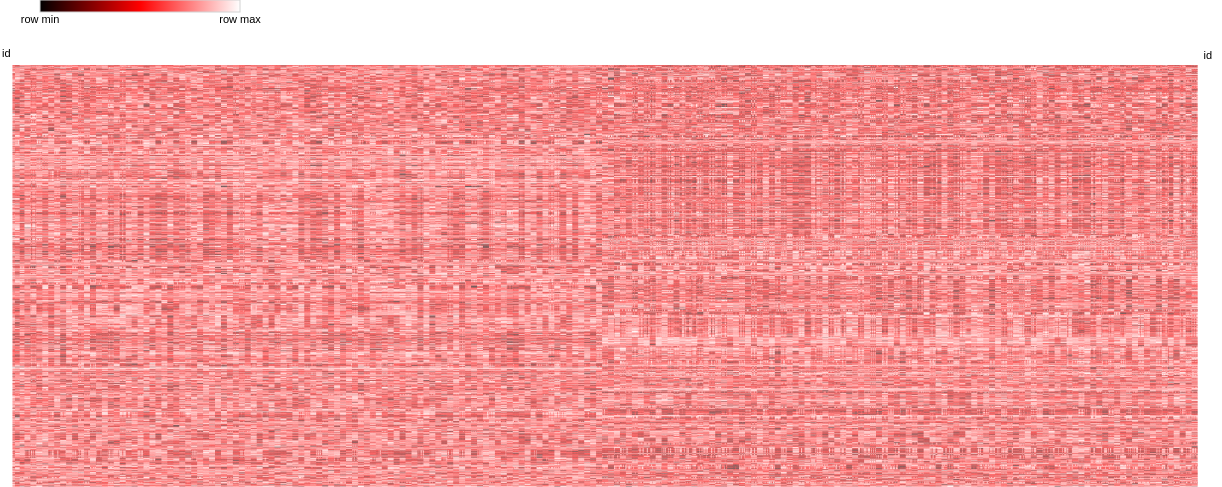
**Data:**

The data used in this study was downloaded from RegulonDB, which includes a comprehensive list of all known transcription factor – gene interactions and transcription factor – transcription factor interactions. These files were downloaded and stored as ‘TF\_Gene Interaction Network.txt’ and ‘TF\_TF Interaction Network.txt’.

**Methodology:**

The ‘TF\_Gene Interaction Network.txt’ and ‘TF\_TF Interaction Network.txt’ files were parsed and analyzed using ‘Network Topography Maker.py’ to create .topo files (‘TF-Gene.topo’ and ‘Omics.topo’) for use in RACIPE. RACIPE was used to perform 10,000 simulations for each of the topography files, generating 990 unique solutions for the ‘TF-Gene.topo’ and 1000 unique solutions for ‘Omics.topo’. Of those 990 unique solutions in ‘TF-Gene.topo’, RACIPE identified 1 unique set of 4-modal steady states, 2 sets of 8 modal steady states, and 97 sets of 10 modal steady states; where as all solutions for the ‘Omics.topo’ were 10 modal steady states.

Each of the results from both sets of simulations were parsed and incorporated into a single data frame using ‘Model Parser.py’. This data frame was 1874 rows, corresponding to genes, by 1990, corresponding to simulation results. The data frame was then exported to ‘Steady\_State\_Analysis.csv’ so that it could be visualized using Broad Institute’s online tool, Morpheus [4]. The result is shown below in Figure 1.

****Figure 1. Steady State Heat Map

In order to analyze the heat map of that scale, the data needed to be bi-clustered. In order to determine the optimal number of the bi-clusters for this heat map, the elbow method was employed. This was performed in ‘Model Parser.py’, by performing Kmeans bi-cluster algorithm from sklearn altering k from 2 to 43. For each model, the sum of squared distance between each point and its closest cluster centroid was calculated and plotted. Using the elbow method, the optimal number of clusters was determined to be 15, see figure 2.

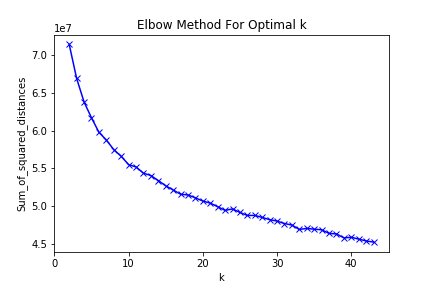


Figure 2. Determination of k for Kmeans bicluster

Using the optimal k, the data was bi-clustered using Kmeans algorithm on Morpheus and exported as ‘Cluster\_Steady\_State.gct’. The bi-cluster was analyzed using ‘Cluster Analyzer.py’, and the composition of each cluster, both genes and steady states, were studied. After identifying the list of genes contained in each cluster, a gene ontology query was performed using Panther[5]. The resulting bicluster image is shown in Supplementary Material labeled ‘Clustered\_Heat\_Map.png’.

In order to reduce the complexity of the heat map, the Chung and Church method was employed, with alpha value as 1, and beta value as 0.2. This resulted in reducing the number of rows from 1874 to 870, and number of columns from 1990 to 1000. The resulting heat map is shown below, Figure 3.

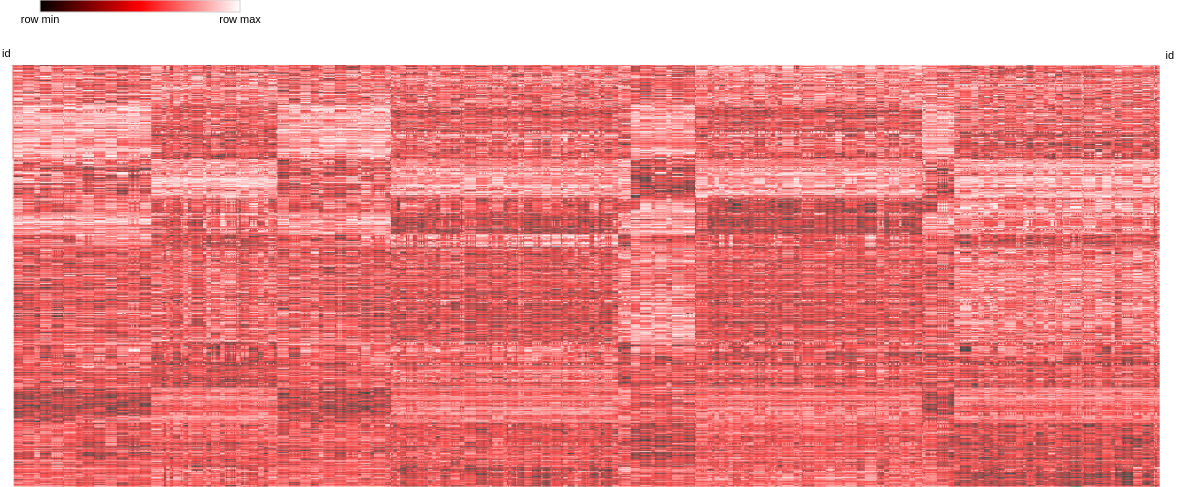


Figure 3. Revised Heat Map

**Conclusion:**

During the analysis of the biclusters, the content of each cluster was studied. The gene ontology study of the clusters showed no significant overrepresented biologic process, see Supplementary Material / Gene Ontologies. It is still possible that there is some biological relevance to these genes having similar expression profiles in various predicted steady states, but that would require additional research beyond the scope of this study. Each of the 4 modal steady states were clustered into a single set with a set of 8 modal steady states, and a variety of 10 modal steady states. The second set of 8 modal steady states were also clustered with a set of 10 modal steady states. These results seem to indicate that the behavior of the lower modal steady states are not unique, and can be captured by higher modal steady states. A more interesting observation however, was that steady state solutions from the different experiments were never clustered together. This would seem to indicate that Larsen’s hypothesis is likely correct; implementing multiple levels of omics data is sufficient to significantly alter the behavior of current models, and perhaps improve the accuracy of *in silico* models.

**Sources:**

[1]. Huang, B., Jia, D., Feng, J., Levine, H., Onuchic, J. N., & Lu, M. (2018). RACIPE: a computational tool for modeling gene regulatory circuits using randomization. BMC systems biology, 12(1), 74. doi:10.1186/s12918-018-0594-6

[2]. Gutiérrez-Ríos, R. M., Rosenblueth, D. A., Loza, J. A., Huerta, A. M., Glasner, J. D., Blattner, F. R., & Collado-Vides, J. (2003). Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome research*, *13*(11), 2435–2443. doi:10.1101/gr.1387003

[3]. Larsen, S. J., Röttger, R., Schmidt, H., & Baumbach, J. (2018). E. coli gene regulatory networks are inconsistent with gene expression data. Nucleic acids research, 47(1), 85–92. doi:10.1093/nar/gky1176

[4]. https://software.broadinstitute.org/morpheus

[5]. Huaiyu Mi, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, Paul D Thomas, PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools, *Nucleic Acids Research*, Volume 47, Issue D1, 08 January 2019, Pages D419–D426 https://doi.org/10.1093/nar/gky1038