Modeling the Dynamics of a 21-Gene, 50-edge Gene Regulatory Network Controlling the Transcriptional Response to Cold Shock in *Saccharomyces cerevisiae* using GRNmap

K. Grace Johnson[1], Natalie E. Williams[2], Kam D. Dahlquist[2], and Ben G. Fitzpatrick[3]

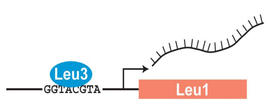
[1]Department of Chemistry and Biochemistry, [2]Department of Biology, [3]Department of Mathematics

Loyola Marymount University, 1 LMU Drive, Los Angeles, CA 90045 USA

LMU|LA Frank R. Seaver College of Science and Engineering

University Honors Program

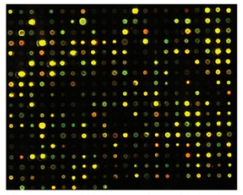
Transcription Factors Control Gene Expression by Binding to Regulatory DNA Sequences Upstream of Genes



- Activators increase gene expression.
- Repressors decrease gene expression.
- Transcription factors are themselves proteins that are encoded by genes.
- A gene regulatory network (GRN) consists of a set of transcription factors that regulate the level of expression of a set of target genes, which can include other transcription factors.
- The dynamics of a GRN is how the expression of genes in the network change over time.

Yeast Respond to the Environmental Stress of Cold Shock by Changing Gene Expression

- Little is known about which transcription factors regulate this response.
- The Dahlquist Lab studies the global transcriptional response to cold shock using DNA microarrays, which measure the level of mRNA expression for all 6000 yeast genes.
- We have collected expression data from the wild type strain and five transcription factor deletion strains (*Δcin5, Δgln3, Δhmo1, Δzap1, Δhap4*) before cold shock at 30°C and after 15, 30, and 60 minutes of cold shock at 13°C.
- The Dahlquist Lab has shown that yeast deleted for the Hap4 transcription factor, a heme activator protein, show impaired growth at cold temperatures, implying that it is important for regulating the response to cold shock.
- We use mathematical modeling to determine the relative influence of each transcription factor in the GRN that controls the cold shock response.



Microarray at 60 minutes after cold shock

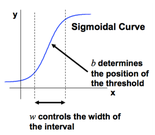
The *Δhap4* Strain Microarray Data Was Used to Derive a Family of Related GRNs from the YEASTRACT Database

- An ANOVA test of the *Δhap4* strain DNA microarray data showed that 1794 genes (29%) had a $\log_2$ fold change significantly different than zero at any of the time points, with a Benjamini & Hochberg corrected p value < 0.05.
- These genes were submitted to the YEASTRACT database, which returned a list of candidate regulatory transcription factors that potentially regulate those target genes, in order of significance.
- The transcription factors for which we had deletion strain microarray data were added to the list of the 29 most significant regulators to generate the largest GRN we modeled with a total of 34 genes and 102 edges. Transcription factors and edges were removed from the GRN in a stepwise fashion in order of least to most significant until the network was pared down to 15 genes and 28 edges.
- The purpose of comparing a family of related networks is to determine which sized network models the experimental data best, accounting for indirect effects of other regulatory transcription factors upon cold shock gene expression.

For Each Gene in the Network, a Nonlinear Differential Equation Determines the Rate At Which the Gene is Expressed

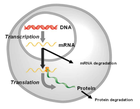
- The model, called GRNmap (Gene Regulatory Network modeling and parameter estimation) was implemented in MATLAB (Dahlquist et al. 2015).
- The MATLAB code and executable are available under an open source license at https://github.com/kdahlquist/GRNmap/.
- Each gene has a differential equation that models the change in expression over time as production – degradation
- Degradation rates for each gene were taken from protein half life data from Belle et al. (2006)
- We use a sigmoidal production function where:
  - P_i is mRNA production rate for gene i
  - d_i is the mRNA degradation rate for gene i
  - w is weight term, determining the level of activation or repression of j on i
  - b is a unique threshold for each gene
- The production rate (P_i), weight (w), and threshold (b) values were estimated from DNA microarray data using a penalized least squares approach.

$$\frac{dx_i(t)}{dt} = \frac{P_i}{1+\exp\left(-\sum_j \left(w_{ij}x_j(t)-b_j\right)\right)} - d_i x_i(t)$$



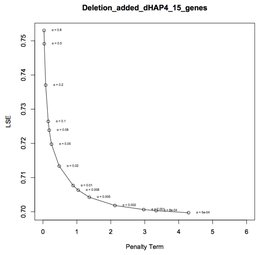
$$E = \alpha\|\theta\|^2 + \frac{1}{Q}\sum_{r=1}^{Q}\left[z^d(t_r) - z^c(t_r)\right]^2$$

- E represents the error between estimated values and microarray data values.
- θ is the penalty term, which is the combined w, P, and b parameter values.



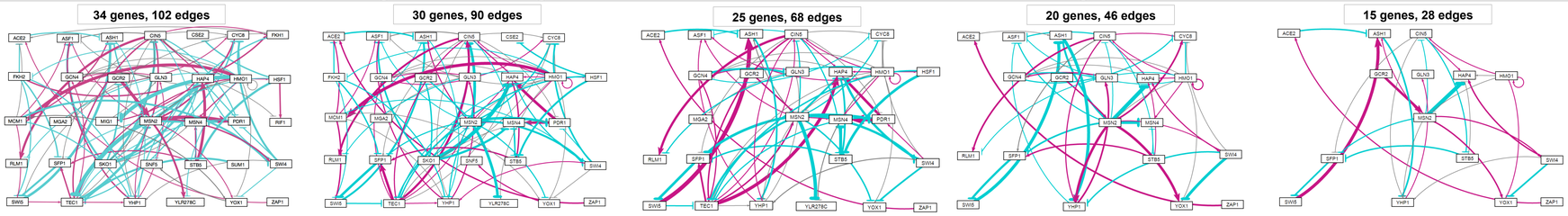
(Freeman, 2002)

L-curve Analysis of the Smallest GRN Suggests a Good alpha Value to be 0.002



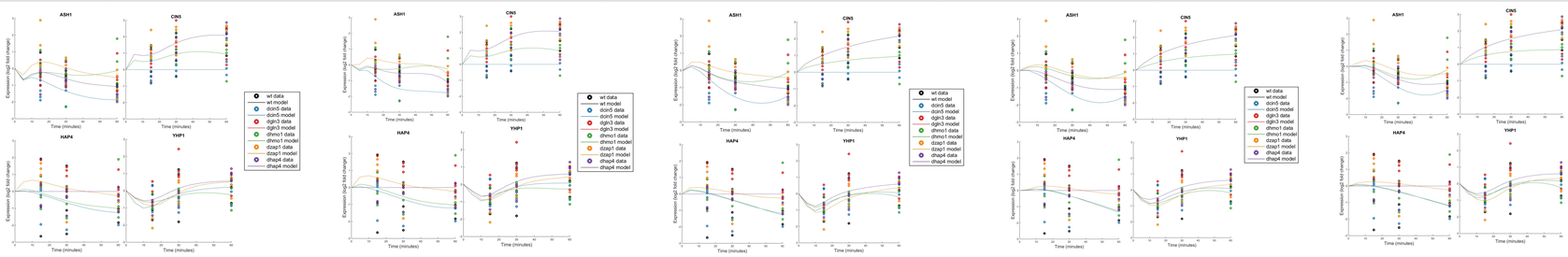
- The alpha value (α) controls the flexibility of the model fit to the data.
- Choosing the best alpha value is best done through iteration.
- The estimation was run iteratively for a series of different alpha values ranging from 0.8 down to 0.0005 where the parameters output from one run was used as the initial guesses for the next run.
- For each alpha value ranging from 0.0005 to 0.8, the Least Squares Error (LSE) was plotted against the penalty term.
- The best alpha is one that minimizes both the LSE and the penalty term, and therefore lies near the "elbow" of the L-curve.

Visualization via GRNsight Reveals Changes in Activation/Repression Relationships as Nodes and Edges Are Removed from the Network



- GRNsight automatically generates weighted network graphs from the output spreadsheets produced by GRNmap.
- The absolute value of the weight parameters are divided by the largest value, which distributes them between 0 and 1. The thickness of the lines is on a linear scale with thin lines for values near 0 and thick lines for values near 1.
- Positive weights are colored magenta to indicate activation, negative weights are colored cyan to indicate repression.
- Weights within ±0.05 of zero are colored grey to denote negligible influence on the target gene.

GRNmap Reveals YHP1 is Modeled Well in All Five Networks, ASH1 and CIN5 are Modeled Best by the Smaller Networks, and HAP4 is Not Modeled Well in Any Network

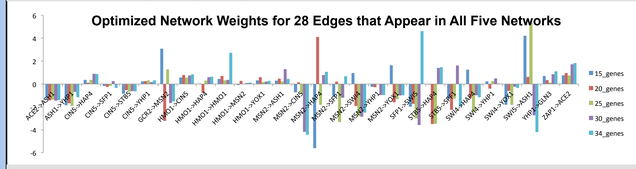


- Individual plots of each gene's expression compare experimental data for each strain (circles) to simulated data based on solving the differential equation with the estimated parameters (lines).
- ASH1, CIN5, and YHP1 were chosen for further examination because they exhibited interesting, significant dynamics ($\log_2$ fold change) across all five networks. HAP4 was chosen because the networks modeled were based on data from its deletion strain.
- YHP1 is modeled well in all five networks. For each strain's data points, the color coded model output easily follows the trend of the data points.
- ASH1 and CIN5 are more realistically modeled by smaller networks. In larger networks, the simulations exhibit extreme dynamics that are likely not biologically relevant. This may be due to using an alpha value that is too small for the larger networks. This warrants further investigation.
- While ASH1, CIN5, HAP4, and YHP1 occur in all five networks, their connectivity to other genes changes as the network is pared down. This may affect their performance in the model.
- As part of the GRNmap output, mean squared errors (MSE) were reported for each gene and each deletion dataset (color coded in the expression plots above).
- For the genes ASH1, CIN5, HAP4, and YHP1, the MSE's were summed for all datasets and reported along with their B&H corrected p values.
- The sum MSE's for ASH1, CIN5, and HAP4 decrease as the network size increases, but at the cost of unrealistic flexibility in gene dynamics revealed in the individual gene plots of the larger networks.
- The sum MSE's for HAP4 are consistently higher than other genes for all five networks. This, however, is due to uncertainty in the data rather than model shortcomings.

Mean Square Errors are Smaller for Larger Networks

Gene	ANOVA B&H p-values	Sum MSE				
		Network size				
		34 genes	30 genes	25 genes	20 genes	15 genes
ASH1	0.0686	3.7168	3.7225	3.8400	3.9071	4.0256
CIN5	0.0100	4.2400	4.2634	4.3219	4.3167	4.3261
HAP4	0.4090	7.1153	7.1152	7.4471	7.5679	7.5163
YHP1	0.0074	3.3188	3.2880	3.2791	3.2770	3.2684

Comparing Output Estimated Parameters Reveals Networks are Modeled Differently Based on Size



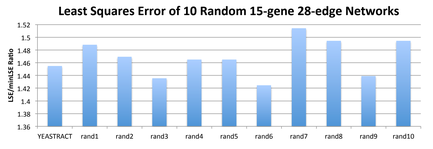
- Drastic changes in sign, signifying a switch from activation to repression or vice versa, occurs with four edges across the five networks.
- The change in sign for some connections may have resulted from a deletion in edges or nodes between networks.
- Both MSN2 and SWI5 exhibit frequent, extreme sign changes as both regulators and targets. This warrants further investigation.

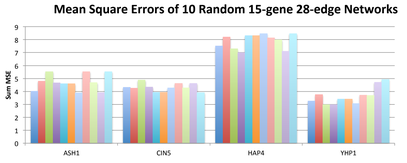
LSE Reveals the Model Performs Consistently For a Range of Network Sizes

- Least squares error (LSE) represents the total error between the model outputs and data points for all five networks. The larger the LSE, the more difficult it was to fit the model to the data.
- The minimum LSE would be the best theoretical model fit for each network based on the average of the data.
- The ratio is the LSE divided by the minimum theoretical LSE and shows how close the LSE is to the ideal minimum LSE.
- As the ratios were relatively similar for all five networks, this indicates the model performs consistently for a range of network sizes.

Network size	34 genes, 102 edges	30 genes, 90 edges	25 genes, 68 edges	20 genes, 46 edges	15 genes, 28 edges
Parameters	170	150	118	86	58
LSE	0.7932	0.7524	0.7048	0.6876	0.7056
Minimum theoretical LSE	0.5467	0.5331	0.4898	0.4776	0.4850
Ratio	1.4510	1.4113	1.4388	1.4397	1.4549

Comparison to 10 Random 15-gene 28-edge Networks Reveals Database-Derived Network is Modeled Well





- Random networks to compare to the smallest YEASTRACT-derived network were created from the same 15 genes, but with 28 random connections.
- Random networks allow for a control against which to compare the performance of the YEASTRACT-derived network.
- The ratios of Least Squares Error to the minimum Least Squares Error were generally higher for the random networks than the database-derived network.
- For the four genes in the network analyzed above, the sum Mean Square Errors were larger for most random networks.
- Both trends indicate that the database-derived network is modeled better by GRNmap than the random networks, although further analysis is needed.

Conclusions and Future Directions

- Expression plots of ASH1, CIN5, HAP4, and YHP1 showed how, in general, the model fit the smaller networks better than the larger ones. This is in direct contrast to the MSE's, which are smaller for the larger networks. Expression plots and MSE's reveal that this inconsistency is likely due to an improper choice in alpha for the larger networks. In the future, a larger alpha should be chosen for the larger networks to dampen the flexibility (seen in ASH1 and CIN5 plots for the larger networks) that is likely biologically irrelevant.
- The LSE and the ratio of output LSE to theoretical minimum LSE for each networks demonstrated that the model works consistently for this range of network sizes.
- Estimated parameter comparisons showed how the parameters can change with node/edge deletion between networks. Extreme fluctuation in estimated parameter outputs was especially frequent with genes shown in the individual gene plots to be modeled poorly, namely ASH1 and CIN5.
- The 15-gene 28-edge network derived from the YEASTRACT database was compared to 10 random networks of the same genes and same number of edges connected randomly. Comparison of output LSE and MSE for four genes revealed that, in general, the database-derived network was modeled better than the random networks.
- In addition to the above, future directions include comparisons of this family of networks to more randomly generated networks with the same nodes and same number of edges. These will provide a control that will allow us to be more confident in the model interpretation and validity of our hypothesis networks.

Acknowledgments

For their work on the GRNmap code, we would like to thank Juan S. Carrillo, Trixie Anne M. Roque, Chukwuemeka E. Azinge, Katrina Sherbina, and Nicholas A. Rohacz. We thank Nicole A. Anguiano, Anindita Varshneya, Mihir Samdarshi, and Britain J. Southwick for their work on the GRNsight visualization software. Microarray data were collected by Cybele Arsan, Wesley Citti, Kevin Entzminger, Andrew Herman, Monica Hong, Heather King, Lauren Kubeck, Stephanie Kuelbs, Elizabeth Liu, Matthew Mejia, Kevin McGee, Kenny Rodriguez, Olivia Sakhon, Alondra Vega, and Kevin Wyllie. This work was supported by NSF-RUI Award 0921038 (K.D.D., B.G.F.), a Kadner-Pitts Research Grant (K.D.D., K.G.J.), and a Loyola Marymount University Honors Program Summer Fellowship (K.G.J.).

References

Abdulrehman, D., Monteiro, P., Teixeira, M., Mira, N., Lourenço, A., Santos, S., Cabrito, T., Francisco, A., Madeira, S., Aires, R., Oliveira, A., Sá-Correia, I., & Freitas, A. (2011). YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in *Saccharomyces cerevisiae* through a web services interface Nucl. Acids Res., 39: D136-D140, Oxford University Press.

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.

Dahlquist, K., Fitzpatrick, B., Camacho, E., Entzminger, S., & Wanner, N. (2015). Parameter Estimation for Gene Regulatory Networks from Microarray Data: Cold Shock Response in Saccharomyces cerevisiae. *Bulletin Of Mathematical Biology*, *77*(8), 1457-1492. http://dx.doi.org/10.1007/s11538-015-0092-6.

Freeman, S. (2002). *Biological science* (First ed.). Prentice Hall.

GRNsight - Home. (n.d.). Retrieved March 10, 2016, from http://dondi.github.io/GRNsight/.

GRNmap – Home.. (n.d.). Retrieved March 10, 2016, from https://github.com/kdahlquist/GRNmap