Dynamical systems modeling and gene regulatory network structure analysis reveals Hap4's role in regulating the response to cold shock in *Saccharomyces cerevisiae*

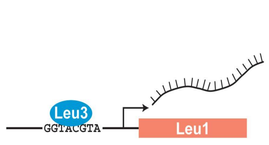
LMU|LA Frank R. Seaver College of Science and Engineering

Kristen M. Horstmann[1,2], Margaret J. O'Neil[1], Ben G. Fitzpatrick[2], and Kam D. Dahlquist[1]

[1]Department of Biology, [2]Department of Mathematics

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

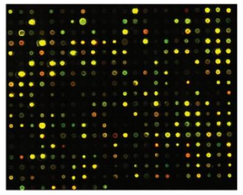
Transcription factors control gene expression by binding to regulatory DNA sequences upstream of genes



- 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

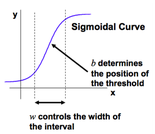
The *Δhap4* strain microarray data was used to derive a family of related GRNs from the YEASTRACT database

- 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

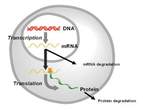
- The model, called GRNmap (Gene Regulatory Network modeling and parameter estimation) was implemented in MATLAB (Dahlquist et al. 2015).
- The MATLAB code and executable are available under an open source license at https://github.com/kdahlquist/GRNmap/.
- Each gene has a differential equation that 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.

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



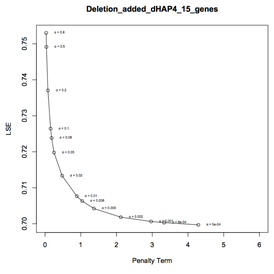
$$E = \alpha\|\theta\|^2 + \frac{1}{Q}\sum_{r=1}^{Q}\left[z^d(t_r) - z^c(t_r)\right]^2$$

- 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.



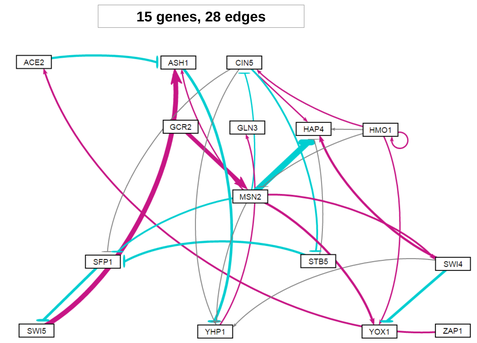
(Freeman, 2002)

L-curve analysis suggests a good alpha value to be 0.002

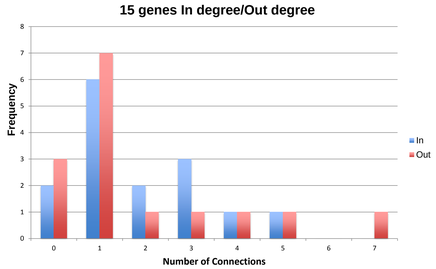


- The alpha value (α) controls the flexibility of the model fit to the data.
- Choosing the best alpha value is best done through iteration.
- The estimation was run iteratively for a series of different alpha values ranging from 0.8 down to 0.0005 where the parameters output from one run was used as 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.

Network derived from *Δhap4* data can be visualized using GRNsight



- GRNsight generates weighted network graphs using 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.
- Arrow heads also represent activation, while blunted heads indicate repression.
- Weights within ±0.05 of zero are colored grey to denote negligible influence on the target gene.



In Degree, Out Degree, and Total Degrees for all 15 Genes in the GRN			
	In-Degree	Out-Degree	Degree
ACE2	1	1	2
ASH1	3	1	4
CIN5	2	4	6
GCR2	0	1	1
GLN3	1	0	1
HAP4	5	0	5
HMO1	1	5	6
MSN2	2	7	9
SFP1	3	1	4
STB5	1	2	3
SWI4	1	3	4
SWI5	1	1	2
YHP1	4	1	5
YOX1	3	0	3
ZAP1	0	1	1

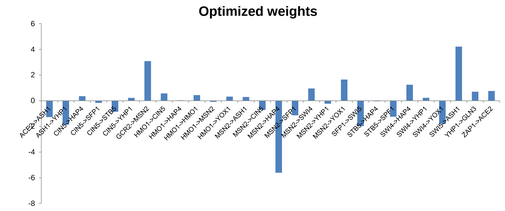
- In- and out-degree distributions were manually plotted. These plots show how many genes are connected to other genes in the network as source (out) or by target (in).
- It makes sense that the fewest number of connections would occur most frequently, as the majority of the transcription factors have a small number of connections to genes in the rest of the network.
- The statistical program Gephi was able to calculate the exact connections of in-degree, out-degree, and total degree for each gene in the GRN.
- From the GRNsight visualized network as well as the Gephi outputs, MSN2, CIN5, HMO1, HAP4, and YHP1 are some of the most active genes, with having total degree of 9, 6, and 5, respectively.
- This level of connectivity can be examined in more detail with further statistical tests.

The fit of the model parameters is close to the minimum theoretical least squares error

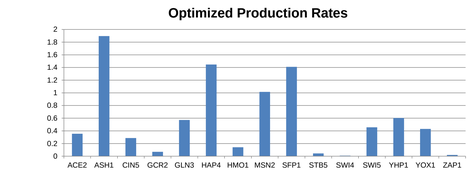
Network	Parameters	LSE	Minimum theoretical LSE	Ratio
15 genes, 28 edges	58	0.706	0.485	1.455

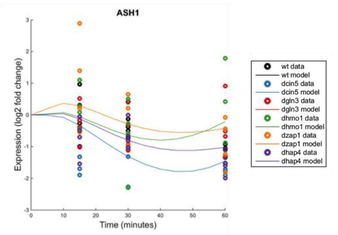
- Least squares error (LSE) represents the total error between the model outputs and data points for all five networks. A large LSE represents difficulty with the model fitting the data.
- The minimum theoretical LSE represents the ideal theoretical model fit for each network based on the average of 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 LSE is fairly small, the model fit well, but did not fit the exact minimum theoretical LSE, as the ratio is larger than 1. So the model had more errors than the ideal theoretical model, but is still considered a good fit on the network.

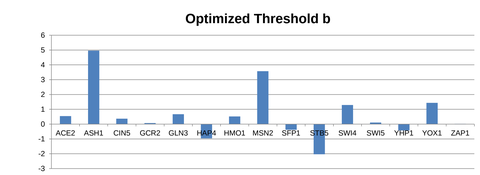
The individual parameters reveal details about the behavior of individual genes



- Upper left. The 28 edges in the network, with the source node listed first, and the target node listed second.
- Negative weight values signify repression, whereas positive values signify activation.
- MSN2 -> HAP4 is the strongest relationship within the network, with a repression weight of nearly 6. This can also be visually confirmed in the GRNsight network.
- SWI5 -> ASH1 is the largest activation relationship, with an estimated weight of 4.







- ASH1 had the highest production rate and threshold b for this network, possibly due to the high activation of it by SWI5.

Gephi was used to compute the graph properties of the *Δhap4* data-derived network

- The eccentricity centrality of a network shows how easily accessible a node is from other nodes (Pavlopoulos et al. 2011)
- The eccentricity is calculated using an algorithm for identifying the *max{dist(i,j)}* where *i* is the node listed in the table and *j* is any other node in the network.
- Eccentricity centrality is a directional statistic, which only takes a node's out degree into account
- To have a high eccentricity centrality means that the gene is connected indirectly to many other genes in the network. This indicates these genes with high eccentricities have a greater impact on other nodes than a node with low eccentricity.

	ACE2	ASH1	CIN5	GCR2	GLN3	HAP4	HMO1	MSN2	SFP1	STB5	SWI4	SWI5	YHP1	YOX1	ZAP1
Eccentricity Centrality	3	2	3	3	0	0	3	2	4	5	2	3	1	0	4

- Closeness Centrality is a centrality measure that indicates how long it will take for information from a node *x* to reach other nodes in the network. (McSweeney, 2009)
- The closeness centrality of a node can be determined using the following formula:

$$C(x) = \frac{1}{\sum_y d(x,y)}(n-1)$$

which is looking at the average shortest path from *x* to all other nodes

- Closeness centrality, like eccentricity centrality is a directional statistic, which only takes a node's out degrees into account.
- Nodes with high closeness centralities are the most central nodes in the network. These are the nodes that can communicate the quickest in the network.
- The directional aspect of the closeness centrality measure means those genes and nodes with no out-degree connections have a closeness centrality of 0.

	ACE2	ASH1	CIN5	GCR2	GLN3	HAP4	HMO1	MSN2	SFP1	STB5	SWI4	SWI5	YHP1	YOX1	ZAP1
Closeness Centrality	0.5	0.667	0.636	0.458	0	0	0.55	0.769	0.4	0.375	0.8	0.5	1	0	0.4

- Betweenness Centrality is a centrality measure that indicates how often a node is found on a shortest path between two nodes, s and t. (McSweeney, 2009.)
- The betweenness centrality of a node can be calculated by the following:

$$g(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

where $\sigma_{st}(v)$ is the number of shortest paths that pass from s to t and σ_{st} is the total number of shortest paths from node s to node t

- Without these "between" nodes, there would be no way for two separated nodes to communicate.
- Betweenness centrality requires there to be at least one in degree and one out degree in order to be calculated, so in cases where there is either no in degree or no out degree, the betweenness centrality of those nodes is 0.
- Nodes with high betweenness centralities are key genes associated with regulating other genes in the network, and are known as "bottlenecks" that must be passed through in the network.

	ACE2	ASH1	CIN5	GCR2	GLN3	HAP4	HMO1	MSN2	SFP1	STB5	SWI4	SWI5	YHP1	YOX1	ZAP1
Betweenness Centrality	3	10	5	0	0	0	0	14	9	0	0	7	11	0	0

Conclusions and Future Directions

- DNA microarray data from the *Δhap4* deletion strain subjected to cold shock was analyzed using an ANOVA test, the YEASTRACT database, and an ordinary differential equations model called GRNmap that modeled the dynamics of each gene in candidate gene regulatory networks. From larger networks, the 15 gene, 28 edges network was determined to be the best candidate for data analysis.
- The weighted output network was visualized using GRNsight.
- The Gephi results are consistent with the in-degree, out-degree statistics, where the genes with the highest degree and overall degree measures are also found to have the highest betweenness centrality measures, and those nodes with the lowest degree measures also have the lowest betweenness centrality. The statistics from Gephi provided useful information through which to view the graph. While MSN2 has the highest betweenness centrality and the highest degree measure, it only has the second highest closeness centrality measure, which indicates that while it is a very important node in the graph, SWI4 is more centralized in the graph.
- The LSE and the ratio of output LSE to theoretical minimum LSE for the network demonstrated that the model has more errors than a theoretically ideal run. However, this is to be expected for any model run, and the ratio demonstrates a close fit as it was only slightly above 1.
- ASH1 had the strongest activation input in the network, from SWI5. This may have affected the size of the production rate and optimized threshold b levels.
- In addition to the above, future directions include running Gephi statistical analysis on the other gene family networks. Then, comparisons of the *Δhap4* network statistics to the other deletion gene networks could be done. It would also be interesting to run Gephi analysis on networks of larger size in order to see how the centrality of nodes and connections change with the deletion of important nodes and edges.

Acknowledgments

For their work on the GRNmap code, we would like to thank Trixie Anne M. Roque, Chukwuemeka E. Azinge, and Justin K. Torres. We thank Nicole A. Anguiano, Anindita Varshneya, Mihir Samdarshi, Edward Bachuora, Jen Shin, and Eileen Choe 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. Further, we would like to thank Natalie E. Williams and Brandon J. Klein for their contributions to the GRNmap data analysis team. This work is partially supported by NSF award 0921038 (K.D.D., B.G.F.) and a Kadner-Pitts Research Grant (K.D.D.).

References

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://doi.org/10.1007/s11538-015-0092-6.

Dário Abdulrehman, Pedro T. Monteiro, Miguel C. Teixeira, Nuno P. Mira, Artur B. Lourenço, Sandra C. dos Santos, Tânia R. Cabrito, Alexandre P. Francisco, Sara C. Madeira, Ricardo S. Aires, Arlindo L. Oliveira, Isabel Sá-Correia, Ana T. Freitas (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.

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

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

Gephi – Home. (n.d.). Retrieved November 15, 2016, from https://gephi.org/.

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

Ma, H. W., & Zeng, A. P. (2003). The connectivity structure, giant strong component and centrality of metabolic networks. Bioinformatics, 19(11), 1423-1430.

McSweeney, Patrick J. (2009). Gephi Network Statistics: Google Summer of Code 2009 Project Proposal. http://web.ecs.syr.edu/~pjmcswee/gephi.pdf

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., ... & Bagos, P. G. (2011). Using graph theory to analyze biological networks. BioData mining, 4(1), 10.