**Sensitivity Assessments of the GRNmap Modelling Software Reveals that Certain Manipulations of the Model Input Impacts the Fit of the Model**

**Alice Finton**

**May 8, 2020**

**Abstract:**

A gene regulatory network (GRN) is a set of transcription factors which regulate the level of expression of genes encoding other transcription factors. The dynamics of a GRN show how gene expression in the network changes over time. GRNmap (Gene Regulatory Network modeling and parameter estimation) is a MATLAB application that uses ordinary differential equations to estimate GRN parameters including gene expression thresholds, production rates, and regulatory weights from DNA microarray data. DNA microarray data from wild type *Saccharomyces cerevisiae* and transcription factor deletion strains (*∆cin5*, *∆gln3*, *∆hap4*, *∆hmo1*, *∆zap1*) in which the strains were subjected to 13oC cold-shock for 15, 30, and 60 minutes was used as input to the model. A Gene Ontology enrichment analysis of the microarray data using the clusters generated through significant STEM profiles and *k-*means clustering showed consistent results with common GO terms associated with down and upregulation during cold shock or recovery timepoints. A candidate cold-response GRN with 15 nodes (transcription factors) and 28 edges (regulatory relationships) was derived from the YEASTRACT database. Through this study, a series of experiments were run to determine the sensitivity of the model parameters to manipulations of the input, including structural changes to the candidate GRN, variable inclusion of strain expression data, changes to the optimization parameters, and fixing or estimating production rates. Through comparing the fit of the model between these experiments, the sensitivity of the model parameters to these changes was assessed. While changes to the optimization parameters did not impact the fit of the model, structural manipulations of the GRN, including edge and node deletions, variable inclusion of strain expression data, and using fixed values for the production rates did have an impact on the model fit.

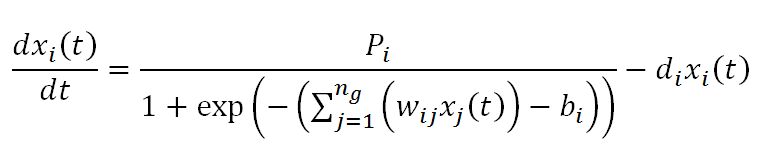
**Introduction**

Microorganisms must be able to adapt to environmental fluctuations in order to survive in natural environments. Environmental stressors, such as temperature, nutrient availability, and acidity, impact the internal homeostasis of the cells, requiring a cellular response to respond to the changing environment (Schade et al*.,* 2004). In yeast cells, internal responses are necessary after sudden environmental stress in order to maintain growth under the specific circumstances (Gasch et al.,2000). The environmental stress response (ESR) is initiated when the cells react to changes in the environment by altering the expression of hundreds to thousands of genes in the yeast genome in response to any type of environmental stress exposed to the cells. These changes to the expression of certain genes allows for the cells to return to a state of homeostasis, allowing for continued cell growth and division (Gasch et al., 2000).

The optimum growth temperature of budding yeast, *Saccharomyces cerevisiae,* is 30°C, with cells impaired for growth at other temperatures (Salvadó et al., 2011). Drastic temperature fluctuations affect the cellular functions and characteristics of yeast, impacting the ability for the cells to divide and grow. When grown in environments at other temperatures, *S. cerevisiae* responds with a change to the transcriptional control of various sets of proteins (Gasch et al., 2002). For example, the response of *S. cerevisiae* to heat shock, where cells are grown at higher than optimal temperatures, has been widely studied. In response, yeast induce a set of heat shock proteins (Mager, 1995). However, the response of yeast to cold shock is less widely studied, and the effects on gene expression is not completely understood. Under cold shock conditions, yeast cells experience lower than optimal temperatures that slow, but still allow cell growth and division (Tai et al.,2007). The cold shock response can be divided into the early cold response, which constitutes the first twelve hours of growth at low temperature, and the late cold response, which is the response after twelve hours of growth (Schade et al., 2004). The late cold response has been shown to be regulated by the transcription factors Msn2 and Msn4, while the transcription factors that regulate the early cold shock response remain unknown (Schade et al., 2004). Under cold conditions, decreases in membrane fluidity and membrane-bound enzymes have been shown to cause issues in membrane diffusion (Al-Fageeh et al.,2005). In addition, at low temperatures, protein synthesis has been shown to decrease due to changes in internal cellular processes (Al-Fageeh et al.,2005). When returned to the optimal growth temperature, yeast cells respond with downregulation of ESR genes, indicating cold shock response recovery (Gasch, 2002).

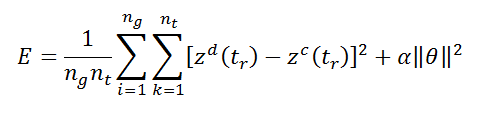
*Saccharomyces cerevisiae* is an ideal model organism for use in systems biology research. Due to the relatively small genome of about 6,000 genes and the availability of molecular genetic tools and datasets, research studies are more easily conducted. One such tool for generating large datasets is DNA microarray technology. The Dahlquist lab has used this technology to measure the global transcriptional response of a *S. cerevisiae* wild type strain and five transcription factor deletion strains to cold shock and subsequent recovery. DNA microarray technology allows for the determination of the induction or repression of genes within an entire genome (Schena, 1996). 6,189 genes from the yeast genome were included on the DNA microarrays used in this study, allowing for the transcriptional response in each of the genes to be determined. From the data, genes with similar expression profiles can be clustered. The Gene Ontology (GO) is a controlled vocabulary used to annotate genes and proteins in three categories, biological process, cellular component, and molecular function (Gene Ontology Consortium, 2015). GO can be used to assign function categories to genes in clusters, but determining how the genes are regulated requires understanding gene regulatory networks.

A gene regulatory network (GRN) is a set of transcription factors which regulate the level of expression of genes encoding other transcription factors. The dynamics of a GRN show how gene expression in the network changes over time. GRNmap (Gene Regulatory Network modeling and parameter estimation) is a MATLAB application that models the change in expression for each gene in the GRN as the production of mRNA minus its degradation using ordinary differential equations with a sigmoidal production function (Equation 1) where *Pi* is the mRNA production rate of the gene, *di* is the mRNA degradation rate of the gene, *b* is the expression threshold, and *w* is the regulatory weight of the gene (Dahlquist et al., 2015)



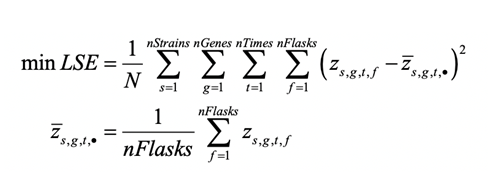
**Equation 1:** Differential equation used in GRNmap which models expression as production minus degradation with a sigmoidal production function.

Given published mRNA degradation rates from Neymotin et al. 2014 and the cold shock microarray data from the Dahlquist lab (t15. t30, t60 for wild type, *Δcin5, Δgln3, Δhap4, Δhmo1,* and *Δzap1* strains), GRNmap then estimates the production rates and expression thresholds for each gene, and regulatory weights for each edge, which denote the direction (activation or repression) and strength of the regulatory relationships. The model uses a penalized least squares approach to estimation (Equation 2). The least squares error assesses the difference between the experimental and simulated model values produced when the differential equation for each gene is solved with the estimated model parameters.



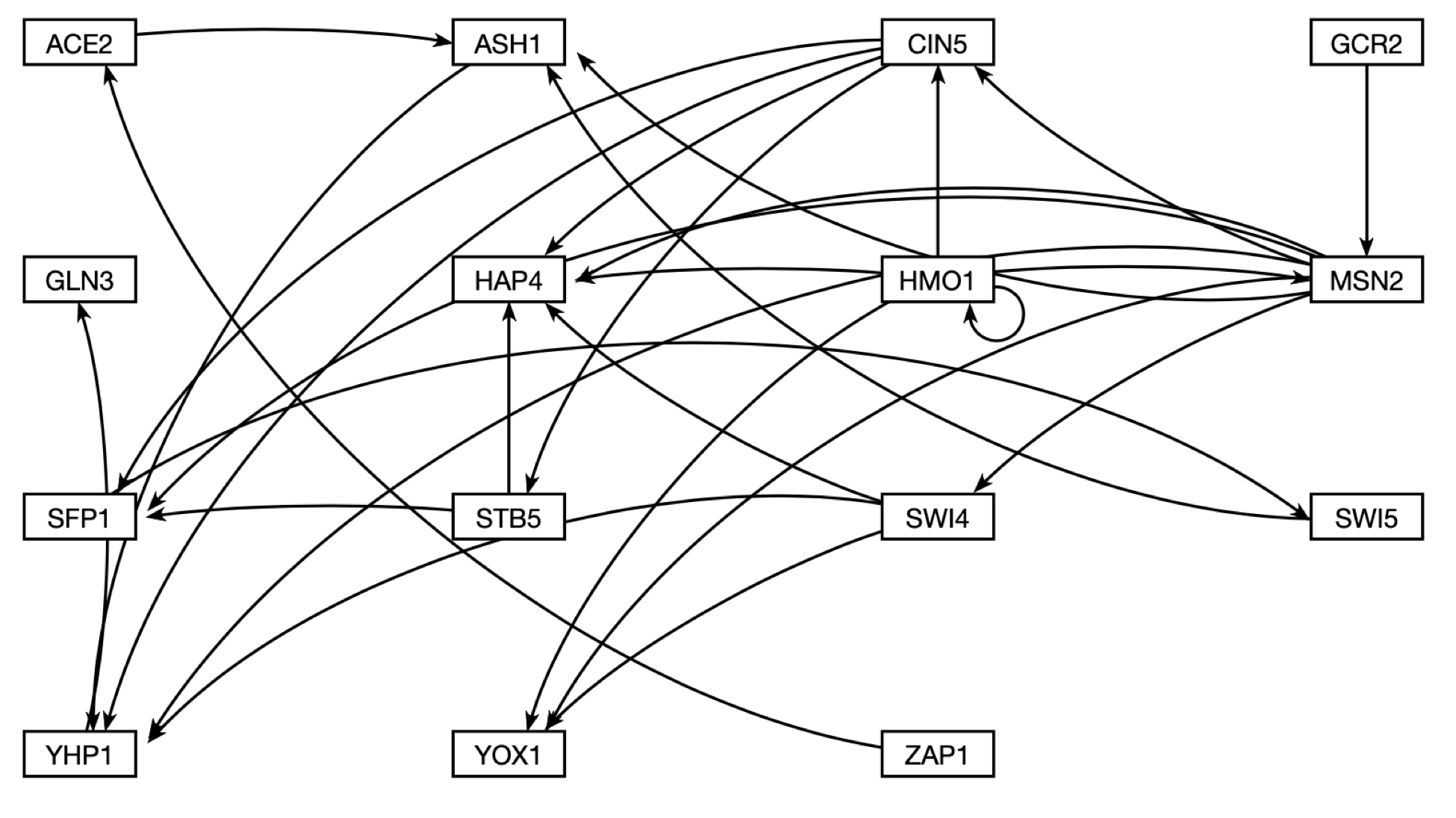
**Equation 2:** Penalized least squares error (LSE) assesses the difference between the experimental and simulated model values.

The fit of the model parameters for a particular estimation run can be assessed and different model runs can be compared using the ratio of the least squares error (LSE) to the minimum theoretical least squares error (minLSE) (LSE:minLSE).



**Equation 3:** Minimum theoretical least squares error (minLSE)

The candidate gene regulatory network (GRN) used in this study was “db5”, which has 15 nodes, representing regulatory transcription factors, and 28 edges, representing the regulatory relationships between them (Figure 1). The GRN was generated by inputting the genes that showed a significant change in expression at any time point (Benjamini and Hochberg correct p value < 0.05) from the *cin5* strain microarray data into the YEASTRACT (Yeast Search for Transcriptional Regulators and Consensus Tracking) database, which generates a list of transcription factors that potentially regulate genes in the data, as well as an adjacency matrix denoting their regulatory relationships (Monteiro et al., 2020).



**Figure 1:** Intact db5 gene regulatory network graph visualized using GRNsight. The GRN has 15 nodes (transcription factors) and 28 edges (regulatory relationships).

In this study, a series of experiments was performed to determine the sensitivity of the model parameters to manipulations of the input files, including structural changes to the candidate GRN, variable inclusion of strain expression data, changes to the optimization parameters, and fixing or estimating production rates and threshold b parameters. By comparing the fit of the model parameters between these experiments, the sensitivity of the model to these changes were assessed.

**1. Gene Ontology Enrichment Analysis of Microarray Data Clusters Generated using STEM and k-means Clustering**

**DNA Microarray Data**

Using DNA microarrays, expression data for 6189 genes from the *Saccharomyces cerevisiae* wild type strain and five transcription factor deletion strains (*Δcin5, Δgln3, Δhap4, Δhmo1,* and *Δzap1*) was collected in the Dahlquist lab (NCBI GEO GSE83656). Eachstrain was subjected to cold shock at 13oC for one hour and subsequent recovery at 30oC for another hour. Samples were taken before cold shock, after 15, 30, and 60 minutes of cold shock, and after 60 minutes of cold shock followed by and 30 and 60 minutes of recovery. A within-strain ANOVA was performed on these t15, t30, t60, t90 and t120 samples to determine which genes had a significant change in gene expression at any time point in the experiment. P values were adjusted for the multiple testing problem using the Benjamini and Hochberg correction. Genes with an adjusted p value < 0.05 were then included in the analysis below.

*STEM Profiles*

The expression patterns for the genes that had a significant change in expression at any timepoint (Benjamini and Hochberg corrected p value < 0.05) for five strains (excluding the *Δhmo1* data) were clustered using the Short Time-series Expression Miner (STEM) software individually for each strain (Ernst and Bar-Joseph, 2006). STEM groups genes by their expression patterns into pre-determined profiles based on the number of timepoints and determines which profiles show a statistically significant enrichment of genes that show that expression pattern (Ernst and Bar-Joseph, 2006).

The resulting STEM profiles for each strain were compared to determine which profiles the strains had in common. An analysis of the STEM profiles revealed that three profiles— profiles 45, 9, and 22— were common to the clustering results of all of the strains. Other profiles were found in several, but not all, of the strains, such as profiles 48, 2, 7, 40, 0, and 28. Finally, other profiles were unique to a single strain, such as profiles 31 and 38 (Figure 2).

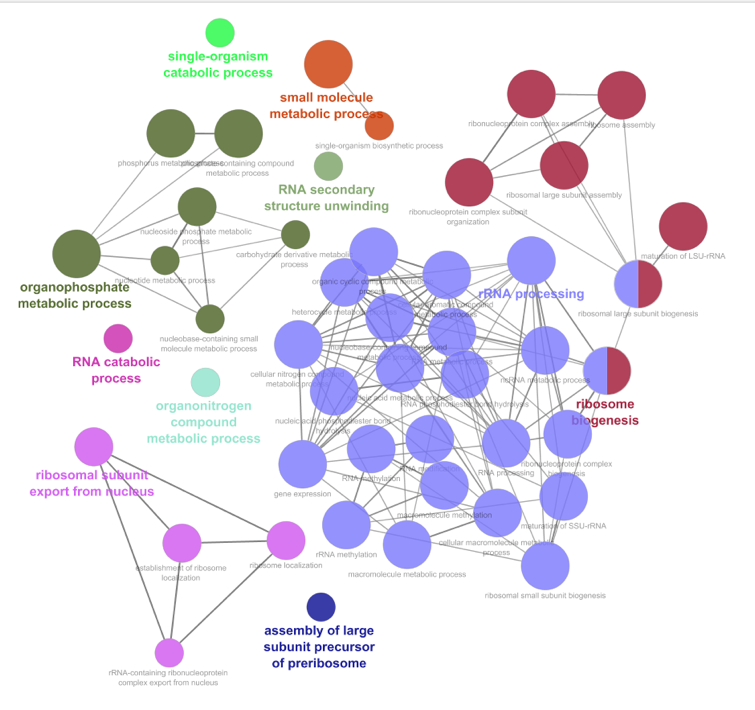
Profile 45 is characterized by an initial increase in gene expression during cold shock followed by a decrease during the recovery period (Figure 2). Profile 9 shows down regulation during cold-shock with a return to the pre-cold shock expression levels in the recovery period. Profile 22 shows no change in expression throughout the cold shock time points followed by upregulation of gene expression during the recovery period (Figure 2). These three profiles are likely to contain genes that are important for the cell’s response to cold shock because they show different levels of expression during the cold shock time points versus the recovery time points.

|  |
| --- |
| **Figure 2:** The significant STEM profiles, with trend lines, indicating the pattern of expression common to the genes in the cluster. Profile identifiers are listed across the top row and strains are listed to the left. |

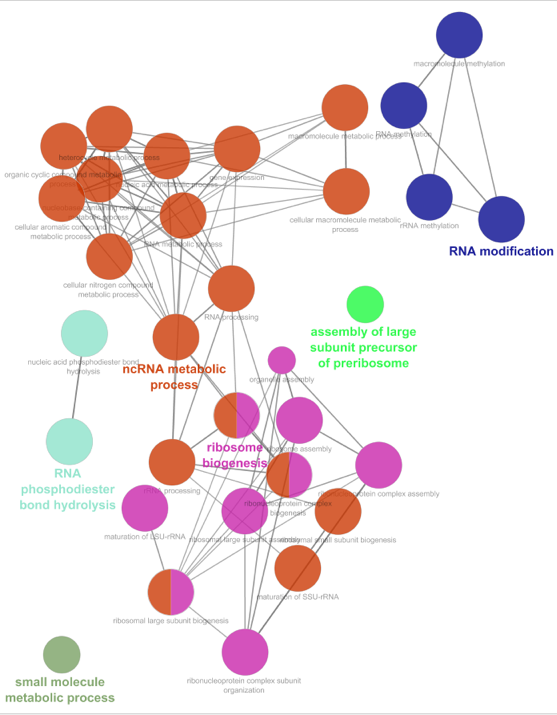
*ClueGO Gene Ontology Analysis of STEM Profiles*

A Gene Ontology (GO) term enrichment analysis for each of the STEM clusters was conducted using ClueGO, a Cytoscape plugin that creates Gene Ontology term networks. ClueGO also facilitates comparisons between two lists of GO terms (Maere et al., 2005). The plugin creates a visualization for groups of GO terms with networks and charts, using kappa statistics to generate links between GO terms. The gene lists and GO terms for STEM profiles 45, 9, and 22 for each strain were input into ClueGO and the resulting GO term networks were used to determine the overrepresented functional categories. The full set of methods used to input gene lists or GO terms into ClueGO and examples of properly formatted input files can be found at <https://openwetware.org/wiki/Alice_Finton_Online_Lab_Notebook#How_to_run_ClueGO>.

For profile 45, the overrepresented GO categories included those involved in ribosome biogenesis and RNA processing (Figure 3). Previous studies have shown that ribosome biogenesis and rRNA-processing genes are upregulated during cold shock, indicating their role in the cell’s response to the low temperature (Al-Fageeh et al., 2006). Upregulation of genes associated with ribosome biogenesis indicates that it is an important factor in the adaptation of the yeast cells to the cold environment. Albert et al. (2019) found that the induction of these genes is necessary for the maintenance of cell proliferation and growth. Therefore, the induction of these genes indicates that in order to maintain this homeostasis, the yeast cells had an increase in expression of ribosome-related genes. In addition, Sahara et al. (2002) found that genes associated with RNA processing, especially RNA polymerase I, which is involved in the production of rRNAs, are upregulated during the early cold shock response. In the present study, it was found that genes associated with ribosome biogenesis and RNA processing were upregulated initially then downregulated during the recovery period, which is consistent with the previous findings.



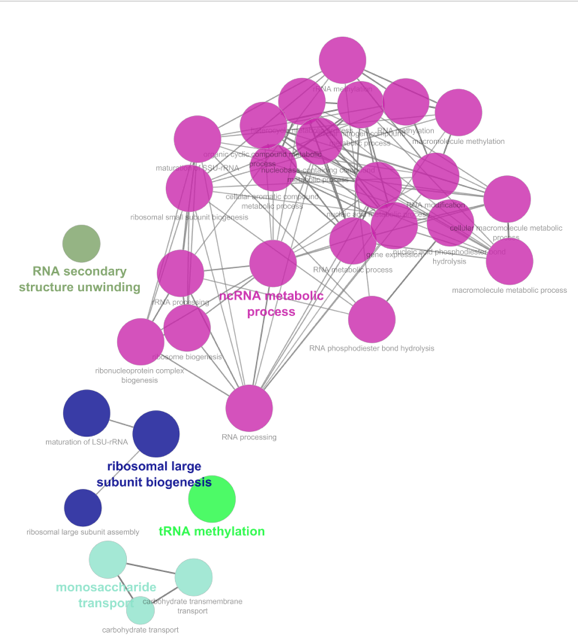
Wild-Type



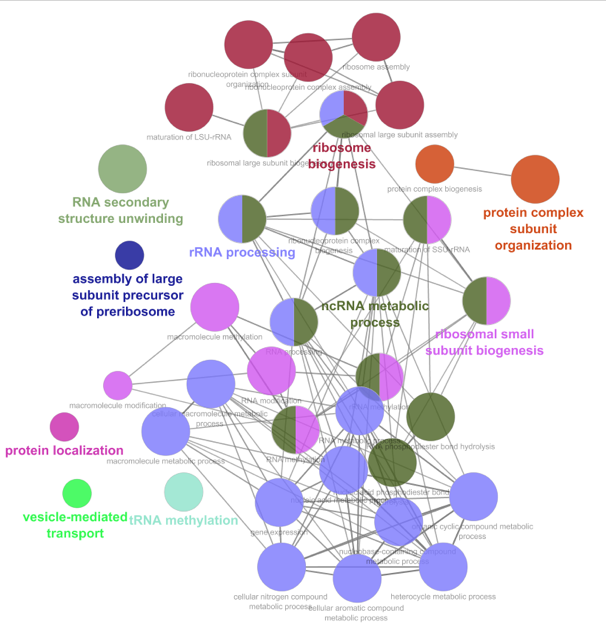
*∆cin5*



*∆gln3*



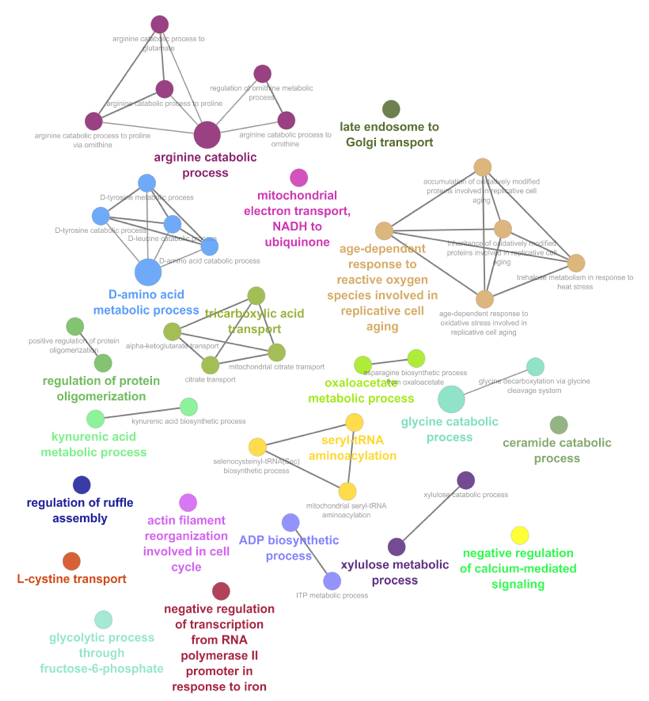
*∆hap4*



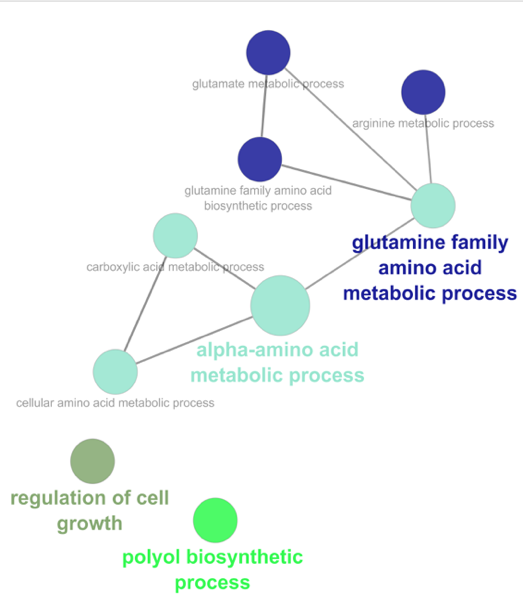
*∆zap1*

**Figure 3:** GO networks showing the overrepresented GO categories in profile 45 for each of the strains. The color of the nodes in the networks correspond to related terms found within each strain. The color of nodes between strain networks does not necessarily mean that the terms are related.

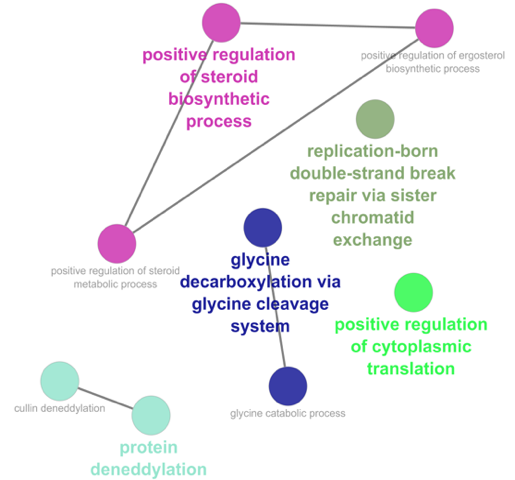
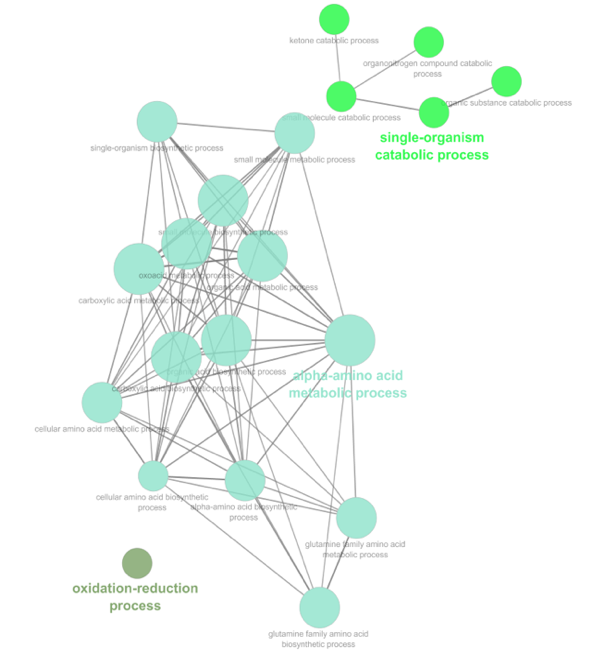
Profile 9 contains genes that were downregulated during the cold shock time points and increased back to pre-shock levels during the recovery period. For this profile, the overrepresented GO categories enriched similarly for each of the strains included catabolic and metabolic processes, such as amino acid and carboxylic acid metabolism, and cell cycle regulation (Figure 4). Schade et al. (2004) found that genes associated with lipid and amino acid metabolism are upregulated during the early cold shock response, which is inconsistent with our findings.



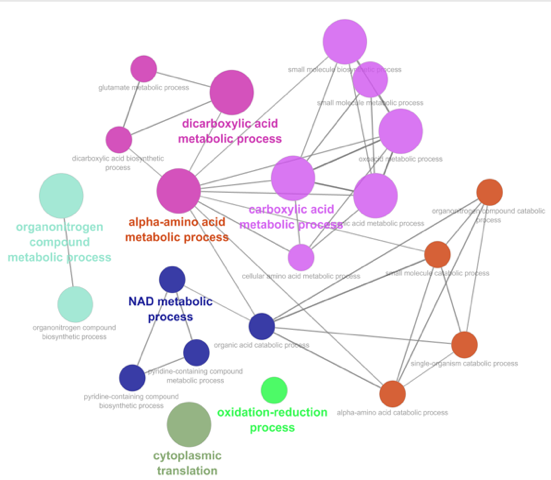
Wild-Type



*∆cin5*



*∆hap4*

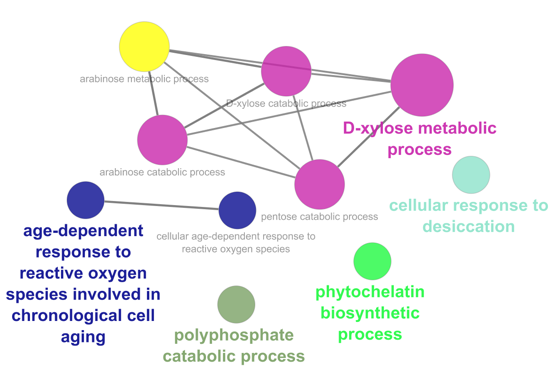
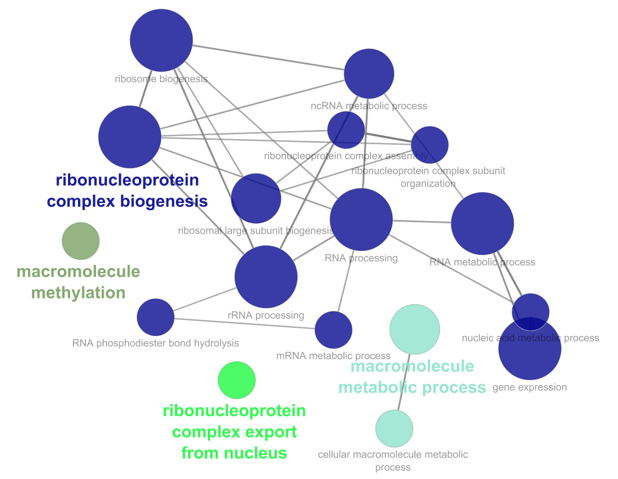
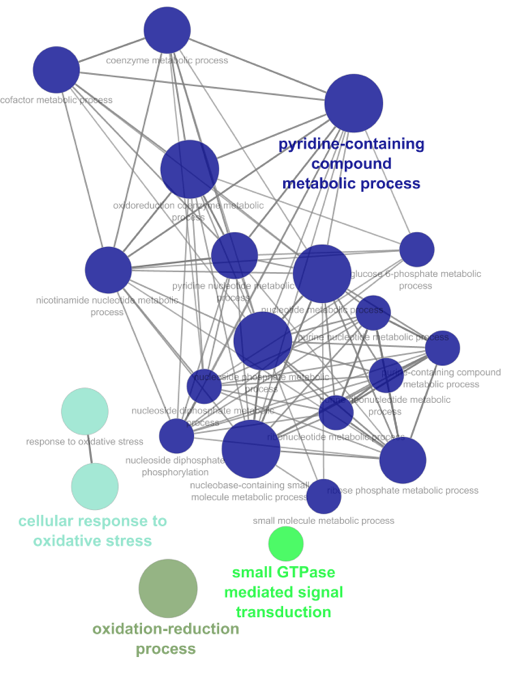
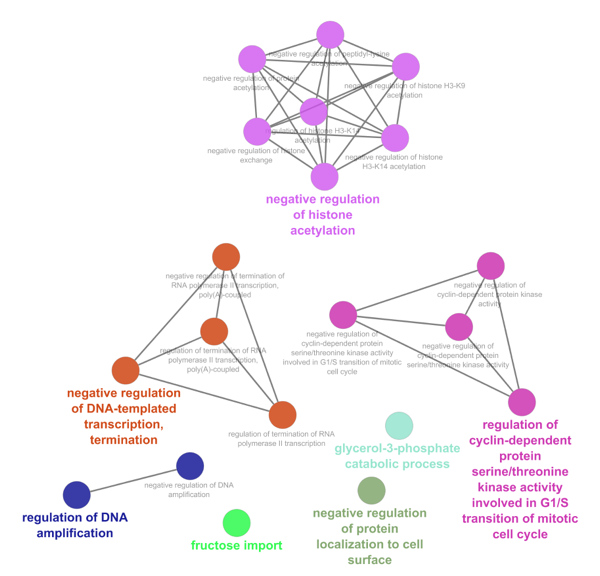
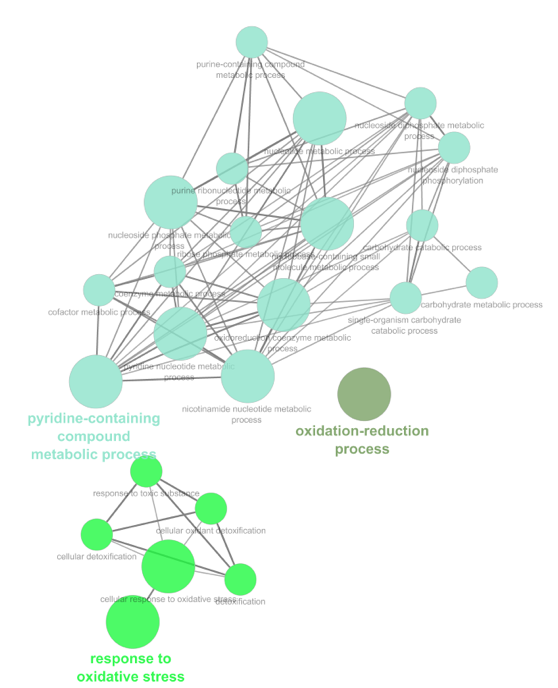


*∆zap1*

*∆gln3*

**Figure 4**: GO networks showing the overrepresented GO categories in profile 9 for each of the strains. The color of the nodes in the networks correspond to related terms found within each strain. The color of nodes between strain networks does not necessarily mean that the terms are related.

Profile 22 is defined by an upregulation of genes during the recovery period. Within this profile, the GO categories enriched in all strains included those involved in cell aging and stress responses, such as desiccation and oxidative stress (Figure 5). Sahara et al. (2002) found that genes associated with cell rescue, defense, and virulence were downregulated throughout the early and late cold shock responses.



Wild-Type

*∆cin5*

*∆gln3*

*∆hap4*

*∆zap1*

**Figure 5:** GO networks showing the overrepresented GO categories in profile 22 for each of the strains. The color of the nodes in the networks correspond to related terms found within each strain. The color of nodes between strain networks does not necessarily mean that the terms are related.

k-*Means Clustering of the Microarray Data*

In addition to clustering the microarray data through STEM, we performed *k-*means clustering in MATLAB. The STEM clustering method takes data time points into account in the clusters, whereas *k-*means clustering simply clusters the complete dataset, without considering time points (Ernst et al., 2005). The *k-*means clustering method partitions the data into a predetermined number of clusters (*k*) based on similarity (Wagstaff et al., 2001). Thus, we implemented *k-*means clustering to cluster the entire microarray dataset independently from the time points to determine whether we could uncover other expression patterns.

The implementation of *k*-means clustering in MATLAB could not handle missing expression values that occurred for some replicates of the microarray data. Thus, before carrying out the clustering, the log fold change values for each gene at each of the timepoints for each strain were averaged and the remaining missing values were set to zero. Genes that showed no significant change in gene expression (Benjamini and Hochberg corrected ANOVA p value > 0.05 across all of the strain data were filtered out, leaving 2441 genes in the analysis. These genes were then grouped into ten clusters using *k*-means clustering (Figure 6). The genes were then reordered manually by similarity (left to right) and the experiments were ordered by strain and time point (top to bottom). A heatmap was generated with green indicating downregulation and red indicating upregulation (Figure 6). (The code and files for clustering and heatmap generation can be found in the appendix.)

|  |
| --- |
| **Figure 6:** Heatmap of microarray clusters with associated cluster number designated below the heatmap. Strain data ordered by strain and time point is listed on the y-axis and the genes which showed significant changes in expression are on the x-axis. Green indicates downregulation and red indicates upregulation. The strains and timepoints on the right hand side are listed top to bottom as follows: ∆*hap4* t120, ∆*zap1* t120*, ∆swi4* t120*, ∆gln3* t120*, ∆cin5* t120*, wt* t120*, ∆hmo1* t120*, ∆hmo1* t90*, ∆hap4* t90*, ∆zap1* t90*, ∆swi4* t90*, ∆gln3* t90*, ∆cin5* t90*, wt* t90*, ∆hmo1* t60*, ∆hmo1* t30*, ∆gln3* t30*, ∆zap1* t60*, ∆zap1* t30*, ∆swi4* t60*, ∆swi4* t30*, ∆hap4* t60*, ∆zap1* t15*, ∆hmo1* t15*, ∆hap4* t30*, ∆hap4* t15*, ∆gln3* t30*, ∆gln3* t15*, ∆cin5* t60*, ∆cin5* t30*, ∆cin5* t15*, wt* t60*, wt* t30*, wt* t15*.* |

The heatmap (Figure 6) shows that there are differences in the expression of genes during cold shock versus the recovery period. Certain clusters showed more extreme changes in expression, such as in clusters 2, 9, 8 and 1, 5, 7 (Figure 6). Other clusters do not display as strong of a change, as shown in clusters 10, 4, and 6 (Figure 6).

There are fifteen regulatory transcription factors in the candidate db5 network modeled in the current study (ACE2, ASH1, CIN5, GCR2, GLN3, HAP4, HMO1, MSN2, SFP1, STB5, SWI4, SWI5, YHP1, YOX1, and ZAP1). Eight of them (ACE2, GCR2, GLN3, HAP4, STB5, SWI4, SWI5, and ZAP1) were not present in the k-means clustering analysis because they did not change expression significantly in at least one strain (Table 1). The remaining seven were spread out among different clusters. ASH1 is present in cluster 10. This indicates that the ASH1 expression change was significant for at least one time point in at least one strain, but there was only a slight downregulation of the gene during cold shock and the recovery period. Ash1 is a transcription factor which negatively regulates mating type switching and the G1/S checkpoint in mitosis (Ash1, SGD). CIN5 and HMO1 are found in cluster 6, which indicates that the expression of these genes was increased during the cold shock time points and returned to baseline during the recovery period. CIN5 is involved in the regulation of DNA binding and the salinity stress response (Cin5, SGD). Hmo1 is associated with DNA structure compaction and the maintenance of the yeast genome (Hmo1, SGD).

|  |  |
| --- | --- |
| **Table 1:** Regulatory transcription factors in the db5 GRN and the *k*-means cluster number in which they are found. | |
| **Gene** | **Cluster** |
| ACE2 | none |
| ASH1 | 10 |
| CIN5 | 6 |
| GCR2 | none |
| GLN3 | none |
| HAP4 | none |
| HMO1 | 6 |
| MSN2 | 4 |
| SFP1 | 4 |
| STB5 | none |
| SWI4 | none |
| SWI5 | none |
| YHP1 | 4 |
| YOX1 | 2 |
| ZAP1 | none |

The list of genes for each of the clusters was loaded into the Gene Ontology term enrichment tool found on the Gene Ontology website (<http://geneontology.org/>) and the list of significant (p<0.05) Gene Ontology terms associated with each cluster was generated. Cluster 4 was not significantly enriched for any GO terms, but the rest of the clusters were significantly enriched for particular GO terms.

Clusters 1, 3, 5, and 7 contain genes that were upregulated during cold shock and downregulated during the recovery period, which is similar to the expression pattern of STEM profile 45. GO categories overrepresented in cluster 1 included ribosome biogenesis and RNA metabolic processes. These results are consistent with the GO terms enriched from STEM profile 45 and the findings of Al-Fageeh et al. (2006) and Albert et al. (2019), which were discussed earlier. Cluster 3 is enriched for GO categories involved in RNA metabolic processes, such as transcription, processing, and degradation. Schade et al. (2004) found that RNA metabolism is increased during the early cold shock response, which is consistent with these findings. Cluster 5 is enriched for GO terms associated with carbohydrate (pentose, mannose, fructose, and glucose) and proton transmembrane transport. The mobilization of carbohydrates in yeast has been shown to increase during the early cold shock response (Schade et al., 2004). Reserves of carbohydrates, specifically glucose and trehalose, have been associated with cold shock (Schade et al., 2004). Therefore, the transport and metabolism of carbohydrates is likely a response by the cells to the cold in order to maintain cell membrane fluidity and resist freezing (Al-Fageeh et al., 2006). Lastly, cluster 7 is enriched for GO categories that are associated with cellular zinc homeostasis and fatty acid metabolic processes. Lipid metabolism has been shown to be upregulated in cold shock, which is consistent with these findings (Schade et al., 2004).

Clusters 2, 8, and 9 contain genes that were downregulated during cold shock and upregulated during recovery, which is a similar expression pattern to STEM profiles 9 and 22. GO terms enriched in cluster 2 include terms that are associated with oxidation-reduction and phosphorus metabolic processes. Cluster 8 is enriched for GO terms involving protein folding, amino acid metabolism, and RNA processing. These results are consistent with the GO term analysis of STEM profile 9, but inconsistent with previous findings, which report an upregulation of genes associated amino acid metabolism during the early cold response (Schade et al., 2004). However, genes involved with protein folding have been shown to be downregulated during the early cold response, but induced during the late cold response (Al-Fageeh et al., 2006). Cluster 9 was enriched for only two GO terms: glucose 6-phosphate metabolic processes and generation of metabolite precursors. Glucose 6-phosphate is an intermediate in glycolysis and its presence negatively regulates the progression of glucose metabolism (Berg et al., 2002). Therefore, when there are high amounts of glucose 6-phosphate, glycolysis rates decrease. The downregulation of metabolite precursors and glucose 6-phosphate indicates that metabolism increases, which is consistent with Murata et al. (2006), which determined that energy and metabolism functions increase during cold shock.

Clusters 6 and 10 contained genes that did not show major changes in expression between the cold shock and recovery time points. There was an overall upregulation for the genes in cluster 6, with more extreme upregulation during the cold shock time points. Only one GO term was enriched in this cluster: mitochondrial translation. There was an overall downregulation for the genes in cluster 10 for both the cold shock and recovery time points. This cluster was enriched for GO categories associated with amino acid metabolism and translational termination. These results are consistent with the findings from profile 9 of the STEM analysis and cluster 8 of the *k-*means analysis, which showed strong downregulation during cold shock, but are inconsistent with Schade et al. (2004), which indicated amino acid metabolism and protein synthesis are upregulated during cold shock.

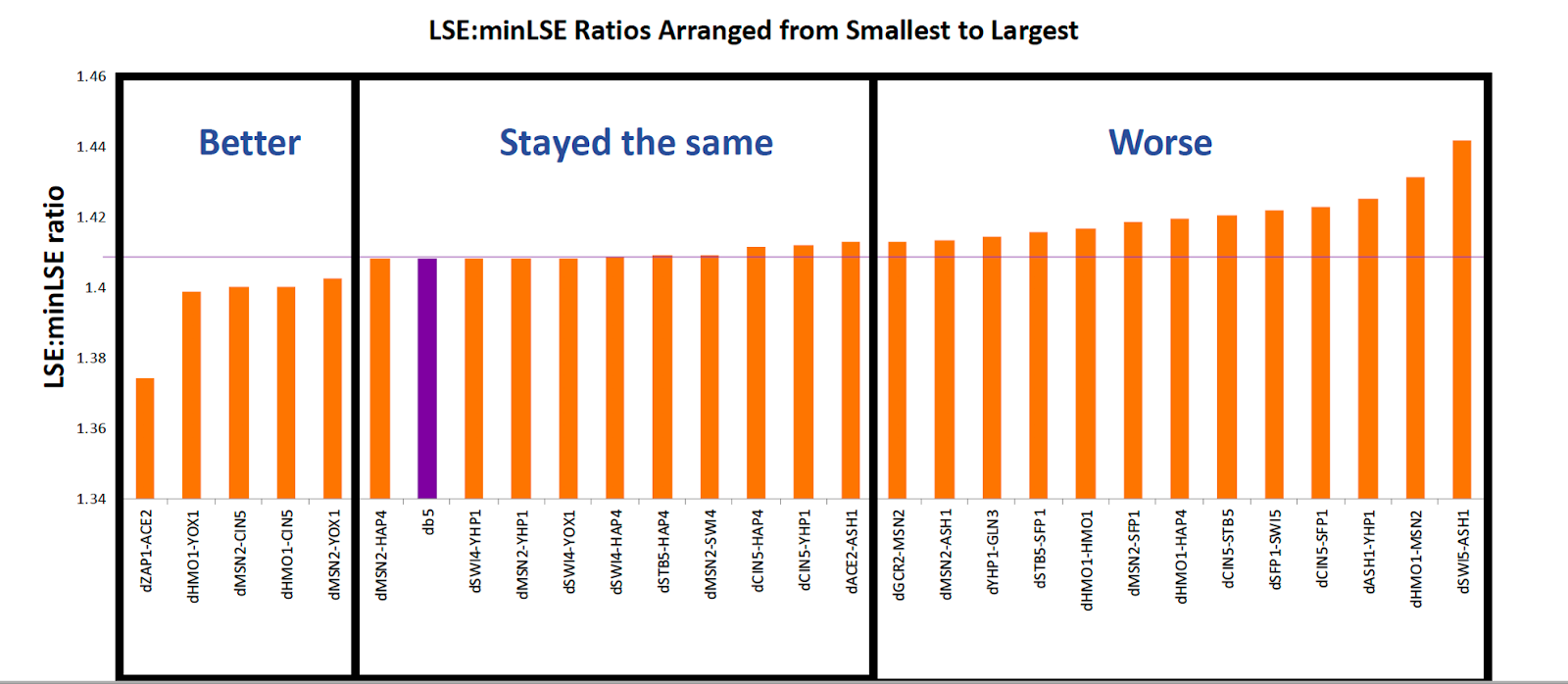
The enrichment analysis of Gene Ontology terms was consistent for the STEM and *k-*means clusters that showed the same expression patterns throughout the timepoints, which was expected because the data used in the clustering was the same. Further, a comparison of these results to previous studies found certain consistencies, which validates some of the trends in expression that were seen. However, certain GO terms were inconsistent with previous studies, such as the downregulation of amino acid metabolism during cold shock, which was found in our study, but is inconsistent with the findings of Schade et al. (2004). These inconsistencies could be explained by the differences in experimental design, such as yeast strain, temperature of growth, type of growth apparatus, RNA processing and chip protocols, and type of data analysis between our study and previous studies. Further, our methods of statistical and bioinformatics analyses could be different from the previous studies which would impact our results. Lastly, GO enrichment analysis of this study was conducted over ten years after most of the previous studies discussed in this paper. Therefore, the annotations are overall better than those from a decade ago.

**2. Sensitivity of the Model Parameters to Network Structure and Strain Data**

Kelly (2019) originally estimated the parameters for the db5 network. To extend this study we performed a series of modifications to the model input to assess the sensitivity of model parameters to different modeling conditions, discussed in this and subsequent sections of this paper.

**Systematic Deletion of Edges from the db5 Network**

To determine the importance of each edge in the intact network, the edges of the GRN were systematically deleted, one at a time, generating 28 new GRNs whose parameters were then estimated with GRNmap. To compare the fit of these runs, the LSE:minLSE ratio of each was determined. A ratio higher than that of the intact network indicates that the model is a worse fit, while a ratio lower than the intact network indicates that the model is a better fit. In these 28 edge-deletion networks, LSE:minLSE ratios indicated that five networks performed better than the intact network, while ten networks performed about the same, and thirteen performed worse (Figure 7). The edge-deletions involving the Hmo1, Msn2, and Cin5 transcription factors resulted in a poor performance of the model, indicating that those edges represent important regulatory relationships in the cold shock response (Kelly, 2019).



**Figure 7:** LSE:minLSE ratios for each of the 28 runs, ordered smallest to largest (Kelly, 2019).

Kelly (2019) grouped the model runs based on an “eyeball” analysis of the LSE:minLSE ratio. To determine whether the edge weight values were similar to each other in these three categories, *k*-means clustering was performed on the weight values from the intact network and edge-deletion networks, resulting in three clusters. Sixteen edge-deletion networks occurred in the same cluster as the intact network, while the other two clusters had four and eight networks, respectively. The “better” category (Figure 7) networks formed cluster 1, except for the network with the smallest LSE:minLSE ratio (dZAP1🡪ACE2), which was found in cluster 3. The “stayed the same” category (Figure 7) networks were all found in cluster 2. Cluster 3 contained networks that were found in all three of the “better”, “stayed the same” and “worse” categories. The results of the *k*-means clustering shows that both the LSE:minLSE ratio and the actual weight values must be taken into account when evaluating network performance. Model runs with either better or worse LSE:minLSE ratios can lead to similar weight values, suggesting that the differences in the LSE:minLSE ratios are not as meaningful as we have made them out to be. In the future, expression plots and MSE:minMSE values for the individual genes should be examined to determine what is contributing to changes in LSE:minLSE ratios.

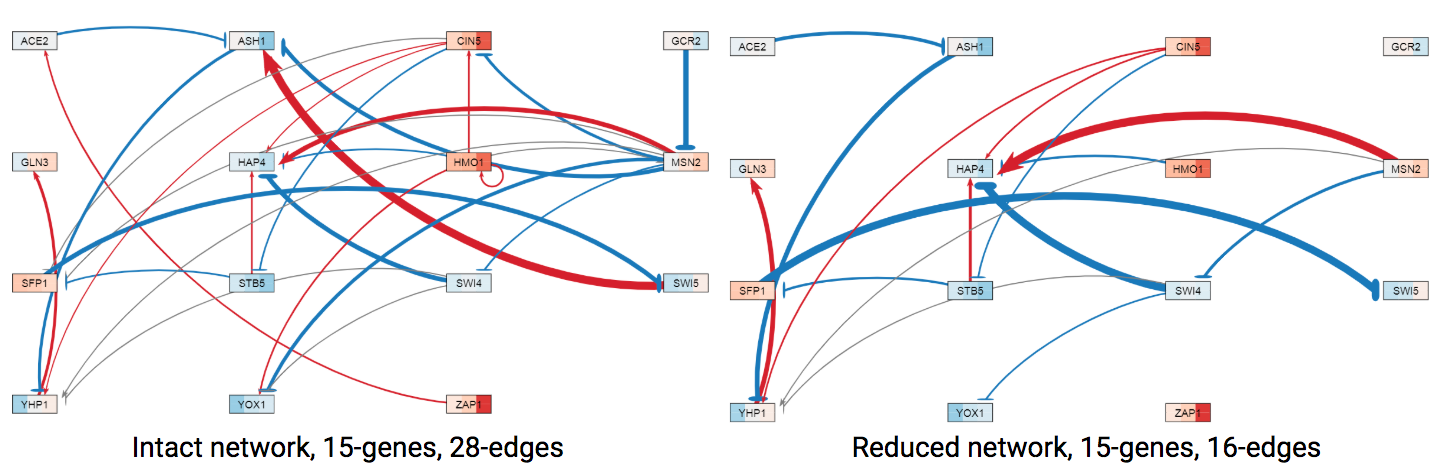
An examination of the genes involved in the edge-deletions for each cluster showed that certain transcription factors were frequently found in clusters 1 and 3 that showed deviations from the pattern of weight values for the intact network. Weight values deviated from those of the intact network for 6 of 9 edge-deletions involving Msn2, 4 of 5 edge-deletions involving Hmo1, and 3 of 6 edge-deletions involving Cin5 (Figure 8). In addition, the edge-deletions involving Gcr2 and Zap1 caused major deviations in weight values from the intact network (Figure 8). This edge-deletion analysis reveals the importance of these regulatory relationships in the cold shock response.



**Figure 8:** *k*-means clustering of the edge-deletion model runs. The three clusters are marked as 1, 2, and 3. The LSE:minLSE ratio for each of the runs is represented by a bar chart. The purple bar and line indicate the ratio value for the intact network. The heat map shows the regulatory weight values for each of the 28 edges in the network in each model run. Red indicates activation and blue indicates repression. The intensity of the color reflects the magnitude of the activation or repression.

These results suggest that the transcription factors Msn2, Hmo1, and Cin5 are important for the cold shock stress response. Gasch et al. (2000) showed that Msn2 and Msn4 are key transcription factors for regulating the general environmental stress response in yeast. Msn2 and Msn4 bind to the stress response element (STRE) sequence, which causes the induction of the stress response (Gasch et al., 2000). In addition, Schade et al. (2004) reported that the late cold response is controlled by the Msn2/Msn4 transcription factors, inducing a large fraction of the genes associated with the late response, such as those involved in glycogen and trehalose synthesis. However, the Msn2/Msn4 complex has not been shown to influence the early cold shock response (Schade, 2004). With respect to HMO1, Berger et al. (2007) reported that the deletion of the HMO1 gene caused a decrease in the transcription of genes indirectly associated with ribosome biogenesis. Thus, Hmo1 could be an important transcription factor for regulating the cold shock response because genes associated with the GO term ribosome biogenesis are overrepresented among genes that are upregulated during cold shock. Finally, Cin5 has been found to be induced upon stress conditions and regulates genes associated with ribosome biogenesis, protein synthesis, and maintaining chromosome stability (Uniprot, 2019; Fernandes et al., 1997). Disruptions to the pattern of weights when edges involving Msn2, Hmo1, or Cin5 are deleted from the network are supported by literature showing the importance of these transcription factors in controlling the stress response.

Msn2, Hmo1, or Cin5 regulate or are regulated by multiple other transcription factors, as shown in the candidate network (Figure 1)., While most of the edge deletions involving these transcription factors caused changes to the regulatory weight pattern compared to the intact network, not all of their associated edge-deletions did. Thus, we wondered whether, a smaller network where the edges in clusters 1 and 3 of figure 8 were deleted from the GRN would show a better model fit. Once these edges were deleted, the resulting network had 16 edges. In this GRN, the Gcr2 and Zap1 nodes were left unattached to other nodes because the edges connecting them to the other nodes in the network were removed (Figure 9).



**Figure 9:** Visualization of the differences of thegene regulatory networks for db5 were modified by manipulating the edges clusters 1 and 3, where GRNsight input files were manipulated with no re-estimation of parameters. The left GRN is the intact network with 15 genes and 28 edges and the right GRN is the network where the edges not included in cluster 2 were deleted.

Because Gcr2 and Zap1 were detached from the network when their edges were deleted, and because their individual edge deletions severely impacted the overall pattern of edge weights (Figure 8), three new model estimations were run to determine the fit of the model when the intact network was made smaller (Table 2). In one model run, the network was left intact, with 15 genes and 28 edges, but the *Δzap1* expression data sheet was completely removedfrom the input. Another model was run where the edges in clusters 1 and 3 were deleted so as to create a 15-gene and 16-edge network (Figure 9, reduced network). The last model involved the deletion of the GCR2 and ZAP1 nodes and their associated edges, resulting in a 13-gene network, 26-edge network. To analyze the fit of the models compared to the intact network, the LSE:minLSE ratios of each of the models were compared.

|  |  |  |
| --- | --- | --- |
| **Table 2:** Models run to determine the importance of Gcr2 and Zap1 in the network. The edges included in each of the networks and the data included in the model run are indicated. | | |
| **Genes** | **Edges** | **Data** |
| **15** | **28-Edge** | All data |
| **15** | **16-Edge:** Those within the same k-means cluster as the intact network were included | All data |
| **13** (No Gcr2, Zap1 genes) | **26-Edge:** Edges involving Gcr2 and Zap1 were deleted | All data |
| **15** | **28-Edge** | *Δzap1* data deleted |

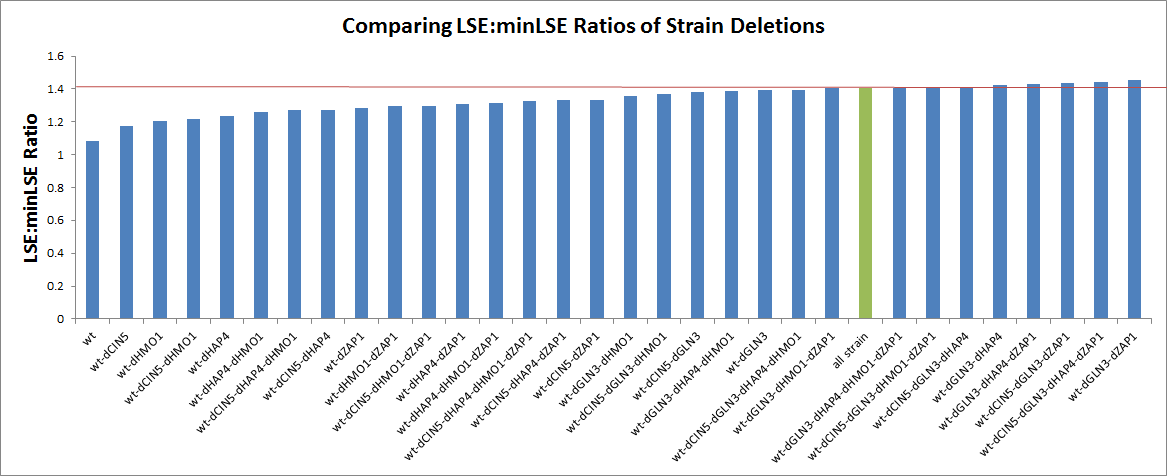
The LSE:minLSE ratio for the 15-gene, 16 edge network was about the same as the intact network, indicating that there was not a major difference in the fit of the model parameters (Figure 10). When only the two edges involving Gcr2 and Zap1 were deleted (13-gene, 26-edge network), the LSE:minLSE ratio decreased even further, indicating a better fit of the model parameters. When *Δzap1* data was deleted from the model, the LSE:minLSE ratio decreased slightly, indicating a slightly better fit (Figure 10). These results indicate that the model is sensitive to the deletion of the Gcr2 and Zap1 edges from the network. Gcr2 is a transcription factor which activates glycolytic genes, thereby increasing carbohydrate metabolism and energy production (Gcr2, SGD). Chen et al. (2009) found that Gcr2 was found to be the only consistently upregulated stress-response regulator in all of the environmental stress conditions they tested. Furthermore, Zap1 is a transcription factor that is itself regulated by the presence of zinc in the environment (Zap1, SGD). In the absence of zinc, Zap1 represses other transcription factors while in the presence of zinc, it induces other transcription factors. Therefore, both Gcr2 and Zap1 regulate other transcription factors when introduced to some environmental stress.

|  |
| --- |
| **Figure 10:** LSE:minLSE ratios for the model runs corresponding to Table 2: the 15-gene, 16-edge reduced network (16-edge), intact network without GCR2 and ZAP1 (13-gene, 26-edge), no dZAP1 data included (15-gene, 28-edge no Zap1), and the intact 15-gene, 28 edge db5 network. |

**Variable Inclusion of Strain Expression Data**

Up until this point, most of the model runs on the db5 network were performed using all available microarray data from each strain (wild type, *Δcin5, Δgln3, Δhap4, Δhmo1,* and *Δzap1* strains). To test the sensitivity of the model parameters to the inclusion of each strain’s data, 32 model runs were conducted with the wild type expression data as input and all possible permutations of inclusion of the other strains’ data. The runs were generated with wild type (*wt*) only, *wt* plus one, *wt* plus two, *wt* plus three, *wt* plus four, and *wt* plus all five deletion strains (A table listing the runs can be found in the appendix). A re-estimation of model parameters, including production rates, expression thresholds, and regulatory weights, was performed in GRNmap.

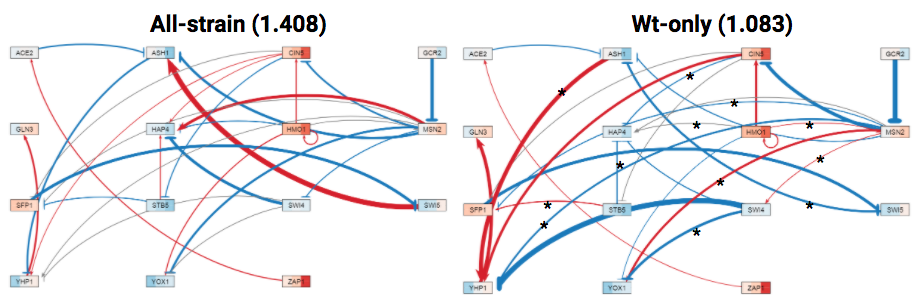
In order to assess the fits of the model parameters, the LSE:minLSE ratio was determined for each run and compared between the 32 models (Figure 11). A ratio higher than that of the intact network indicates that the model has a worse fit, while a lower ratio indicates that the model has a better fit than the intact network. The LSE:minLSE ratio increased for eight of the new models. Of these eight models, all of them included expression data from the *Δgln3* strain. The LSE:minLSE ratio decreased for 23 of the models, including the wild-type only model and all of the wild-type plus one deletion strain models. Overall, the model run with wild-type only expression data resulted in the best fit, while the inclusion of the *Δgln3* strain expression data resulted in a fit that was about the same as the all-strain network or worse. Further research is necessary to determine why the inclusion of *Δgln3* strain expression data causes a worse fit. This could be the result of either technical or biological reasons. Technically, the expression data from the *Δgln3* strain may be of a poorer quality than that of the other strains. Biologically, the deletion of the Gln3 transcription factor may have a stronger effect on gene expression than the other deletion strains, forcing the model to have to fit very different patterns of expression for each gene.

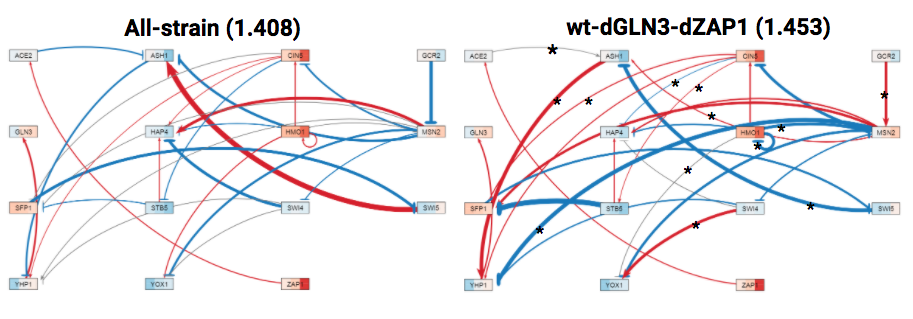


**LSE:minLSE Ratios of Model Runs with Variable Inclusion of Strain Data**

**Figure 11:** LSE:minLSE ratios for each of the model runs that had variable inclusion of strain expression data. The models are ordered by increasing LSE:minLSE ratio. The green bar represents the all-strain network and the red line indicates the ratio value for the all-strain network.

The networks were then visualized with GRNsight (Figures 12-14). The all-strain intact network was compared to the network model with the lowest LSE:minLSE ratio (wt-only; Figure 12), the network model with the highest LSE:minLSE ratio (wt-dGLN3-dZAP1; Figure 13), and the network that included expression data from all strains except the *Δgln3* strain (wt-dCIN5-dHAP4-dHMO1-dZAP1; Figure 14). Compared to the all-strain network, the “wt-only” GRN had 12 edges that changed the sign of their weight values (Figure 12). The “wt-dGLN3-dZAP1” GRN had the highest of the LSE:minLSE ratios. When compared to the all-strain network, 13 of the 28 edges changed the sign of their weight values (Figure 13). Furthermore, whenever *Δgln3* expression data was included, the model consistently performed worse. Thus, a comparison was made between the all-strain data model and the model where only the *Δgln3* expression data was not included. The network run using the wt-dCIN5-dHAP4-dHMO1-dZAP1 expression data had only two edge weights that changed sign compared to the all-strain network (Figure 14). However, both of these edges were thin in the intact and wt-dCIN5-dHAP4-dHMO1-dZAP1, indicating that the sign flipping did not have a strong effect the network dynamics. These resgults indicate that the model is sensitive to which strains’ data is used to estimate the model parameters, particularly the *Δgln3* expression data.

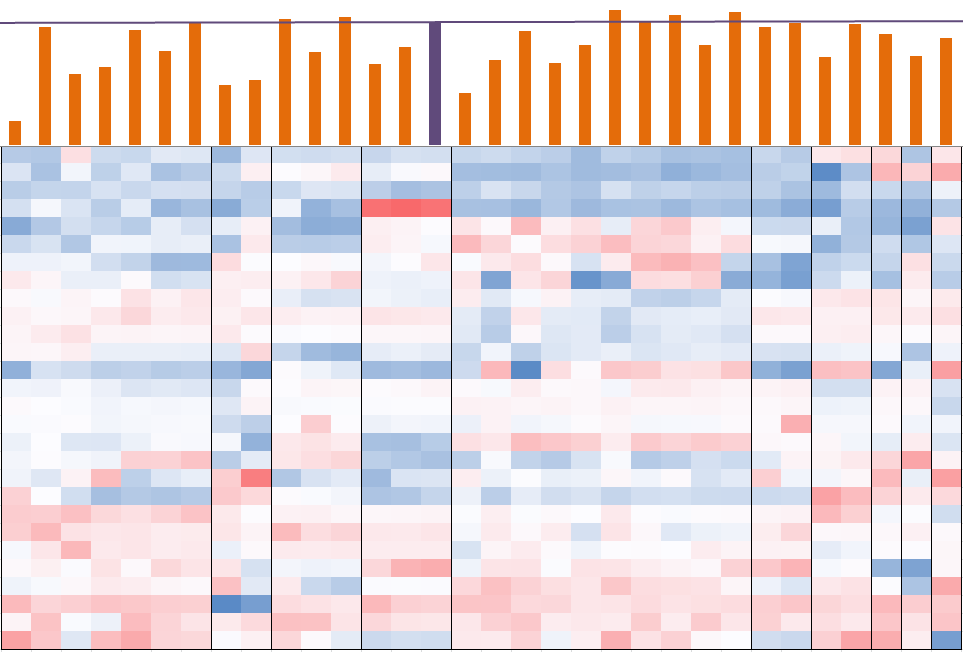
**Figure 12:** GRNs visualized with GRNsight. The red edges indicate activation and the blue indicate repression. The thickness of the edge represents the strength of that relationship. In the GRN to the left, all expression data was used in estimation, and in the network to the right, only wild type expression data was used in the estimation. The LSE:minLSE ratio is given in the parentheses. The asterisks on the edges of the wt-only GRN indicate edges that have changed sign from the all-strain network. 15 of the 28 edges changed.

**Figure 13:** GRNs visualized with GRNsight.The red edges indicate activation and the blue indicate repression. The thickness of the edge represents the strength of that relationship. In the GRN to the left, all expression data was used in estimation, and in the network to the right data from wt, dGLN3 and dZAP1 was used in the estimation. The LSE:minLSE ratios are given in the parentheses. The asterisks refer to the edges that changed.

|  |
| --- |
| **Figure 14:** GRNs visualized with GRNsight.. The red edges indicate activation and the blue indicate repression. The thickness of the edge represents the strength of that relationship. In the GRN to the left, all expression data was used in estimation, and in the network to the right data from wt, dCIN5, dHAP4, dHMO1, dZAP1 was used in the estimation. The LSE:minLSE ratios are given in the parentheses. The asterisks indicate edges that have changed in the new model. |

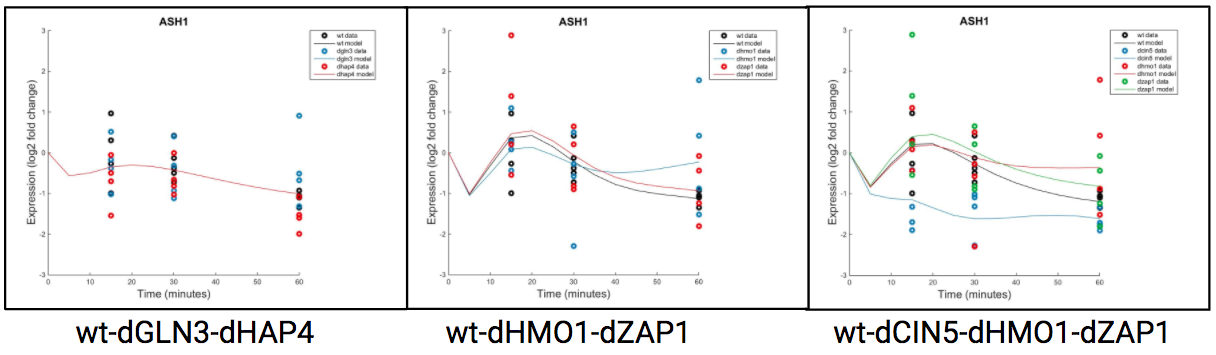
To discern patterns in the weight values, it was necessary to use *k*-means clustering in MATLAB and generate a heat map (Figure 15). The weight values were clustered in both dimensions. Cluster 4 contained the model run with all strain expression data and had a pattern of weight values that was different than all other clusters (Figure 15). The two model runs that were most similar to the all-strain’s regulatory weights were the models that either excluded the *Δgln3* strain expression data or both the *Δgln3* and *Δzap1* expression data (Figure 15). As noted before, the runs that included *Δgln3* data had the highest LSE:minLSE ratio, indicating the worst fit. This indicator is reflected in the weight clusters, where the regulatory weights values did not deviate from the all-strain network as severely when *Δgln3* and *Δzap1* data were not included in the model (Figure 15).

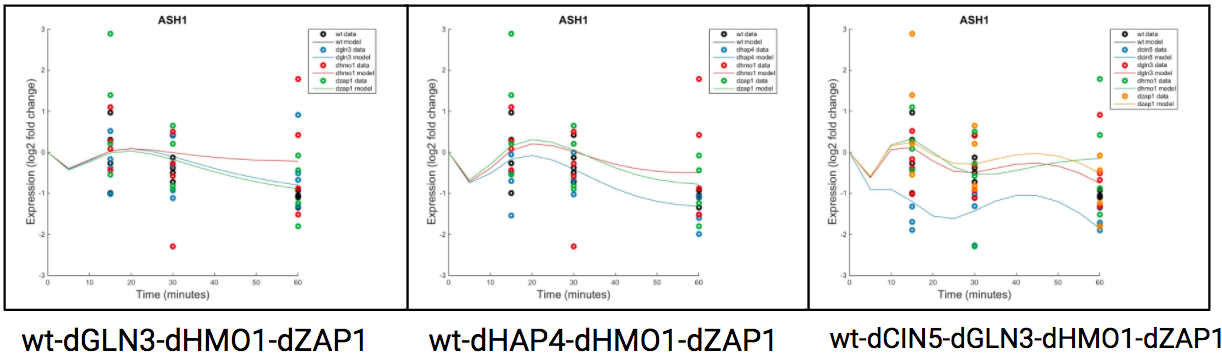
Another way to compare models is to look at the expression plots for individual genes. Figure 16 shows a sampling of expression plots for the ASH1 gene for different model runs. The dots represent the replicate experimental data points for each timepoint and strain, and the lines represent simulated expression when the differential equation for ASH1 is solved with the estimated model parameters (production rate, weights, and threshold b). When the simulated data lines line up on top of each other and look like a single line, the model shows that there are no differences in the expression pattern between strains. When the simulated lines diverge and multiple lines appear, the model shows that the expression for that gene is different between strains. In Figure 16, when the expression data from *Δcin5*, *Δhmo1*, and *Δzap1* data were not included, the simulated model data did not cause for ASH1 to diverge from each other, indicating that the expression is similar to that of the wild type strain. However, when *Δ*cin5, *Δhmo1*, and *Δzap1* data were included, the simulated model data did diverge, indicating that the expression of ASH1 is dissimilar to the other strains (Figure 16). These results indicate that the strains have different expression patterns for ASH1. (The rest of the expression plots can be found in the appendix.)

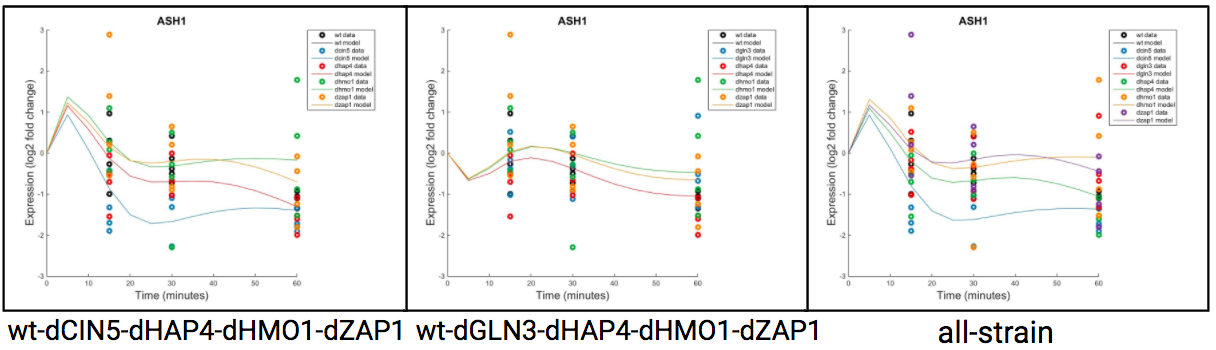


**1 2 3 4 5 6 7 8 9 10**

**Figure 15:** Heatmap of regulatory weight clusters. The heatmap shows the results of k-means clustering in both the x and y dimensions, using ten clusters. The bars indicate the LSE:minLSE ratios for the models in that column. The purple bar and line indicates the ratio for the intact network. Clusters are demarcated by the black lines. Red indicates activation and blue indicates repression. The intensity of the color reflects the magnitude of the activation or repression.







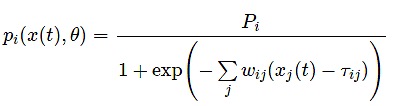
**Figure 16:** Expression plots the ASH1 gene for model runs with variable inclusion of strain data. The dots represent the replicate experimental datapoints for each timepoint and strain, and the lines represent simulated expression when the differential equation for ASH1 is solved with the estimated model parameters (production rate, weights, and threshold b). The color of the dots indicates the strain data, which is given in the legends. Divergence is indicated by multiple simulated expression lines, while no divergence results in a single expression line.

The variable inclusion of strain data impacted the fit of the model, with models including *Δgln3* data performing worse than the intact network. This trend is reflected in the LSE:minSLE ratios, GRNs, *k-*means clustering, and expression plots from the model runs. Gln3 is a transcription factor that regulates glutamine metabolism and has been found to be associated with the nitrogen catabolite repression system (Tate et al., 2007; SGD). Schade et al. (2004), showed that upregulated functional categories during the early cold shock response included those that are associated with amino acid metabolism. Both Schade et al. (2004) and the current study show that Gln3 plays a role in the cold shock response in *S. cerevisiae.*

**3. Sensitivity Assessment of the Model to Production Rates**

GRNmap performs estimations of regulatory weight, expression threshold, and production rate parameters. Up to this point, each of these parameters has been estimated in each model run, but options in GRNmap exist to input the threshold and production rate parameters as fixed values.  To determine the impact of fixing or estimating these parameters, models were run where the production rates and thresholds were either estimated or fixed, alone or in combination.

These parameters are modeled using ordinary differential equations in GRNmap. The production rate is modeled using the sigmoidal function:

, where *Pi* is the initial guess production rates and *pi*is the estimated rate (Dahlquist et al., 2015). The expression over time is then modeled by production minus degradation rate.

The threshold indicates the point at which activators either induce production, when expression is above the threshold, or stop production, when expression is below the threshold (Dahlquist et al., 2015). Repressors induce production when the expression is below the threshold level, and stop production when the expression is above threshold (Dahlquist et al., 2015). In the model, the threshold is modeled using:

https://lh4.googleusercontent.com/AGQtm2MrzQPC6JV_VCxIj2n_ydkK4ZMGBl7TnlnEc4CnSFxR8eQ5JAQdPwm7oBKojKbhlkR3Pqu5dWJPurGWnuuxf7uDea-pXg-GTWRNylR2X8Uyf5BegqQgBgPwvjzqLdZjZlZP, where *wi j* is the weight of a regulatory relationship and τ*i j* is the threshold at which the production turns off or on (Dahlquist et al., 2015). In standard model runs, the initial guess for the threshold is set to zero.

**Comparison of Estimated Production Rates with Published Experimentally-derived Rates**

For the model runs described previously, the production rates were always estimated and the initial guesses were set as two times the degradation rate for each gene. The degradation rates were experimentally determined by Neymotin et al. (2014) who measured mRNA half-life data for all genes in yeast using a method called RATEseq. Assuming that the system is in steady state with the relative expression of all genes equal to 1, (P/2) - lambda = 0, P is the production rate and where lambda is the degradation rate, is a reasonable initial guess. We wondered how closely the production rates estimated by GRNmap matched the experimentally-derived rates measured by Neymotin et al. (2014). To perform this comparison, production rates were derived from the Neymotin et al. (2014) data. The production rates were calculated using k = α[RNA], where k is the rate constant for synthesis and α[RNA] is the RNA abundance (Neymotin et al., 2014). This RNA abundance was calculated as α = αRNA + αgrowth, where α is the RNA concentration and αRNA is the degradation rate constant and αgrowth is the cell’s division rate constant (Neymotin et al., 2014). Using this equation, the Neymotin et al. (2014) production rates were derived from their dataset (Table 3).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 3:** Comparison of production rates. Left to right: initial guesses based on 2X the degradaton rate from Neymotin et al. (2014), estimated production rates from the GRNmap all-strain data model run, Neymotin et al. (2014) production rates, the GRNmap estimated rate minus the Neymotin et al. (2014) rate, and the percent change. | | | | | |
| **Gene** | **Initial Guess Production Rate** | **Estimated Production Rate** | **Neymotin Production Rate** | **Estimated - Neymotin** | **Percent Change (%)** |
| **ACE2** | 0.224 | 0.202 | 0.180 | 0.022 | 12.22 |
| **ASH1** | 0.433 | 1.677 | 1.037 | 0.640 | 61.72 |
| **CIN5** | 0.201 | 0.656 | 0.063 | 0.593 | 941.27 |
| **GCR2** | 0.193 | 0.232 | 0.327 | -0.095 | -29.05 |
| **GLN3** | 0.322 | 0.302 | 0.365 | -0.063 | -17.26 |
| **HAP4** | 0.272 | 1.302 | 1.827 | -0.525 | -28.74 |
| **HMO1** | 0.099 | 0.307 | 1.406 | -1.099 | -78.17 |
| **MSN2** | 0.408 | 2.557 | 0.487 | 2.070 | 425.05 |
| **SFP1** | 0.693 | 1.555 | 1.199 | 0.356 | 29.69 |
| **STB5** | 0.140 | 0.120 | 0.080 | 0.040 | 50.00 |
| **SWI4** | 0.283 | 0.316 | 0.157 | 0.159 | 101.27 |
| **SWI5** | 0.322 | 1.921 | 0.340 | 1.581 | 465.00 |
| **YHP1** | 0.173 | 0.208 | 0.283 | -0.075 | -26.50 |
| **YOX1** | 0.730 | 1.391 | 1.028 | 0.363 | 35.31 |
| **ZAP1** | 0.104 | 0.128 | 0.082 | 0.046 | 56.10 |

In comparing the estimated production rates and the Neymotin et al. (2014) rates, there are differences between the values for each of the genes. For ten of the fifteen genes, the estimated production rate was greater than the Neymotin rate, while five of the genes had a lower estimated rate (Table 3). Certain genes, such as CIN5, MSN2 and SWI5 showed a major difference between the Neymotin et al. (2014) and estimated production rates (>400% change), revealing inconsistency in the rates between them (Table 3). However, other genes, like GLN3 and ACE2, had relatively consistent production rates between the estimated and Neymotin et al. (2014) production rates (< 20% change; Table 3).

Figure 17 shows a scatter plot comparing the GRNmap-estimated production rates and the Neymotin et al. (2014) rates. A slightly positive linear relationship can be seen, indicating there is some relationship between the production rates, but the R2 value of 0.1631 is small. CIN5, MSN2, SWI5, noted above are contributing to the skew of the line away from a 1:1 relationship.

|  |
| --- |
| **Figure 17:** Neymotin et al. (2014) production rates compared to with the estimated rates. The associated gene is labeled to the right of the data points and the solid blue line is the linear trendline for the data points. The dotted gray line is a reference line with a slope of 1. |

For a group of genes in the lower-right quadrant of the graph, production rates match quite well. Further analysis of why the production rates for these seven genes better match experimentally-derived data is necessary.

**The Effect of Fixing or Estimating Thresholds and Production Rates**

Besides db5, there are six other candidate GRNs previously generated in the Dahlquist lab (db1-db7). All seven of these GRNs were used in conjunction with all-strain expression data in a set of experiments to determine the effect of either fixing or estimating the production rates and thresholds. Six model runs were conducted per network where the weights (*w*) were estimated and thethresholds *(b)* and production rates *(P)* were either estimated or fixed in the model (Table 4). For the models where the production rates were fixed, two runs were conducted with the initial guesses from or the experimentally-derived Neymotin et al. (2014) production rates (Table 4). For the models that were run with fixed threshold values, the b value was set to 0. The models were run in GRNmap and from the output, the LSE:minLSE ratio was determined for each model in order to assess and compare the fit of the models.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 4:** List of model runs indicating which parameters were fixed or estimated. Weights were estimated for all model runs, while the threshold and production rate were variably estimated or fixed between the runs. The fixed production rates were derived either by doubling the Neymotin et al. (2014) degradation rates or using the Neymotin et al. (2014) production rates as discussed in the text.  The fixed threshold value was set to 0. | | | |
| **Model Run** | **Weight *(w)*** | **Threshold *(b)*** | **Production Rate *(P)*** |
| **Est *P, b, w*** | Estimated | Estimated | Estimated |
| **Est *P, w;* Fix *b*** | Estimated | Fixed | Estimated |
| **Est *b, w;* Fix *P*** | Estimated | Estimated | Fixed |
| **Est *b, w;* Fix *P*** (using Neymotin et al. 2014 Production Rates) | Estimated | Estimated | Fixed |
| **Est *w;* Fix *P, b*** | Estimated | Fixed | Fixed |
| **Est *w;* Fix *P, b*** (using Neymotin et al. 2014 Production Rates) | Estimated | Fixed | Fixed |

Variation in the production rates for each of the runs occurred after the re-estimation of parameters (Table 5). To determine the effect of estimating or fixing parameters on the fit of the model, the LSE:minLSE ratios from each model run were compared. The lower the LSE:minLSE ratio, the better the fit of the model. Db5 was chosen as the candidate network for analyzing the effect on model sensitivity, as db1-db7 model runs resulted in the same trends for increased or decreased LSE:minLSE ratio among the estimated/fixed model runs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 5:** Production rates from model runs with fixed parameters (left two columns) or estimated (right two columns). | | | | |
| **Gene** | **Fix *P* (2x *d*)** | **Fix *P* (Neymotin)** | **Est *P, b, w*** | **Est *P, w*, Fix *b*** |
| **ACE2** | 0.224 | 0.180 | 0.202 | 0.124 |
| **ASH1** | 0.433 | 1.027 | 1.677 | 0.785 |
| **CIN5** | 0.201 | 0.063 | 0.656 | 1.247 |
| **GCR2** | 0.193 | 0.327 | 0.232 | 0.236 |
| **GLN3** | 0.322 | 0.365 | 0.302 | 0.263 |
| **HAP4** | 0.272 | 1.827 | 1.302 | 0.884 |
| **HMO1** | 0.099 | 1.406 | 0.307 | 0.160 |
| **MSN2** | 0.408 | 0.487 | 2.557 | 3.316 |
| **SFP1** | 0.693 | 1.199 | 1.555 | 1.598 |
| **STB5** | 0.140 | 0.080 | 0.120 | 0.205 |
| **SWI4** | 0.283 | 0.157 | 0.316 | 0.528 |
| **SWI5** | 0.322 | 0.340 | 1.921 | 3.806 |
| **YHP1** | 0.173 | 0.283 | 0.208 | 0.257 |
| **YOX1** | 0.730 | 1.028 | 1.391 | 1.664 |
| **ZAP1** | 0.104 | 0.082 | 0.128 | 0.129 |

The lowest ratio occurred for the models where *w, b,* and *P* were all estimated. In other words, fixing any parameter decreased the overall fit of the model (Figure 18).  Estimating or fixing the threshold did not change the LSE:minLSE ratio value greatly, which indicates that the threshold does not have a major impact on the fit of the model. Conversely, fixing the production rates made the LSE:minLSE ratio increase, indicating their importance on the fit of the model (Figure 1). In addition, the use of the Neymotin et al. (2014) production rates as the fixed values caused the LSE:minLSE ratio to increase even more, indicating that the initial guesses for the  production rates are a better fit for the model than the Neymotin et al. (2014) values.

|  |
| --- |
| https://lh3.googleusercontent.com/d8SLZTq85iL4uoIm7OTqLabfKN1OpAW27T-EiOJlj1ViVtQ15vKM4mhAQ61MwtgGNFOK4Xz0IYwzK6jNuu1exG9otXA4jfWqD4kPxRUoTlGJEtioyCpyxjSiQMTcRLcAMKqwxP8g  **Figure 18:** The candidate db5 network LSE:minLSE ratios for the estimate/fix production rate (*P*), weight (*w)*, and threshold (*b)* model runs. Production rates used are described in Tables 4 and 5. |

From these results, we found that fixing the production rates, in general, caused the model to perform worse, but the worst fit came from the models that were run with the Neymotin et al. (2014) production rates. This indicates that our initial guess production rates are a better fit for the model than the Neymotin values. Overall, by allowing the production rates to be estimated, the model has more degrees of freedom for finding an optimal fit to the experimental data.

**4. Sensitivity Assessment of the Model to Manipulation of Optimization Parameters**

GRNmap uses the optimization parameters kk\_max, MaxIter, TolFun, MaxFunEval, and TolX to optimize the fit of the model. The parameter kk\_max is the number of times that the optimization loop is rerun, which can improve the model performance. The MaxIter is the maximum iterations that the model will run. This parameter should be high enough that the model does not stop before completing the estimation. Therefore, if the MaxIter is too low, the model may not complete the optimization. The TolFun is the difference between the least squares error of each rerun that signifies when the model is no longer making improvements. The MaxFunEval is the maximum number of times that the function will evaluate the least squares cost. Lastly, the TolX indicates the maximum difference between the least errors cost before the program signifies there is no more improvement of the model (Dahlquist et al., 2018).

The standard optimization parameter values used in the model are 1 for kk\_max, 1.00x108 for MaxIter, 1.00x10-6 for TolFun, 1.00x108 for MaxFunEval, and 1.00x10-6 for TolX.  To assess the sensitivity of the model to changes in these optimization parameters, each parameter was changed by an order of magnitude in either direction for a new set of runs. The LSE:minLSE ratios of the models were compared to that of the standard model run with the midpoint values. The iteration counts of the models were also compared to determine whether the changes to the optimization parameters affected the number of times the model evaluated the least squares error before stopping the model run.

Using the db5 network and all-strain expression data, the optimization parameters in the input Excel workbook were changed by one order of magnitude in each direction for kk\_max, MaxIter, TolFun, MaxFunEval, and TolX (Table 6).  One optimization parameter was changed per new model, with the other parameters set with their standard values, resulting in ten new models.

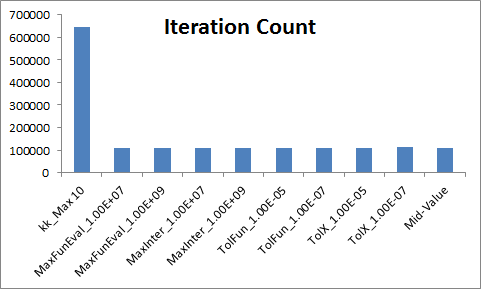
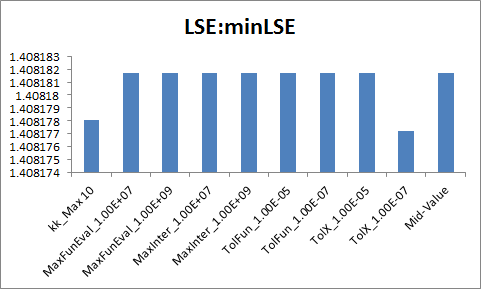
|  |  |  |  |
| --- | --- | --- | --- |
| **Table 6:** Optimization parameter changes. | | | |
| **Optimization Parameter** | **Low** | **Mid** | **High** |
| **kk\_max** | 0.1 | 1 | 10 |
| **MaxFunEval** | 1.00x107 | 1.00x108 | 1.00x109 |
| **MaxIter** | 1.00x107 | 1.00x108 | 1.00x109 |
| **TolFun** | 1.00x10-7 | 1.00x10-6 | 1.00x10-5 |
| **TolX** | 1.00x10-7 | 1.00x10-6 | 1.00x10-5 |

The model run where kk\_max was set to 0.1 failed to complete the estimation, therefore, it was not included in the comparison.

The LSE:minLSE ratio for the run under standard conditions (mid-point value for each parameter) was 1.4081816 (Table 7, Figure 19). When the parameters were changed by one order of magnitude in either direction, the ratio changed only for kk\_max at 10 and TolX at 1.00x10-07, but the change was minimal with a 2.56x10-4% decrease for kk\_max at 10 and a 3.18x10-4% decrease for TolX at 1.00x10-07.   The other runs caused no difference in the ratio (Table 7, Figure 19). Therefore, changing the value for the optimization parameters made very minimal difference to the fit of the model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 7:** Iteration count and LSE:minLSE ratio for the low, mid, and high values for each of the optimization parameters. The asterisks indicate the values that differed from the mid-point values. | | | |
| Optimization Parameter | Value Magnitude | Iteration Count | LSE:minLSE |
| kk\_max | 0.1 | - | - |
| 1 | 111242 | 1.4082 |
| 10 | 645513\* | 1.40817\* |
| MaxFunEval | 1.00x107 | 111242 | 1.40818 |
| 1.00x108 | 111242 | 1.40818 |
| 1.00x109 | 111242 | 1.40818 |
| MaxIter | 1.00x107 | 111242 | 1.40818 |
| 1.00x108 | 111242 | 1.40818 |
| 1.00x109 | 111242 | 1.40818 |
| TolFun | 1.00x10-7 | 111242 | 1.40818 |
| 1.00x10-6 | 111242 | 1.40818 |
| 1.00x10-5 | 111242 | 1.40818 |
| TolX | 1.00x10-7 | 115171\* | 1.40817\* |
| 1.00x10-6 | 111242 | 1.40818 |
| 1.00x10-5 | 111242 | 1.40818 |

The iteration count of the model run indicates the number of times the least squares error is evaluated by the program before the model run stops (Dahlquist et al., 2018). The iteration count for the run under standard conditions (mid-point value for each parameter) is 111,242 (Table 7, Figure 19). Changing the kk\_max to 10 and the TolX to 1.00x10-7 caused a change in the iteration count. The change in iteration count for these parameters was greater than the change that occurred for the LSE:minLSE ratio, especially for kk\_max which showed a 480% increase, while the change in TolX caused a 3.5% increase (Figure 19).  This increase could be expected however, because changing kk\_max to 10 increases the number of times the optimization loop is rerun, which in turn can increase the number of iterations the model runs through. Likewise, decreasing the TolX by one magnitude decreases the maximum allowed difference in the least squares cost before the model stops, which may cause the model to run through more iterations.



**Figure 19:** Comparison of the LSE:min LSE ratio (left) and iteration count (right) for the different optimization parameter manipulations. The mid-value model is the one run with the standard values for each optimization parameter.

 From these sensitivity assessments of the model, it can be determined that changes in the optimization parameters of the model do not impact the fit of the model, but an increase in the kk\_max value does increase the iteration count. In addition, changing the optimization parameters did not impact the production rates, thresholds, or weights compared to the standard model, except for when kk\_Max was increased to 10 and TolX was decreased to 1.00x10-07. However, the changes in values for each of these parameters was minimal, which indicates that manipulations of the optimization parameters does not significantly impact the model. (Tables comparing production rates, thresholds, and weights for the standard, kk\_Max 10, and TolX 1.00x10-07 are given in the appendix.)

**5. Revisiting the Dahlquist et al. (2015) GRN model**

An early version of the GRNmap software was reported in Dahlquist et al (2015). A 21-gene, 31-edge network was derived from the ChIP-chip data reported in Lee et al. (2002) and Harbison et al. (2004). In that study, production rates, weights, and threshold b values were estimated using microarray data from Schade et al. (2004) and protein degradation rates from Belle et al. (2006). Schade et al. (2004) subjected *S. cerevisiae* cells to a cold shock environment (10°C) for 10, 30, and 120 minutes and subsequently harvested the cultures. Results from Dahlquist et al. (2015) can be visualized in GRNsight (Dahlquist et al., 2016; https://dondi.github.io/GRNsight/; demo #4). Since Dahlquist et al. (2015) was published, the GRNmap software evolved to have the following additional features. It can now handle replicate values for the input expression data instead of just the average log2 fold change values, including missing values in the data. It can also handle data from both the wild type strain and multiple deletion strains as explored above. Because of the differences in both the data and the GRNmap software, we wanted to revisit the Dahlquist et al. (2015) model to see how changes to the model and data affect the parameter estimation. A series of model runs was undertaken in a stepwise fashion to examine this (Table 8).

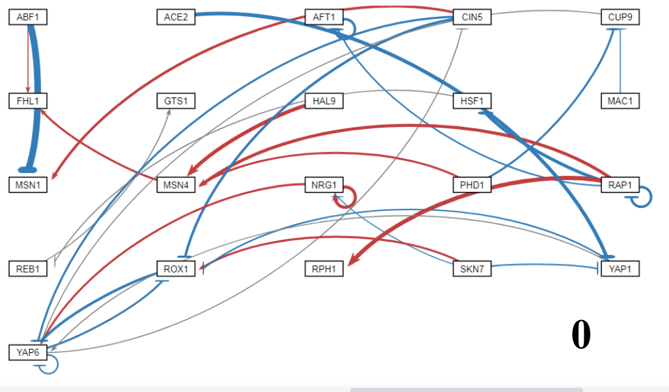
The first model run took the original input file used in Dahlquist et al. (2015) and formatted it for GRNmap v1.10. The second run was performed with the optimization parameters updated to the values used in the current db5-based models: the alpha to 0.02, MaxIter to 1.00E+08, TolFun to 1.00E-06, MaxFunEval to 1.00E+08, and TolX 1.00E-06. This modification was carried throughout subsequent model runs. The third model run included the replicate expression data from Schade et al. (2004) instead of the averages. In the fourth and fifth model runs, the Belle et al. (2006) protein degradation rates were replaced with the Neymotin et al. (2014) mRNA degradation rates. The last model was run with the Schade et al (2004) expression data replaced with the Dahlquist lab wild type expression data. The Dahlquist lab data was compiled using the MS Access Database found at: <https://github.com/kdahlquist/DahlquistLab/blob/master/data/Spring2019/Expression-and-Degradation-rate-database_2019.accdb>.

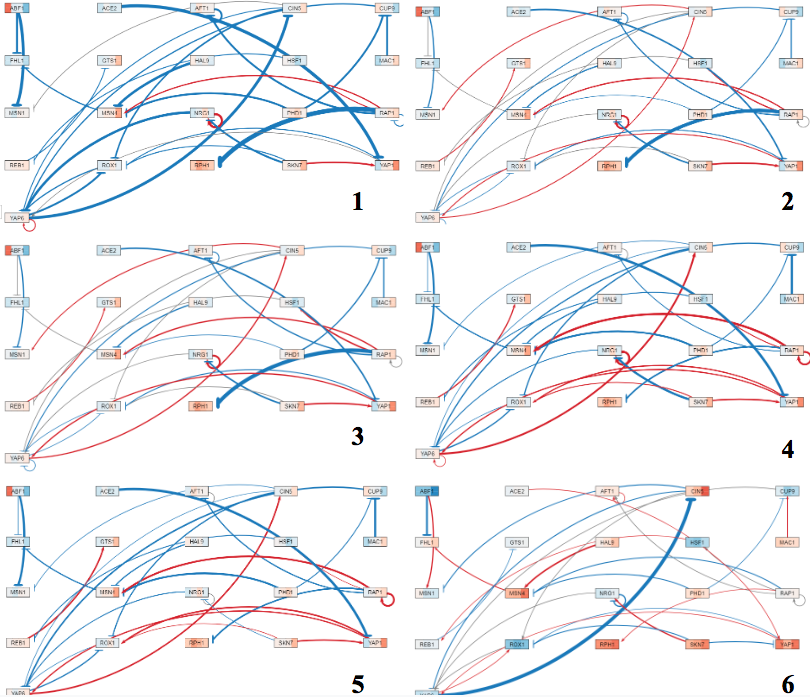
|  |  |
| --- | --- |
| **Table 8:** Model runs for the 21-gene, 31-edge network reported in Dahlquist et al. (2015) using GRNmap v1.10 | |
| **Model Run** | **Manipulation to Input Sheet** |
| **1** | Change format to GRNmap v1.10, Schade et al., 2004 microarray data, Belle et al. 2006 degradation rates |
| **2** | Changes from 1 and Optimization Parameters changed (alpha= 0.02, MaxIter = 1.00x10+8, TolFun = 1.0x10-06, MaxFunEval = 1.00x10+08, and TolX = 1.00x10-06) |
| **3** | Changes from 2 and Schade et al. (2004) replicate log2 fold change expression data instead of averages used |
| **4** | Changes from 2 and Belle et al. (2006) degradation rates and production rate initial guess replaced by those from Neymotin et al. (2014) |
| **5** | Changes from 3 and Belle et al. (2006) degradation rates replaced by those from Neymotin et al. (2014) and Schade et al. (2004) replicate log2 fold change expression data used instead of averages used |
| **6** | Schade et al (2004) expression data replaed with Dahlquist wt-only expression data, Belle et al. (2006) degradation rates replaced by those from Neymotin et al. (2014) |

The resulting six model outputs were visualized in GRNsight to compare changes to the regulatory weights between runs (Figure 20). For each of the visualizations, the edge weight normalization factor in GRNsight was set to 2.971, which was the maximum magnitude weight value from the Dahlquist et al. (2015) model (GRNsight demo #4). Certain edge weights changed sign from activation to repression or vice versa (Figure 20). Compared to the 31 edges of the demo network (Figure 20, network 0), 14 changed sign when the file was reformatted to GRNmap v1.10 (network 1), 8 changed sign when the optimization parameters were changed (network 2), 9 changed sign when replicate data was added (network 3), 10 changed sign when the degradation rates were changed to Neymotin et al. (2014) values with no replicate Schade et al. (2004) expression data (network 4), 14 changed sign when the degradation rates were changed to Neymotin et al. (2014) values and replicate Schade et al. (2004) expression data was added (network 5), and 16 edges changed sign when the Dahlquist lab wt-only expression data was used (network 6). Unexpectedly, reformatting the input file to support GRNmap v1.10 greatly impacted the edge weight signs.

Another comparison between the network created when the file was reformatted to GRNmap v1.10 and the rest of the networks (2-6). Compared to the reformatted network (Figure 20, network 1), changing the optimization parameters changed only 4 of the edge weight signs, which was expected based on the previous runs where the optimization parameters were manipulated (Section 4). Further, compared to the same reformatted network (network 1), 5 edges changed sign when the Schade et al. (2004) replicate expression data was added (network 3), 6 edges changed sign when the replicate expression data was added and the degradation rates were changed to Neymotin et al. (2014) (network 4), the same 6 edges changed no replicate expression data was added and the degradation rates were changed to Neymotin et al. (2014) values (network 5), and 17 edges changed sign when only the Dahlquist wt-only expression data were used (network 6).

It was not expected that changing the format of the input file to support GRNmap v1.10 would cause almost half of the edges to change signs. However, changing the optimization parameters did not have as much of an effect, which was expected. Changing the data from Schade et al. (2004) expression data to Dahlquist data and also changing the degradation rates caused more edges to change sign, which was again expected because the data were from experiments with different temperatures, strains, and time points. Therefore, the differences in experimental setup impacts the similarities between the weight values, which was reflected in the edges changing signs.





**Figure 20:** GRNsight visualizations of the model runs from Table 8. 21-gene, 31-edge model runs where: demo network (0), reformatted to GRNmap v1.10 (1), optimization parameters changed (2), Schade et al. (2004) replicate data added (3), Neymotin et al. (2014) degradation rates with no replicate data (4), Neymotin et al. (2014) degradation rates with Schade et al. (2004) replicate data (5), and Dahlquist wt-only expression data with Neymotin et al. (2014) degradation rates (6). For the edges, a blue edge represents repression and a red edge indicates activation. For the nodes, the same color pattern is used to denote activation and repression. The node is divided into the timepoints of the experiment going from left to right: t15, t30, t60, t90, and t120.

One edge that consistently changed sign is the regulatory relationship between Rap1 and Rph1, which changed from activation to repression for all of the networks, except for the network using the Dahlquist lab expression data. Another edge, YAP6 → CIN5, consistently showed activation for four of the five new networks, which is different from the published model results (Dahlquist et al. 2015). For the network using the Dahlquist lab data, however, the edge weight became a strong repression. This edge is involved in a complex feedforward loop with ROX1, so it is not surprising that the weight values would be sensitive to changes in input data (both degradation rates and expression data).

From these analyses, it can be determined that certain regulatory relationships in the 21-gene, 31-edge network are sensitive to changes in the input data. Changing the optimization parameters, adding replicate expression data, or changing the degradation rates greatly impacted the network, changing the sign or magnitude of the edge weight values.

**Conclusion:**

*Saccharomyces cerevisiae* must respond to temperature downshifts through a global transcriptional response in order to maintain homeostasis and continue their proliferation and growth. Using wild type and five transcription factor deletion strains, the global transcriptional response of yeast to cold shock was assessed in this study. By clustering the microarray data, through both STEM and *k-*means clustering methods, the Gene Ontology terms associated with genes that share similar expression patterns were determined. Furthermore, the regulation and dynamics of the cold shock response can be modeled using GRNmap. Using ordinary differential equations, GRNmap performs estimations of the regulatory weight, production rate, and gene expression threshold parameters. The model runs can be compared to determine which conditions improve the fit of the model.

In this study, multiple sensitivity experiments were performed on the model to determine how manipulations to the model input, such as changing optimization parameters and production rates, deleting edges or nodes of the candidate GRN, or variable inclusion of strain expression data, impact the fit of the model. Through these assessments, we determined that the model is not sensitive to changing the optimization parameters, but is sensitive to fixing the production rate values. Changing the structure of the GRN or including different strains’ expression data also impacted the model. Certain edges of the candidate GRN were determined to be important in the cold shock response in yeast, such as those involving Msn2, Cin5, or Hmo1. In addition, through deleting the Gcr2 and Zap1 nodes from the network, the fit of the model increased indicating that the model is sensitive to changes in those nodes. Furthermore, through the variable inclusion of strain data, we determined that the inclusion of ∆*gln3* data decreased the fit of the model, indicating that Gln3 plays an important role in the cold shock response. These results indicate that the model is sensitive to changes in GRN structure, strain data, and fixing production rates; but is not sensitive to changes in the optimization parameters.

**Appendix:**

<https://github.com/kdahlquist/DahlquistLab/blob/master/documents/theses/Alice%20Finton%202020/AF_Thesis_Appendix.docx>

**Acknowledgments**

I would like to thank Lauren Kelly for her work in running the models for the edge-deletion experiment, the GRNmap lab alumni (Brandon J. Klein and Lauren Kelly), the current Dahlquist lab researchers (Alyssa Weisblatt and Susanne Carpenter), Dr. Kam Dahlquist, Dr. Ben Fitzpatrick, and the LMU Biology Department.

**References:**

Al-Fageeh MB, Smales CM. (2006) Control and regulation of the cellular responses to cold shock: the responses in yeast and mammalian systems. *Biochem J*. 2006;397(2):247–259. doi:10.1042/BJ20060166

Albert, B., Kos-Braun, I. C., Henras, A. K., Dez, C., Rueda, M. P., Zhang, X., ... & Shore, D. (2019). A ribosome assembly stress response regulates transcription to maintain proteome homeostasis. *Elife*, *8*, e45002. DOI: 10.7554/eLife.45002

ASH1 (n.d.). Retrieved on February 28, 2020 from <https://www.yeastgenome.org/locus/YKL185W>

Belle, A., Tanay, A., Bitincka, L., Shamir, R., & O’Shea, E. K. (2006). Quantification of protein half-lives in the budding yeast proteome. *Proceedings of the National Academy of Sciences*, *103*(35), 13004-13009. DOI: 0.1073/pnas.0605420103

Berg JM, Tymoczko JL, Stryer L. Biochemistry. 5th edition. New York: W H Freeman; 2002. Section 16.2, The Glycolytic Pathway Is Tightly Controlled. Available from: https://www.ncbi.nlm.nih.gov/books/NBK22395/

Berger, A. B., Decourty, L., Badis, G., Nehrbass, U., Jacquier, A., & Gadal, O. (2007). Hmo1 is required for TOR-dependent regulation of ribosomal protein gene transcription. *Molecular and cellular biology*, *27*(22), 8015-8026. DOI: 10.1128/MCB.01102-07

Chen, T., Li, F., & Chen, B. S. (2009). Cross-talks of sensory transcription networks in response to various environmental stresses. *Interdisciplinary sciences: computational life sciences*, *1*(1), 46-54. DOI: 10.1007/s12539-008-0018-1.

CIN5 (n.d.). Retrieved on February 28, 2020 from <https://www.yeastgenome.org/locus/S000005554>

Dahlquist, K. D., Dionisio, J. D. N., Fitzpatrick, B. G., Anguiano, N. A., Varshneya, A., Southwick, B. J., & Samdarshi, M. (2016). GRNsight: a web application and service for visualizing models of small-to medium-scale gene regulatory networks. *PeerJ Computer Science*, *2*, e85. DOI: 10.7287/peerj.preprints.2068v1

Dahlquist, K. D., Fitzpatrick, B. G., Camacho, E. T., Entzminger, S. D., & Wanner, N. C. (2015). Parameter estimation for gene regulatory networks from microarray data: cold shock response in *Saccharomyces cerevisiae*. *Bulletin of mathematical biology*, *77*(8), 1457-1492. DOI: 10.1007/s11538-015-0092-6

Dahlquist, K., Fitzpatrick, B., & Dionisio, J. (2018, February 2). *GRNmap - Documentation*. GitHub Pages. <https://kdahlquist.github.io/GRNmap/documentation/>

Dahlquist, K. (2018, June 22). *GEO accession viewer*. National Center for Biotechnology Information. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE83656>

Ernst, J., Nau, G. J., & Bar-Joseph, Z. (2005). Clustering short time series gene expression data. *Bioinformatics*, *21*(suppl\_1), i159-i168. DOI: 10.1093/bioinformatics/bti1022

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

Fernandes, L., Rodrigues-Pousada, C., & Struhl, K. (1997). Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. *Molecular and cellular biology*, *17*(12), 6982-6993. Doi: 10.1128/mcb.17.12.6982

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., ... & Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. *Molecular biology of the cell*, *11*(12), 4241-4257. DOI: 10.1091/mbc.11.12.4241

Gasch, A. P., & Werner-Washburne, M. (2002). The genomics of yeast responses to

environmental stress and starvation. Functional & integrative genomics, 2(4-5), 181-192. DOI: 10.1007/s10142-002-0058-2

GCR2 (n.d.). Retrieved on February 28, 2020 from

<https://www.yeastgenome.org/locus/S000005143>

Gene Ontology Consortium. (2015). Gene ontology consortium: going forward. *Nucleic acids*

*research*, *43*(D1), D1049-D1056. DOI: 10.1093/nar/gku1179

GLN3. (n.d.). Retrieved on February 28, 2020 from

<https://www.yeastgenome.org/locus/S000000842>

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., ... & Jennings, E. G. (2004). Transcriptional regulatory code of a eukaryotic genome. *Nature*, *431*(7004), 99-104. DOI: 10.1038/nature02800

HMO1 (n.d.). Retrieved on February 28, 2020 from <https://www.yeastgenome.org/locus/S000002581>

Kelly, L. M. (2019). *Modeling of Gene Regulatory Network Dynamics Predicts which Regulatory Relationships are Important for Controlling the Cold Shock Response in Saccharomyces cerevisiae* (Unpublished thesis). Loyola Marymount University, Los Angeles, CA.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., ... & Zeitlinger, J. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, *298*(5594), 799-804. DOI: 10.1126/science.1075090

Maere S, et al. (2005) ClueGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. *Bioinformatics*. 21:3448–3449. DOI: 10.1093/bioinformatics/btp101

Mager, W. H., & De Kruijff, A. J. (1995). Stress-induced transcriptional activation. *Microbiol.*

*Mol. Biol. Rev.*, *59*(3), 506-531.

Monteiro, P. T., Oliveira, J., Pais, P., Antunes, M., Palma, M., Cavalheiro, M., ... & Mota, M. N. (2020). YEASTRACT+: a portal for cross-species comparative genomics of transcription regulation in yeasts. *Nucleic acids research*, 48(D1), D642-D649. [doi:10.1093/nar/gkz859](https://doi.org/10.1093/nar/gkz859)

Murata, Y., Homma, T., Kitagawa, E. et al. Extremophiles: Genome-wide expression analysis of yeast response during exposure to 4C (2006) 10: 117. doi:10.1007/s00792-005-0480-1.

Neymotin, B., Athanasiadou,R., & Gresham, D. (2014). Determination of in vivo RNA kinetics using RATE-seq. *RNA*, 20(10), 1645-1652. doi:10.1261/rna

Sahara, T., Goda, T., & Ohgiya, S. (2002). Comprehensive expression analysis of time-dependent genetic responses in yeast cells to low temperature. *Journal of Biological Chemistry*, *277*(51), 50015-50021. DOI: 10.1074/jbc.M209258200

Salvadó, Z., Arroyo-López, F. N., Guillamón, J. M., Salazar, G., Querol, A., & Barrio, E. (2011).

Temperature adaptation markedly determines evolution within the genus Saccharomyces. *Appl. Environ. Microbiol*., 77(7), 2292-2302. DOI: 10.1128/AEM.01861-10

Schade, B., Jansen, G., Whiteway, M., Entian, K. D., & Thomas, D. Y. (2004). Cold adaptation in budding yeast. *Molecular Biology of the Cell*, *15*(12), 5492-5502. DOI: 10.1091/mbc.E04-03-0167

Schena, M. (1996). Genome analysis with gene expression microarrays. Bioessays, 18(5), 427

431. DOI: 10.1002/bies.950180513

Tai, S. L., Daran-Lapujade, P., Walsh, M. C., Pronk, J. T., & Daran, J. M. (2007). Acclimation of

Saccharomyces cerevisiae to low temperature: a chemostat-based transcriptome analysis. *Molecular Biology of the Cell*, 18(12), 5100-5112. DOI: 10.1091/mbc.E07-02-0131

Tate J. J., & Cooper T. G. (2007). Stress-responsive Gln3 localization in *Saccharomyces cerevisiae* is separable from and can overwhelm nitrogen source regulation. *J. Biol. Chem*. DOI: 18467–18480.

UniProt (2019). AP-1-like transcription factor YAP4. Retrieved on February 28, 2020 from https://www.uniprot.org/uniprot/P40917

Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (2001, June). Constrained k-means clustering with background knowledge. In Icml (Vol. 1, pp. 577-584).

ZAP1 (n.d). Retrieved on February 28, 2020 from

<https://www.yeastgenome.org/locus/S000003592>