5-2-2017

Comparison of the regulatory dynamics of related small gene regulatory networks that control the response to cold shock in Saccharomyces cerevisiae

Natalie Williams
Loyola Marymount University, nwilli31@lion.lmu.edu

LMU | LA
University Honors Program

Comparison of the regulatory dynamics of related small gene regulatory networks that control the cold shock response in *Saccharomyces cerevisiae*

A thesis submitted in partial satisfaction

of the requirements of the University Honors Program

of Loyola Marymount University

by

Natalie Williams

April 28, 2017

Introduction

Saccharomyces cerevisiae, baker's yeast, is a model organism for systems biology. As such, knowledge about biological mechanisms observed in yeast can be applied to other eukaryotic organisms, including humans. Yeast serves as a good a model organism because its relatively small genome of approximately 6000 genes is easier to study than the human genome of approximately 22,000 genes (Goffeau et al. 1996). Further, because of the vast number of experiments performed by the yeast community, deletion strains,other molecular tools, and datasets are readily available.

Many of these tools were amassed by subjecting yeast to varying environmental, genomic, and growth conditions. To better understand the stress response, investigations have grown yeast cultures in different environments, including anoxic conditions and media with different pH and salt concentrations, and nutrient loads (Geistlinger et al. 2013). For example, ter Schure et al. 1995 investigated how altering the ammonia concentrations of the media affected yeast's metabolism of nitrogen (ter Schure et al. 1995). A common examination investigates temperature's effects on yeast.. Specific heat shock proteins have been identified in yeast and act as chaperones to stabilize other molecules and proteins to increase yeasts' survival in warmer temperatures (Feder & Hofmann 1999; Jakob et al. 1993). While the effect of environmental stresses on specific cellular functions is generally well studied, yeast's responses to cold temperature stressors remain unknown.

Characterization of the cell's response to cold shock is limited. When yeast was introduced to cold shock at 10°C , a physiological change in the rigidity of the phospholipid bilayer was one of the many observed results (Aguilera et al. 2007). Other responses include impairment of ribosome function and protein synthesis as well as a decrease in enzymatic

activities (Schade et al. 2004). From the limited studies subjecting yeast to cold shock, there is no equivalent set of proteins that controls the response to this environmental stress for all organisms. However, similar to any other environmental stress response, yeast responds to cold shock conditions by changing its level of gene expression (Schade et al. 2004).

Yeast cells alter levels of gene expression by using regulatory transcription factors. Transcription factors are proteins that bind to regulatory DNA sequences to influence a specific gene's expression (Chen & Rajewsky 2007). Transcription factors can act as either activators, which increase a gene's expression, or act as repressors, which decrease a gene's expression. Transcription factors themselves are proteins encoded by genes (Chen & Rajewsky 2007). In a process called combinatorial control, these regulatory proteins, acting as activators and repressors, bind to DNA sequences of a particular gene to "vote" on whether to alter the gene's expression as a response to environmental changes (McKenna & O'Malley 2002).

To better understand yeast's response to cold shock, the transcription factors involved in this regulatory response must be identified. In the Dahlquist Lab, growth experiments were performed on strains that had a deletion of a transcription factor believed to be involved in the cold shock response. If the yeast deletion strain has impaired growth at cold temperatures, it implied that the transcription factor plays a key role in yeast's survival in cold conditions via regulating gene expression. Microarray data was obtained from the deletion strains that showed impaired growth at cold temperatures to identify which genes' expression changed. A DNA microarray is a technology that binds thousands of nucleic acid sequences in a mixture via hybridization and later detection of that hybridization (Bumgarner 2013). Genes' significant changes in expression are then used to construct a gene regulatory network for cold shock response. Microarray data for the wild type and five deletion strains is obtained from the wet lab.

The yeast undergoes cold shock (13°C) during their exponential growth phase. Microarray collections at cold shock occur at 15, 30, and 60 minutes; after 60 minutes, recovery begins as the yeast is placed back in optimal growth temperatures. From the microarray data, the changes in gene expression compared to t_0 are measured.

Gene regulatory networks (GRNs) are the set of transcription factors that control the level of expression of genes encoding other transcription factors. Yeast, as a model organism, contains approximately 250 transcription factors that regulate its roughly 6000-gene genome. Information about GRNs is contained in the YEASTRACT database, which combines different conditions to construct the network. This information ranges from DNA-binding evidence, expression evidence, and regulatory motif sources (YEASTRACT). With this data, GRNs are generated from the clusters of genes with similar differences in their expression. From the YEASTRACT database, six related small GRNs were generated, ranging in size from 15 – 20 genes and 27 – 30 edges.

Genes within GRNs have connections with other genes within the network such that when expression of a transcription factor changes, the expression of the target gene will also be affected. Mathematical systems have been used to describe the relationship among genes within GRNs through modeling the dynamics. Biological systems behave in a nonlinear manner; thus, mathematical models, which take advantage of ordinary differential equations, serve to describe the system's dynamics (Smolen et al. 2000; Ingalls 2013). Vu & Vohradsky (2007) utilized ordinary differential equations (ODEs) to model the dynamics of transcription factors during the yeast cell cycle. Their model successfully recapitulated what is known about cell cycle regulatory dynamics. Similarly, the Dahlquist Lab uses ordinary differential equations to

describe the dynamics of small GRNs in yeast. However, the Dahlquist Lab is attempting to model a system where behavior is less well known (Dahlquist et al. 2015).

These GRNs are then put into GRNmap, the Dahlquist Lab's code in MATLAB that uses ODEs to estimate parameters that affect the overall expression levels of an individual gene. These parameters include the production rates, weight parameter that denotes one transcription factor's influence on another in the network, and threshold expression. The threshold b is the point at which the production of the gene is switched on or off (Dahlquist et al. 2015). The challenge in estimating these parameters is fitting the equation to gene expression data from the microarray. Because of the combinatorial control of transcription factors in a GRN, the individual weight's effect on the target gene has an indefinite number of possibilities. The threshold of when production is switched on or off also has a multitude of possible values. The least squares approach combats this estimation problem by comparing model outputs to the observed data and minimizing this discrepancy (Dahlquist et al. 2015). The same method, using observed data to help guide model outputs, was used by Kim et al., whose team sought to infer a GRN from noisy and temporal data points (2007).

In the course of my investigation, I have performed statistical analyses of the recurring data. I used the data to infer six related small GRNs from the YEASTRACT database. I ran GRNmap to estimate parameters for regulatory dynamics in this network. To validate my results, I generated more related GRNs for 30 random networks. I found that the database-derived networks performed better than the random networks. This conclusions suggests that the databased-derived GRNs explain at least part of the transcriptional response to cold shock in yeast.

Materials & Methods

Statistical Analysis of Dahlquist Lab Microarray Data

The Dahlquist lab has produced DNA microarray datasets for six strains of yeast – the wild type (wt) and five deletion strains, Δcin5, Δgln3, Δhap4, Δhmo1, and Δzap1. To be modeled, the data must be normalized. R and the limma package were used for within array and for between array normalization for all chips. The R code and more details can be found on the Dahlquist Lab's OpenWetWare site, Dahlquist:Microarray Data Analysis Workflow (www.openwetware.org/wiki/Dahlquist:Microarray_Data_Analysis_Workflow).

After within and between array normalization of the microarray data, , a within-strain ANOVA test was performed using Microsoft Excel. The Benjamini & Hochberg p value correction was used to identify genes with significantly significant different log fold expression changes due to the multiple testing problem.

STEM Clustering of Microarray Data

After the statistical analysis, I focused on the results from the wt strain. Genes that met the Benjamini & Hochberg $p < 0.05$ were selected for clustering. The average $\log_2$ fold gene expression values for each time point were input into the STEM software (Ernst & Bar-Joseph 2006) The STEM software separates genes into clusters based on shared expression profiles (e.g.. activation for the first 30 minutes and no change from 60 minutes until recovery). STEM classifies clusters on significance if more genes belonged to the cluster than would be expected by chance. One of the significant profiles was then chosen for further analysis.

Construction of Gene Regulatory Networks

Genes belonging to the selected cluster from STEM were then submitted to the YEASTRACT database to determine potential regulators of these target genes. Potential regulators were queried based on the setting, "Document" and "DNA binding plus expression evidence". Potential regulators of the target genes were then returned in order of significance. A

regulator was considered significant if it regulated more genes in the cluster than would be expected by chance.

Next, the 25 most significant transcription factors were submitted to YEASTRACT to determine the regulatory relationships between them. If the following genes were not in the list of significant transcription factors, then they were added to the list: CIN5, GLN3, HAP4, HMO1, SWI4, and ZAP1. These genes are over for which the Dahlquist lab has cold shock DNA microarray data. YEASTRACT generated an adjacency matrix with the set of transcription factors, placing "1's" in cells where a regulatory relationship was found with another transcription factor and "0's" in the cells where there was no relationship. This adjacency matrix of regulatory relationships of genes within the same network is the gene regulatory network provided by the microarray data. Because symmetry is required for adjacency matrices, if any gene as a regulator and target did not have a sum of 1 or greater across the rows and columns, it was not correct in the network and was deleted from the adjacency matrix. The regulators were located in the columns while the target genes were placed in the row.

Construction of the Input Workbook for GRNmap

The adjacency matrix was then placed into a new Excel workbook. The input workbook contained worksheets with the following data: initial guesses for the production rates of each gene; degradation rates for each gene; and initial guess for the threshold b values. The $\log_2$ fold expression data for each gene and strain for the cold shock time points (15, 30, and 60); two sheets containing an identical adjacency matrix for the GRN. The optimization parameters sheet contains parameters for MATLAB such as max iterations, the penalty term, and which model to run – Sigmoidal or Michaelis-Menten. For this analysis, the sigmoidal was used for the modeling.

After searching the literature, RNA half lives reported by Neymotin et al. (2014) were used to compute degradation rates. To obtain the degradation rate, the natural log of 0.5 was divided by the gene's reported mRNA half-life. For genes with missing mRNA half-life data, the median half-life of 202 regulatory TFs listed in Harbison et al. (2004) was used. To calculate the initial guesses for the production rate, the degradation rate was used. The production rate for each individual gene was twice the degradation rate.

Log_2 fold change expression data was kept for certain genes, time points, and strains. For example, for the wt strain, there were four 15 time points. For one gene in the network, the third 15 minute replicate value was missing. Because the model cannot compute with missing data points, the three other expression levels for the 15 time point was averaged and used for the missing value. By using the average, it allowed for the model to continue to run. Further, depending on the genes within the network, the expression data for a specific strain was either included or excluded. For instance, in Db4, the network lacked ZAP1 as a transcription factor. As a result, the Δzap1 deletion strain data was not included in the input workbook.

The two network sheets are next. The first network sheet includes the symmetrical adjacency matrix. The regulators or transcription factors are the column headings while the target genes, the same as the transcription factors acting as regulators, are the rows. The sheet after network contains the same adjacency matrix; however, instead of showing the regulatory relationships between transcription factors and their target genes, this sheet contains the initial guesses of the influence of regulatory proteins on their target genes. For this analysis, all the initial guesses were set to "1".

The optimization parameters sheet includes parameters necessary for MATLAB to run the GRNmodel.m code. The first parameter is the alpha value, which is the penalty term weight.

The next parameter is kk_max. kk_max is the number of times allotted to re-run the optimization loop because restarting this process can improve the model's performance. MaxIter is the number of time MATLAB iterates through the optimization scheme. The TolFun tells the program the smallest difference between two least squares evaluations before it determines that there will be no improvement in modeling the dynamics. MaxFunEval is the maximum number of times the software will evaluate the cost of the least squares output. TolX alerts the program how close successive least squares cost evaluations should be before it determines that no improvements are being made. The production_function will either be sigmoidal or Michaelis-Menton. This function determines the type of production used to model a gene's production and whether it will be activation or repression. The L-curve will run sequential rounds of values to estimate various alpha values. For this analysis, it was set equal to 0 to turn this function off. estimate_params is the function used to estimate the parameters, such as production rate or threshold b; for this analysis, it was set to "1" so that the program would estimate the parameters. The make_graphs parameter determines whether or not figures will be output or not; if graphs are wanted, then it should be set equal to "1". The fix_P and fix_b are used to estimate the production rate and the threshold b value, respectively, and both should be equal to "1" so as to allow for their estimation to occur. The expression_timepoints parameter is a row of containing a list of time points when the data was collected (15, 30, and 60). The strain row consists of the strains with their expression data included in the workbook. The last parameter is the simulation_timepoints. This parameter is a row that has time points to evaluate the differential equations to generate simulated data.

The last sheet is the threshold_b sheet. In this sheet, the initial guesses for the threshold value for each gene are listed. The threshold value is the point at which gene expression is either turned on or off. In the value column, all the initial guesses should be set to "0".

Generation of Random Networks Related to Db5 Network

To generate random networks, an R script written by B. Klein was utilized (https://github.com/kdahlquist/DahlquistLab/tree/master/R_scripts). Random networks are networks that are similar to a "real" network; however, they are random because although they contain the same numbers of nodes (genes) and edges (Alon 2007), the connections between these nodes are randomized. The data used to generate these random networks came from Db5 network. 21 random networks were generated via the R script with 15 nodes and 28 edges. Nine random networks were previously generated in Excel using the following formula in each cell of the adjacency matrix: =IF(RAND()<0.1134,1,0).

Ordinary Differential Equations Model: GRNmap

To model the dynamics of the GRN, GRNmap uses a script in MATLAB to calculate the parameters that best fit the observed data. The mathematical model uses the following basic equation:

$$x_i(t) = p_i(x(t)) - d_i x_i(t) \qquad \text{eq (1)}$$

in which, P_i is the production rate of the specific transcription factor and $d_i x_i(t)$ is the degradation rate of the transcript multiplied by the concentration of the particular gene (Dahlquist et al. 2015). The degradation rate was computed from mRNA half-lives taken from the literature (Neymotin et al. 2014). The production rate, P_i, of the gene transcript depends on the influence of transcription factors within the network that regulate it. The production rate is based on a sigmoidal model proposed by Vu and Vohradsky (2007):

$$p_i(x(t), \theta) = \frac{p_i}{(1+e^{-\sum w_{ij}(x_j(t)-\tau_{ij}i)})} \qquad \text{eq (2))}$$

where P_i is the maximum expression rate for a particular transcript. w_{ij} is the weight of influence gene j has on gene i (can be activation or repression); and τ_{ij} is the threshold (threshold b) at which production is turned on or off (Dahlquist et al. 2015).

To force the function above to a singular value, a least squares error approach was used with an alpha and penalty term.

$$E = \alpha\|\theta\|^2 + \frac{1}{Q}\sum_{\tau=1}^{Q}[z^d(t_r) - z^c(t_r)]^2 \qquad \text{eq (3)}$$

The alpha value was determined empirically from the "elbow" of an L-curve. The elbow of the L-curve allowed for the best estimation of parameters without taking too much time. The penalty term combined the estimated parameters: production rate, weight value, and threshold b value. The first term in the summation is the experimental or observed data from the microarray. The second term in the summation is the simulated data, which is the solution to the ordinary differential equation. This E or LSE value describes how well GRNmap modeled the overall dynamics of the entire network.

$$\text{MSE} = \frac{1}{Q}\sum_{t=1}^{Q}\left[z^d(t_r) - z^c(t_r)\right]^2 \qquad \text{eq (4)}$$

The mean square error, MSE, showed how well the model fit each individual gene. This value is automatically included in the output workbook. However, to calculate the minimum (min) MSE, the average $\log_2$ fold change for each gene across each time replicate (15, 30, and 60) was computed. Then, the difference between each individual time point and the average $\log_2$ fold change for the time points was calculated for individual genes. This difference was then squared for each gene. After the differences were squared, they were summed across all time points. For wt, there are 13 time points, so the sum of 13 individual squared differences was

computed. Next, this sum was divided by the number of time points for the specific data. For wt, there were 13 time points, so the sum of the squared differences was divided by 13.

Running GRNmap and GRNmodel in MATLAB

The most recent published version of GRNmap v.1.4.4 was used for running the model. GRNmap can be downloaded from the GRNmap website (http://kdahlquist.github.io/GRNmap/downloads) under the Executable or Source Code headings. To run the GRNmodel, GRNmap source code must be open in MATLAB. For this procedure, MATLAB 2014b was used. After GRNmap is opened in MATLAB, the GRNmodel.m code was selected and the Run button was clicked. A file dialogue box opened to select the input workbook file. Run time for the model varied, being as short as 25 minutes, and as long as 3 to 4 hours for larger networks.

Visualization of Result with GRNsight

GRNsight visualizations were also used to analyze the weight parameters. GRNsight provides automatic layout of medium-scale networks and can be found at the following address: http://dondi.github.io/GRNsight/index.html (Dahlquist et al. 2016). GRNsight shows the strength or magnitude of the weight of influence a transcription has on its target gene. The thicker the edge, the stronger the regulatory relationship. The colors of the edges provide information about the regulatory relationship. Cyan represents repression, and a blunt arrow head also signified repression. Magenta represents activation, and the arrowhead is a normal, pointed arrow head.

Results & Discussion

Statistical Analysis of Dahlquist Lab Microarray Showed Almost a Third of the Genome had Significant Changes in Gene Expression

The focus of any statistical analysis was on the wild type strain. Table 1 lists the number and percentage of genes with expression significantly different than at any time point. These are the Benjamini & Hochberg corrected p-values. These results suggest that drastic changes of gene

expression occur in yeast cells as a response to cold shock. The percent difference between each strain compared to the wild type shows that without the specific transcription factor in that particular strain, there is detrimental growth in the yeast cell.

Table 1: Number of and Percentage of Genes with Significant Changes in Expression

Strain	Wild type	Δcin5	Δgln3	Δhap4	Δzap1
Significant genes	1936 (31%)	1683 (28%)	1683 (28%)	1794 (29%)	1859 (30%)

STEM Clustering of Microarray Data Clustered Six Significant Profiles with Genes that had Significant Changes in Gene Expression

The STEM output for the wt data showed six distinct significant gene profiles (Fig 1) shown in color. The profile chosen for analysis was Profile 45, which can be seen in Figure 2. The following Gene Ontology Caregories were over-represented amongst genes in this cluster.

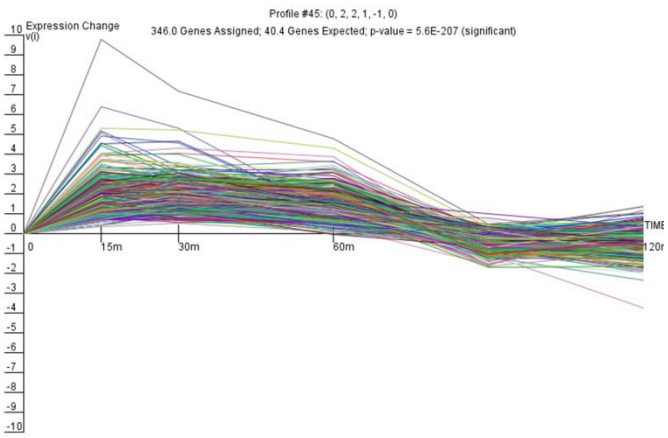
Figure 1: STEM Output with Six Significant Gene Profiles

RNA binding, RNA modification, hexose transport, and macromolecule metabolic process occur as a response to cold shock because yeast changes it gene expression in two stages: early and late (Schade et al. 2004). Schade et al. characterizes the late response to environmental stress with the activation of Msn2 and Msn4 proteins. However, the regulation of the early response remains unknown (2004). The 60-minute cold shock treatment from the microarray

resulted in gene ontology terms associated with potential early response genes and their effects on the cell. Schade et al.'s study (2004) explains that one of the first changes that may occur due to cold shock is the stabilization of RNA and DNA secondary structures, which is supported by genes clustered in Profile #45. For that reason, RNA binding and modification appeared as GO terms for the profile. The most fundamental processes of the cell, which includes transcription and translation of mRNA into proteins, must change to accommodate to these environmental conditions to allow the cell's continued survival.

Hexose transport and macromolecule metabolic processes, two other GO terms associated with Profile #45, may increase as yeast's response to cold shock. For the Dahlquist lab's growth experiments, cell growth was interrupted during its log or exponential growth phase. Utilization of larger macromolecules and enzymes that metabolize them, especially sugars, may increase significantly. Sugars (typically glucose) are yeast's preferable energy source (Berthels et al. 2004). With increased amounts of glucose catabolism, yeast grows and increases in number at faster rates. Due to cold shock, the colder temperatures not only decrease enzymatic activities, but they may also force yeast to reallocate more energies to survival than growth via the cell cycle. Thus, more ATP is needed to allow for proper enzymatic functions. Because cell growth is halted, the amount of sugars transported into the cell may decrease. However, profile #45 saw genes with initial up-regulation. Hexose transport was a GO term because genes encoding these increased within the 60 minutes, suggesting more sugar may be needed to create more energy to support and sustain the cell during colder conditions.

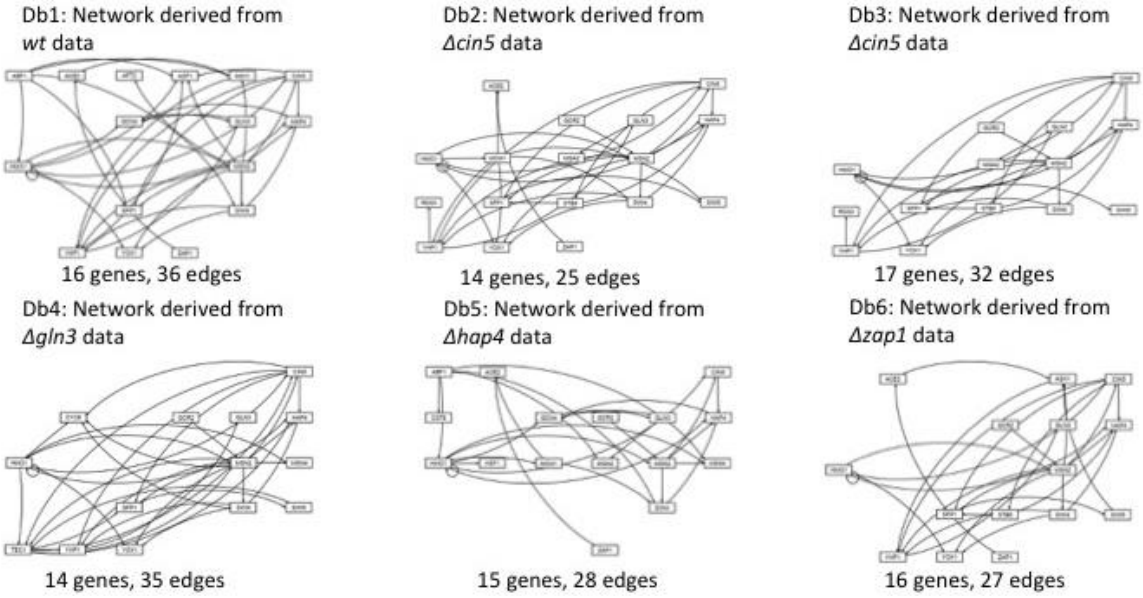
Figure 2: Profile #45 Had Genes with Initial Up-regulation and then Down-Regulation during Recovery Times



Construction of Gene Regulatory Networks and Input Sheets from the Microarray Data Using YEASTRACT

In visualizing the unweighted Db networks, we see common genes amongst all the networks (Fig. 3). This observation suggests that specific transcription factors are key in regulating the response to cold shock. Granted, six of those TFs were added due to the Dahlquist lab having deletion strain microarray data for the genes. However, different genes exist in each of the networks. To avoid dependence on one specific regulator, yeast cells divide the regulation of cellular processes. Biologically, this division of labor allows for the continuation of cellular processes regardless of the deletion of genes.

Figure 3: Visualization of the Six Db networks unweighted, meaning that all regulatory relationships are of the same value

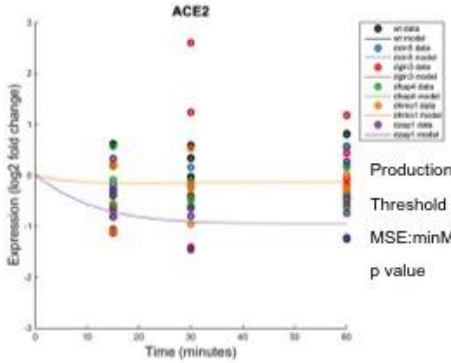


No Relationship between MSE:minMSE Values and P Values for Db5 Network

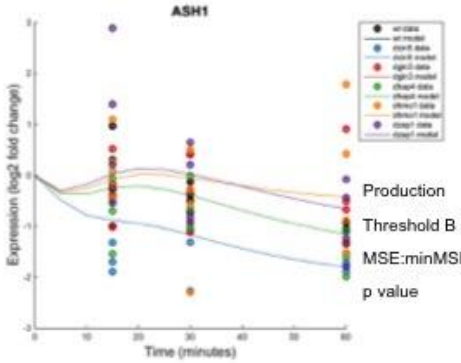
After calculating the minMSE values, the ratio between the MSE and minMSE was compared. The closer this ratio is to one, the better the modeled performed in modeling the dynamics of an individual gene. What we hoped to find was that the better ratios corresponded to the genes with significant Benjamini & Hochberg p values < 0.05. However, a relationship between the MSE:minMSE ratio and the p value was not identified.

Expression plots for the genes within Db5 are provided. In these plots, the dots at the 15, 30, and 60 time points signify the $\log_2$ expression values for each of the strains. The lines correspond to the dynamics the model computed to describe the changes in gene expression with simulated data. Although there are MSE:minMSE values for each gene's behavior for each strain, the ratios in Fig. 4 are from the gene's behavior in the *Δzap1* deletion strain. The p values for each gene in Fig. 4 also correspond to the p values from the *Δzap1* deletion strain.

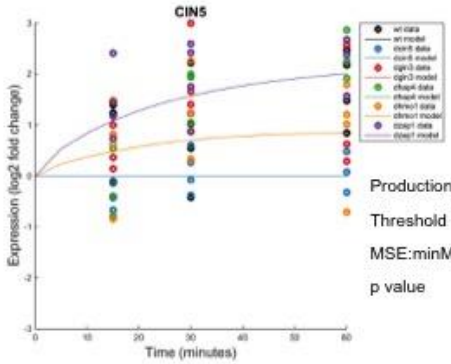
Figure 4: Example of Expression Plots of the Genes Within Db5 with Each Gene's Production and Threshold b Values in addition to the Δzap1 MSE and P Value Comparisons



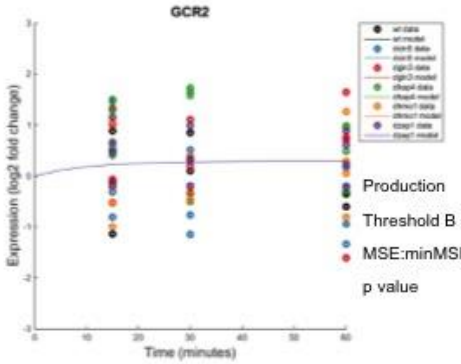
Production	0.1651
Threshold B	0.6179
MSE:minMSE	7.9482
p value	0.0457



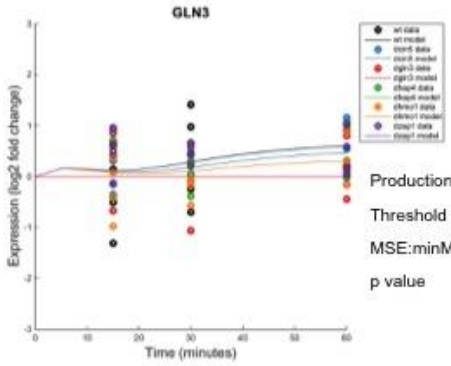
Production	2.0328
Threshold B	-1.7817
MSE:minMSE	6.6229
p value	0.4535



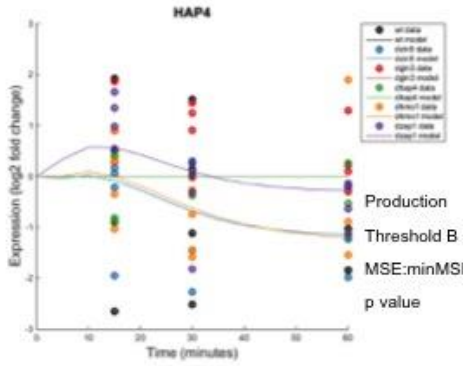
Production	0.958
Threshold B	2.3761
MSE:minMSE	0.7486
p value	0.026



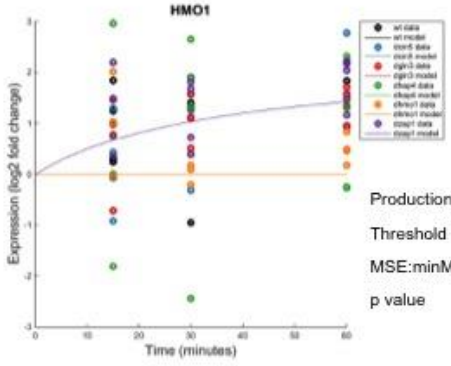
Production	0.2351
Threshold B	0
MSE:minMSE	0.5048
p value	0.3479



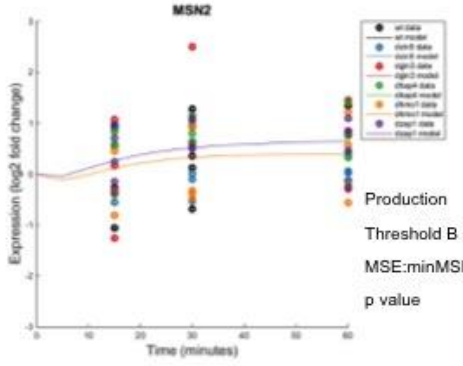
Production	0.2884
Threshold B	0.8198
MSE:minMSE	1.039
p value	0.1171



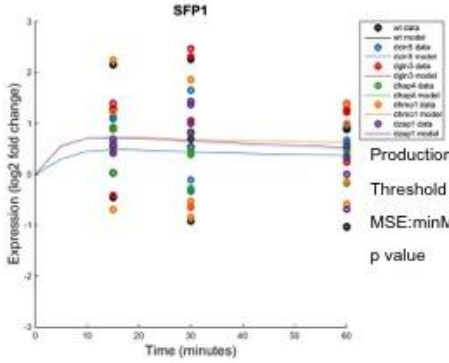
Production	1.6878
Threshold B	0.7298
MSE:minMSE	1.7218
p value	0.0157



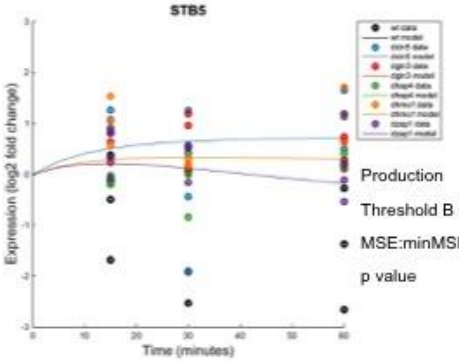
Production	0.2414
Threshold B	0.9792
MSE:minMSE	0.196
p value	0.0011



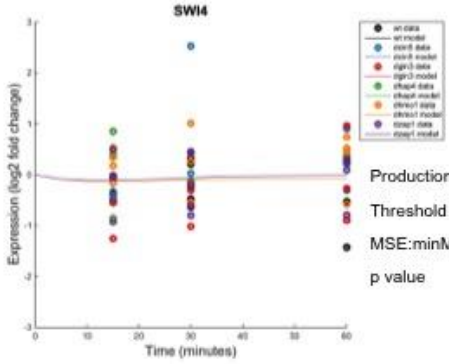
Production	0.8731
Threshold B	4.7432
MSE:minMSE	1.6503
p value	0.0836



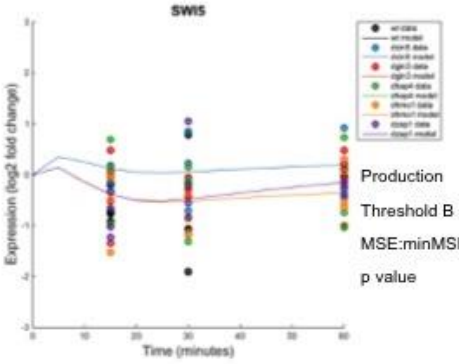
Production	1.2322
Threshold B	1.0618
MSE:minMSE	2.1974
p value	0.0021



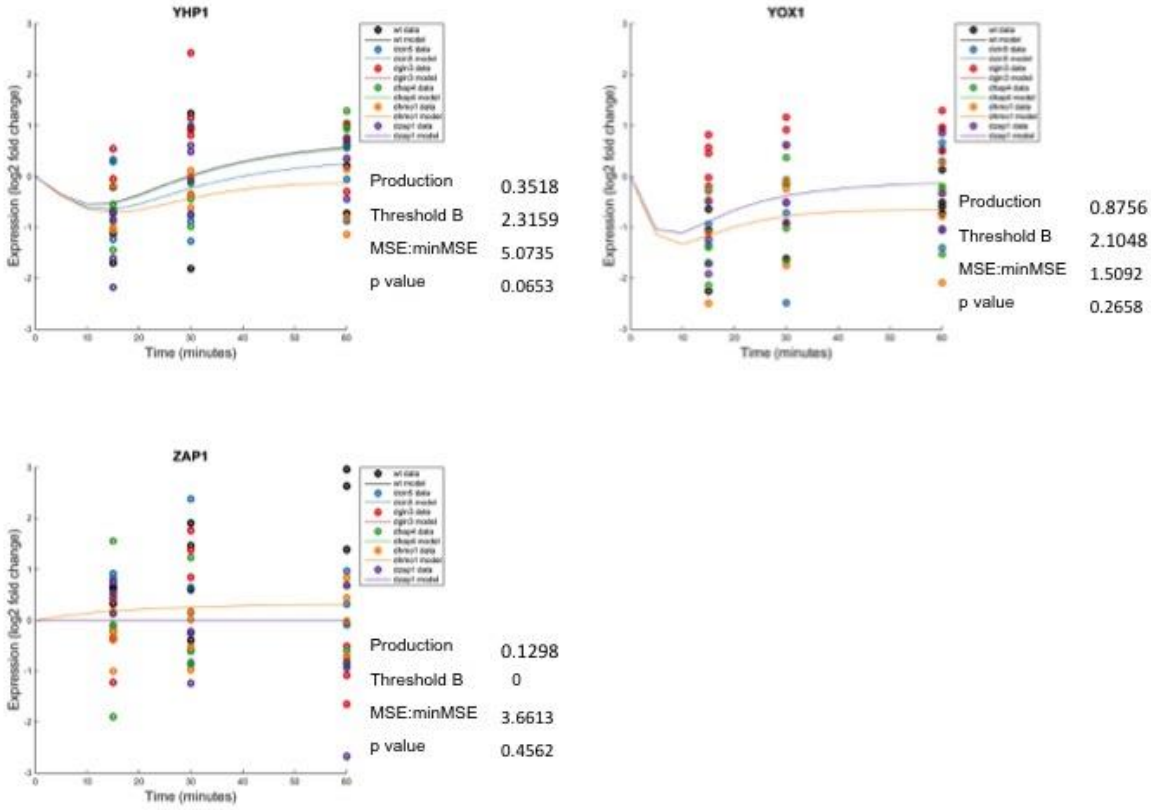
Production	0.3361
Threshold B	0.6494
MSE:minMSE	11.9057
p value	0.3296



Production	0.2436
Threshold B	0.3436
MSE:minMSE	1.8262
p value	0.7799



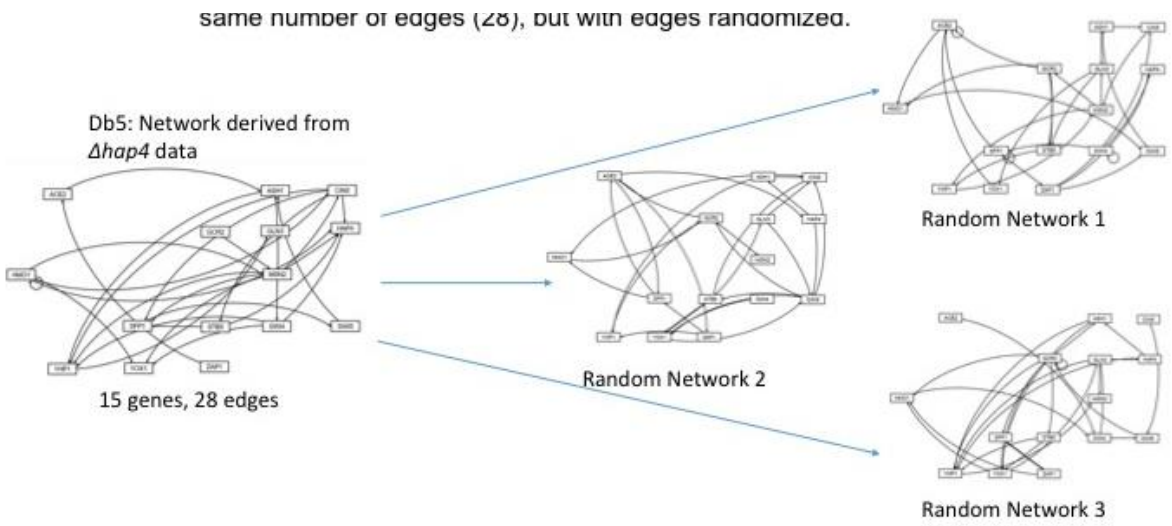
Production	1.258
Threshold B	-0.5855
MSE:minMSE	0.7573
p value	0.2975



Generation of Random Networks Related to Db5 Network

Derived from the same data as the Db5 network, the 31 random networks had the same number of nodes (15) and edges (28). However, the connections or edges between these nodes were randomized. An example of a few random networks and Db5 can be found in Fig. 5 to show that these connections are randomized.

Figure 5: Visualizations of Db5 and Three of the Random Networks Generated from the Same Data



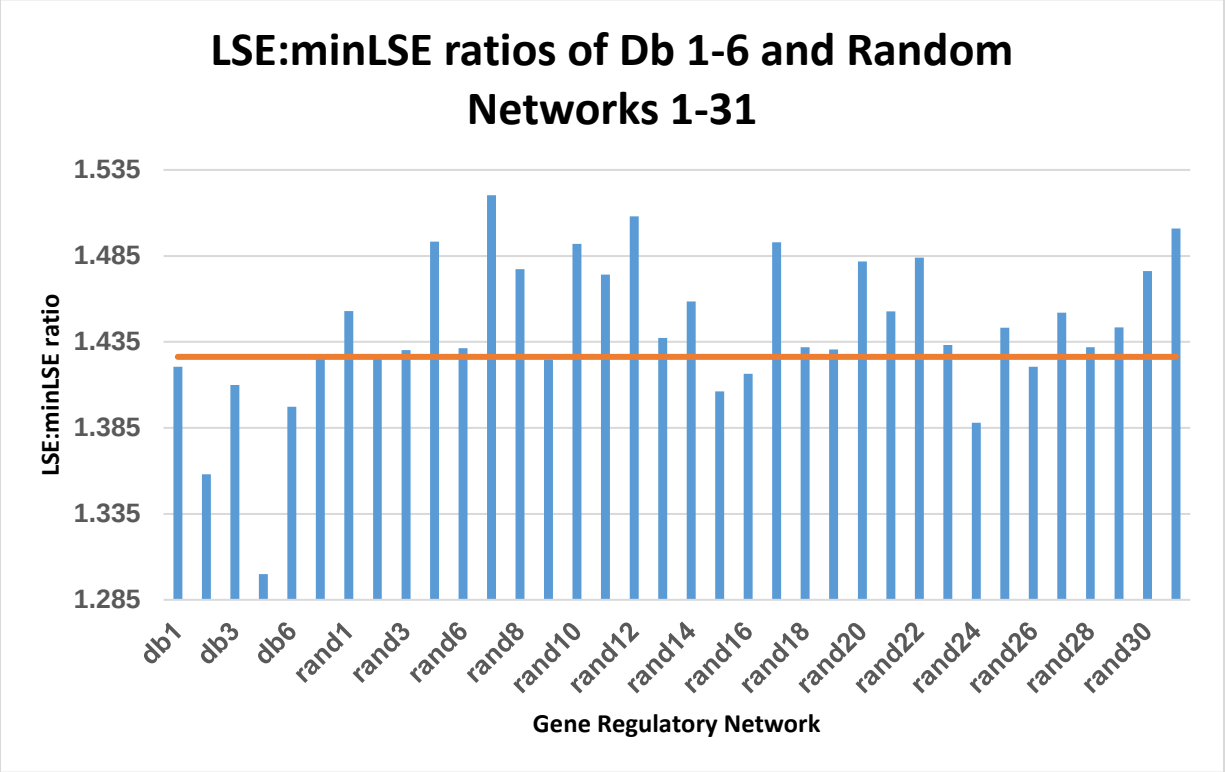
Comparison of Db1-6 and Random Network LSE:minLSE Values Showed that the Database-Derived Networks Performed Better

To analyze the model's output from Db networks 1-6 and some of the random networks, various procedures were taken. For this paper, the Db5 network was compared to the three best and three worst random networks. One of the first tools used to compare the Db5 network to the 31 random networks was the ratio of the observed LSE to the minimum LSE. These ratios were calculated by dividing the LSE by the minLSE that appear on the optimization diagnostics sheet in the network's output. The closer this ratio was to one, the better GRNmap modeled the dynamics of the overall GRN.

In Fig. 6, the LSE:minLSE ratios for each of the networks (Db1-6 and random 1 – 31) can be seen. The average LSE:minLSE ratio for the random networks was 1.455 while the databased-derived networks' average LSE:minLSE ratio was 1.3853. The horizontal line across the plot is the value of Db5's LSE:minLSE ratio because for further analysis, Db5 was used. Db5's LSE:minLSE ratio was 1.4263. This result of these averages suggests that the random networks perform worse than the biologically relavant networks. To compare Db5 to random networks, the three random networks with the best and worst LSE:minLSE ratios were identified.

The three random networks that performed better were networks 15, 16, and 24; those that performed worse were 7, 12, and 31. The three best random networks had LSE:minLSE ratios lower than that of Db5 while those that performed worse had ratios higher than Db5's.

Figure 6: LSE:minLSE ratios for the Six Db Networks and the 31 Random Networks showed that the Database-derived Networks Performed Better on Average

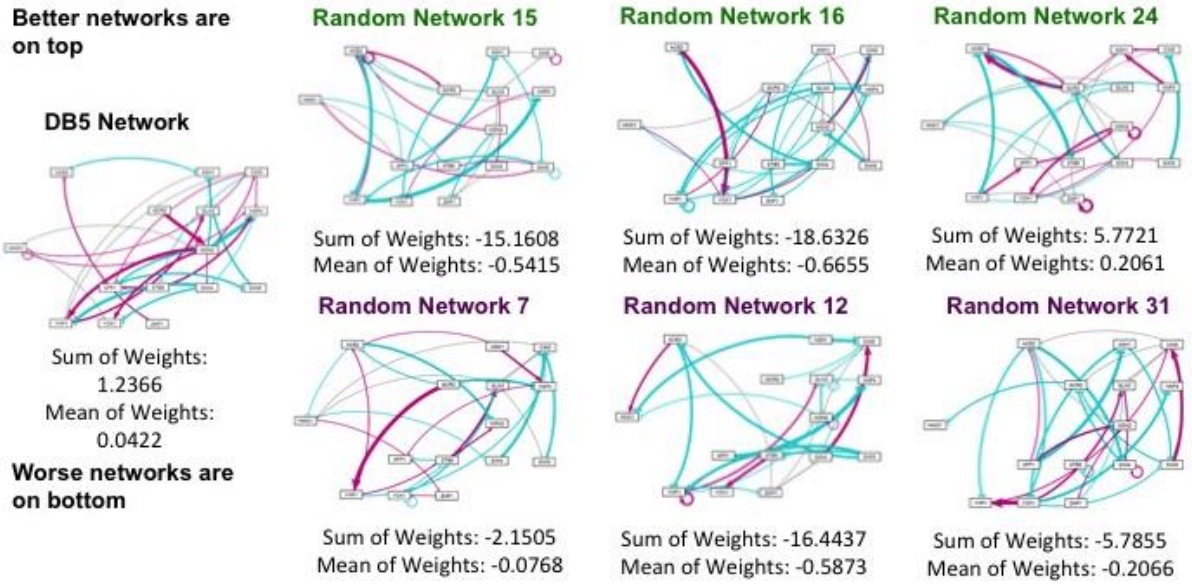


More Repressive Regulatory Relationships Occurred in the Random Networks than Db5

To compare the overall network, the sum of all the weights of each transcription factor on its target gene was computed. Using GRNsight for visualization, a SIF file was exported that isolated each regulatory relationship that provided the magnitude and direction. More repressive regulatory relationships were found in the random networks compared to Db5 (Fig. 7). For example, all but one random network had a positive sum of weights compared to Db5. To compute the sum of the weights, the magnitude and direction (+ or -) of each regulatory relationship in the network was added together. Further, the mean of the weights was also computed, showing the same pattern – overall repression in the random networks compared to

Db5. To calculate the mean of weights, the sum of the weights was first computed, and then the total number of values divided the sum. This observation suggests that the biologically relevant networks have more activation. As Alon 2007 mentioned, certain network motifs are seen in "real" networks compared to randomized ones comprised of the same data.

Figure 7: Comparison of the Sum of Weights and Mean of Weights for Db5 and the Three Best and Three Worst Random Networks

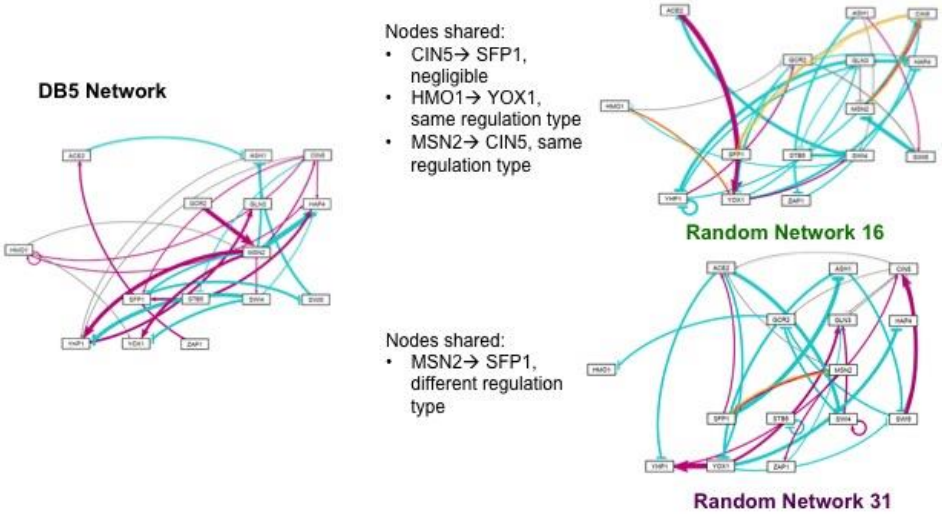


One of these motifs is a positive feed-forward loop, which can be seen in Db5. Network motifs combine to generate larger-scale networks, and are found primarily in real networks compared to random networks (Alon 2007). In most networks, the feed-forward loop consists of transcription factor A that controls transcription factor C and transcription factor B, which also regulates transcription factor C (Ahnert & Fink 2016). According to Milo et al., a possible function of feed-forward loops in sensory component networks may be to signal output if the input stimulus is continuous and allow for rapid deactivation when the stimulus ends (2002). The most common feed-forward loop is a coherent (indirect path results in the same effect as the direct path), three node circuit where all relationships are activating.

Networks that Performed Better than Db5 had More Shared Edges than the Random Networks that Performed Worse

To better understand why the networks performed better or worse as seen by their LSE:minLSE ratios, regulatory relationships were compared. The random networks with similar edges to Db5 performed better than the random networks that did not have similar edges, which performed worse. In the networks that performed better, random network 16 shared three edges with Db5 while random network 31, which performed worse, shared one edge (Fig. 8). While similar edges marked similarities between the random networks and Db5, the regulatory relationship (activation or repression) was the key indicator of performance.

Figure 8: GRNsight Visualization of Db5, Random Networks 16, and Random Network 31 Clearly Showed Similarities and Differences in Regulatory Relationships between Shared Edges



When analyzing the Db5 and random network 16, which performed better in the model, three edges were shared. One was the regulation of SFP1 by CIN5. However, CIN5's regulation of SFP1 was negligible when normalized in GRNsight. However, the next two connections were stronger with the same regulatory relationship (activation) between the Db5 network and random network 16. HMO1 activates YOX1's production. HMO1 maintains the genome by regulating transcription of RNAs via its interaction with DNA and the nucleolus (Saccharomyces Genome Database 2012). YOX1, as a homeobox transcriptional repressor, has a direct role in the

transcription of genes involved in the transition between M and G_1 phases (Saccharomyces Genome Database 2012). HMO1 activates transcription of YOX1 so as to negate too much transcription of unnecessary genes that promote cell growth. This relationship occurs to halt the continuation of cell division and focus on maintenance of the cell while in colder conditions.

The next regulatory relationship shared between random network 16 and the Db5 network was activation of CIN5 by MSN2. MSN2 is known to be part of the late cold shock response (Schade et al. 2004); however, it appears to have a role in the early response to environmental stress. MSN2 may be part of a general environmental stress mechanism (Petrenko et al. 2013; Martinez-Pastor et al. 1996); however, in Schade et al.'s study, MSN2 may have had the highest levels of prolonged gene expression, which suggested its role in the late response to cold shock. With its activation of CIN5, MSN2 recruits CIN5 to increase the expression of TUP1, a stress response gene, in the cell (Hanlon et al. 2011). Together, these transcription factors may be crucial for the cell's immediate response to cold shock.

For the network that performed worse than DB5, only one edge was shared between them; however, the regulatory relationship differed between the two networks. In Db5, MSN2 represses SFP1, but in random network 31, activation was seen between the regulator and its target gene. As previously mentioned, MSN2 is an environmental stress response transcription factor while SFP1 regulates the transcription of ribosomal proteins and biogenesis genes. In the biologically relevant or database-derived network, we see activation of SFP1 by MSN2. Because continued maintenance of proteins that will aid the cell's survival in the cold will be needed (Schade et al. 2004), ribosomal proteins and biogenesis genes will constantly be needed to assist with the generation of gene products. Further, in the random network, the repression of SFP1

would result in decrease in ribosome availability thus halting protein production, reducing the metabolic, enzymatic, and signaling pathway capacities of the cell.

Conclusion

In the course of my investigation, I have performed statistical analyses of the recurring data. I used the data to infer six related small GRNs from the YEASTRACT database. I ran GRNmap to estimate parameters for regulatory dynamics in this network. From the Db5 network, I generated more related GRNs for 30 random networks.

For validation, I performed various analyses using the output workbook from the GRNmap software. Using the observed LSE and comparing it to the minimum theoretical LSE, I found that the overall dynamics of the database-derived networks were lower. This result suggests that the database-derived networks have more biological relevance than the random networks comprised of the same number of nodes and edges as Db5. With direct comparison of Db5 and the three best (15, 16, and 24) and three worst (7, 12, and 31) random networks, I saw that activation was more prevalent in the database-derived network, using the sum of weights and mean of weights. The most common motif in biologically relevant GRNs was the feed-forward loop between three genes, with all activating regulatory relationships (Alon 2007). Finally, all the three best networks that performed better than Db5 shared more edges and regulatory relationships with Db5 than those random networks that performed worse. This conclusion suggests that the databased-derived GRNs explain at least part of the transcriptional response to cold shock in yeast and are biologically relevant.

In the future, there are many directions to take these results. For instance, comparisons between the Db4, the GRN with the lowest LSE:minLSE ratio of all the database-derived networks, and the random networks can also occur. It would be interesting to see if the same

three networks that performed better related to Db5 would share more edges and regulatory relationships with Db4 than those random networks that performed worse.

It would also be interesting to include all the genes from the database-derived networks into one large GRN. By constructing a network that consists of all the genes, it may provide insights on how a broader network in yeast cells would respond to cold shock conditions. After being input into GRNmap, this GRN's results could then be compared to the individual Db 1-6 networks as well as the three best and three worst random networks.

References

Aguilera, J., Randez-Gil, F., and Prieto, J.A. (2007). Cold response in Saccharomyces cerevisiae: new functions for old mechanisms. *FEMS Microbiology Review*, 31, 327-34.1. doi:10.1111/j.1574-6976.2007.00066.x

Ahnert, S.E. and Fink, T.M.A. (2016). Form and function in gene regulatory networks: the structure of network motifs determines fundamental properties of their dynamical steady state. *Journal of the Royal Society Interface*, 13, 1-8. doi:10.1098/rsif.2016.0179

Alon, U. (2007). "The feed-forward loop network motif." *An introduction to systems biology: Design principles of biological circuits*. Boca Raton, FL: Chapman & Hall/CRC, pp 41 – 69. Print.

Berthels, N.J., Otero, R.R.C, Bauer, F.F., Thevelein, J.M., and Pretorious, I.S. (2004). Discrepency in glucose and fructose utilisation during fermentation by Saccharomyces cerevisiae wine yeast strains. *FEMS Yeast Research*, 4, 683-689. doi: 10.1016/j.femsyr.2004.02.005

Bumgarner, R. (2013). DNAmicroarrays: types, applications and their future. *Curr Protoc Mol Biol.* 22. doi: 10.1002/0471142727.mb2201s101

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics*, 8, 93-103. doi: 10.1038/nrg1990

Dahlquist, K.D., Fitzpatrick, B.G., Camacho, E.T., Entzminger, S.D., and Wanner, N.C. (2015). Parameter estimation for gene regulatory networks from microarray data: cold shock response in Saccharomyces cerevisiae. *Bulletin of Mathematical Biology*, 77, 1457-1492. doi: 10.1007/s11538-015-0092-6

Desai, N.S., Agarwall, A.A., and Uplap, S.S. (2010). Hsp: evolved and conserved proteins,

structure and sequence studies. *International Journal of Bioinformatics Research*, 2, 67-87.

Ernst, J. & Bar-Joseph, Z. (2006). STEM: a tool for the analysis of short time series gene expression data. *BMC Bioinformations*. 7:191-202. doi: 10.1186/1471-2105-7-191

Feder, M.E., and Hofmann, G.E. (1999). Heat-shock proteins, molecular chaperones, and the stress response: evolutionary and ecological physiology. *Annual Review of Physiology*, 61, 243-282. doi:10.1146/annurev.physiol.61.1.243

Geistlinger, L., Csaba, G., Dirmeier, S., Küffner, R., and Zimmer, R. (2013). A comprehensive gene regulatory network for the diauxic shift in Saccharomyces cerevisiae. *Nucleic acids research*, 41, 8452-8463. doi:10.1093/nar/gkt631

Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M., Louis, E.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S.G. (1996). Life with 6000 genes. *Science*, 274, 546-567.

Hanlon, S.E., Rizzo, J.M., Tatomer, D.C., Lieb, J.D., and Buck, M.J. (2011). The stress response factors Yap6, Cin5, Phd1, and Skn7 direct targeting of the conserved co-repressor Typ1-Ssn6 in S. cerevisiae. *PLoS One*, 6, e19060. doi: 10.1371/journal.pone.0019060

Ingalls, B.P. (2013). "Introduction." *Mathematical modeling in systems biology: An introduction.* Cambridge, MA: MIT, pp 1-19. Print.

Jakob, U. Gaestel, M., Engel, K., and Buchner, J. Small heat shock proteins are molecular chaperones. *The Journal of Biological Chemistry*, 268, 1517-1520.

Kim, S., Kim, J., and Cho, K. (2007). Inferring gene regulatory networks from temporal expression profiles under time-delay and noise. *Computational Biology and Chemistry*,

31, 239-245. doi: 10.1016/j.compbiolchem.2007.03.013

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.T., Thompson, C.M., Simon, I., Zeitlinger, J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J., Volkert, T.L., Fraenkel, E., Gifford, D.K., and Young, R.A. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298, 799-804. doi:10.1126/science.1075090

Martinez-Pastor, M.T., Marchler, G., Schüller, C., Marchler-Bauer, A., Ruis, H., and Estruch, F. (1996). The Saccharomyces cerevisiae zinc finger proteins Msn2p and Msn4p are required for transcriptional induction through the stress response element (STRE). *EMBO J*, 15, 2227-2235.

McKenna, N.J., and O'Malley, B.W. (2002). Combinatorial control of gene expression by nuclear receptors and coregulators. *Cell*, 108, 465-474. doi:10.1016/S0092-8674(02)00641-4

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298, 824-827. doi:10.1126/science.298.5594.824

involves a histone deacetylase complex. *Nature*, 393, 386-389. doi: 10.1038/30764

Neymotin, B., Athanasiadou, R., and Gresham, D. (2014). Determination of in vivo RNA kinetics using RATE-seq. *RNA*, 10, 1645-1652. doi:10.1261/rna.045104.114

Petrenko, N., Chereji, R.V., McClean, M.N., Morozov, and Broach, J.R. (2013). Noise and interlocking signaling pathways promot distinct transcription factor dynamics in response to different stresses. *Mol. Biol. Cell*, 24, 2045-2057. doi: 10.109/mbc.E12-12-0870

Saccharomyces Genome Database. 2012. Saccharomyces cerevisiae genome database: the

genomics resource of budding yeast.

Schade, B., Jansen, G., Whiteway, M., Entian, K.D., and Thomas, D.T. (2004). Cold adaptation in budding yeast. *Molecular Biology of the Cell*, 15, 5492-5502. doi: 10.1091/mbc.E04-03-1067

Smolen, P., Baxter, D.A., and Byrne, J.H. (2000). Mathematical modeling of gene networks. *Neuron*, 26, 567-580. doi:10.1016/S0896-6273(00)81194-0

ter Schure, E.G., Silljé, H.H.W., Verkleij, A.J., Boonstra, J., and Verrips, C.T. (1995). The concentration of ammonia regulates nitrogen metabolism in Saccharomyces cerevisiae. *Journal of Bacteriology*. 177, 6672-6675.

Vu, T.L. and Vohradsky, J. (2007). Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of Saccharomyces cerevisiae. *Nucleic Acids Research*, 35, 279-287. doi:10.1093/nar/gkl1001