Asha Dorsey and Khayla McCoy

December 9, 2012

Dr. Qin

Final Project Submission

Gene Expression Correlations with Data Sets Succumbing to Heat Stress and Oxidative Stress

**Abstract**

In recent studies, the genomic expression patterns in the yeast *Saccharomyces cerevisiae* responding to diverse environmental to heat shocks and oxidative stress were analyzed. Physiological themes in the genomic responses to speciﬁc environmental stresses provided insights into the effects of those stresses in the yeast gene1. Scientific impacts are attributed to the importance of gene networks, cellular aging and robustness in many other biological processes analyzed growth fitness measures in various conditions and morphological robustness In this experiment the standard deviation (or coefficient of variation) for a gene expression data set was calculated to estimate the robustness effect of every yeast gene2. The correlation among the expression robustness to RLS, fitness, protein interaction network, genetic interaction network, evolution was also examined. Then permutation was a technique was used to evaluate distribution of CV (or stddev) in protein networks and genetic networks3. A statistic measure of lifespan of each module, mean, median, or standard deviation, was evaluated based on null distributions generated by permutation of lifespans. Missing lifespan values will be treated as ‘NA’s during analysis and permutations.

**Introduction**

The topic of cellular aging still holds great ambiguity within the world of biology. Several models, including the Gompertz model, have made attempts to accurately define cellular aging. The most commonly used and seemingly accurate definition of aging currently states that aging is an exponential increase of mortality rate over time2. An organism commonly used to study this idea of aging in the lab is yeast cells (*Saccharomyces cerevisiae).* Yeast cells have the ability to live to different ages despite each cell’s extremely similar genotype. A second benefit of using yeast as a model organism is the fact that yeast have obtained the ability to adapt to drastic environmental changes. Via this adaptation, the organism is able to maintain the specific internal conditions for optimal growth and function, much like multi-cellular organisms.

Currently, there are only a few genes within the yeast organism that have been implicated in modulating responses to the environment1. A current understanding of general stress regulation and other research has allowed for the observation of several different pathways related to yeast aging. These finding indicate that cellular aging is an emergent property of gene networks2. Full understanding of stress regulators is yet to be acquired, however High-throughput screens have begun to reveal the protein interaction network (PIN) that gives some aspect about some of the cellular functions in the yeast Saccharomyces cerevisiae3. The structure of this network affects the evolution of the proteins that compose it is a fundamental question in molecular evolution. Preliminary studies show the protein’s number of interactors is correlated with its effect on organism fitness, which in turn affects rate of evolution. Fitness is measures as the reduction in relative growth rate due to deleting or disrupting the gene that encodes the protein4.

The developing concept between cellular growth rate/aging and network robustness, measure in coefficient of variation (CV), presents a connecting framework for aging in both dividing (replicative life span) and non-dividing cells (chronological life span). The CV is the standard deviation divided by the mean, which calculates a normalized robustness proxy to compare phenotypic measures at different levels. A statistic measure of lifespan of each module, mean, median, or standard deviation, will be evaluated based on null distributions generated by permutation of lifespans2.

This offers insight on aging from the quantitative genetics perspective, and expand the current knowledge on gene networks and complex traits such as adjustment to environmental stressors1. Throughout this course, students were able to study the ideas of cellular aging via understanding of adjustment to environmental stressors via computer simulations and R and R Studio software.  These cellular organisms need specific atmospheric environments in order to grow and be properly functional. For most cellular organisms, adaptive measures have taken place to conserve homeostasis conditions in the midst of severe harsh external conditions. The *Saccharomyces cerevisiae* has mechanisms which allows it to adjust to the surrounding changes, such as withstanding fluctuations in the kinds and the amount of nutrients, acidity, temperature, radiation and toxic chemicals1. It is necessary for the organism to adapt its genome expression scheme when the outside conditions rapidly change.

In this experiment, the yeast gene expression data came from the NCBI Gene Expression Omnibus (GEO) dataset and the Genomic Spatial Event (GSE) database in order to store, retrieve and analyze all types of high-throughput microarray data3. The gene expressions in this database demonstrated adaptation to stressful environments, two in particular was the focus of this study: Genes GSE33276, which showed gene expression profiles of S. cerevisiae under heat stress and GSE18334, which profiled oxidative stress in Saccharomyces cerevisiae by oxidation of transcription factor Swi6p. For each gene, its observed phenotypic and mutational variations could be due to either its intrinsic robustness or extrinsic robustness of its interacting genes4. This genetic experimental data was analyzed in R, to study the change of robustness during aging between the permutated positive/negative protein and genetic interacting networks.

**Materials and Method for GSE33276**

***NCBI GEO analysis in R***

To first measure the gene expression of GSE33276 and GSE18334 the gene expression matrix data set from NCBI GEO was retrieved. From the data set the expression in the interacting network was recovered. Next the probes and the yeast open reading frames were sorted. These ORFs were separated in 16 columns and 9000 rows. Each row has probe IDS, which confers for each individual gene expression data set. After the probes IDs was converted to ORF IDs based on the structure of the bioconducter gene expression class. This was converted using the match function, which extracted the probe IDs and the ORFs. Since the focus was on the yeast gene, afterwards only the yeast ORFs were extracted. The yeast ORFs were identified as a “Y” followed by a sequence of numbers. After the codes were ran, some of the yeast gene had multiple probes and some only had one probe. For this experiment only one probe was needed for each yeast ORF. To do this an expression matrix was created with ORFs as row names. This approach took only one probe for each ORFs, which is often true for cDNA arrays.

Next the expression matrix is named by ORFs. Another approach is to calculate the average signals for all the probes in the same ORFs. To do this technique it was important to keep track of ORF as the multiple probes. Next the R codes looped over every ORF to calculate the average of the probe signal put into the matrix. To apply to original expression data the probe and row position were calculated by columns, where the gene was observed for every probe. Next the gene expression was normalized, the total expression signals for the genes were measured.

Afterwards everything was normalized to one total maximal signal for the experiment, divided it by its own total then scaled back to its maximal one. This was then looped over both of the experiments.

The coefficient of variation calculated by row for every gene, that provided the variants for every gene in the data set. Next the quantity calculated to indicated the extent of deviation for a group as a whole gave the standard deviation (stddev). This was then incorporated in the data frame, outputted into a csv file, and tested. The CV, stddev and mean was tested through the histograms figures.

***Correaltion of the CV to RLS, fitness, interaction degree of protein and genetic networks***

To examine the correlation of expression CV to RLS, fitness, interaction degree of protein and genetic networks, the expression CV from the analyzed data was calculated. The RLS and the CV were merged together to apply the linear regression analysis. A large CV or stddev meant that the data was noisy and less robust, which could be based on the height of the fitness results. The correlation of expression CV, stddev, mean was then investigated with the interaction degree of positive genetic interaction degree. And the multiple regressions and results were recorded.

***Permutation***

Finally permutation was used to calculate the different of expression CV or stddev in protein networks and genetic networks. The transcription profile was analyzed in the protein network in pairs and then it was read in the data. The genetic network in pairs has epsilion value in sttddev and p-value. The epsilon tells whether it is positive or negative interaction. Epsilion equals the double deletion mutant minus the product of the single deletion mutant. The genetic interaction network was then read into R from the sgadata\_costanzo2009\_stringentCutoff\_101120.csv. First a generic function was defined to calculate difference in interacting pairs of proteins, using the values. This was calculated for one interacting network. The first attempt for a generic permutation test function analyzed the original interacting pairs. Next the two columns were merged to a single column for a 'sample' function. Afterwards the amount of simulations were quantified and stored. Next the sample function was used to permute the pairs. For each permutation sample list, the ID repetition was incorporated in the two columns (ORF 1 and ORF 2) and merged them which made a new interaction network. After the permutation a different value function was ran. To define the permutation function the newids were form in a single column vector. After the newids were reformatted to two columns (random pairs). It was then split into two columns; the first half and then second hand. Next the delta.K was calculated for one random network, and that signified the end of the permutation function. The observed difference in PIN was converted in the CV data to form the "ORF" and "Value". The CV table had 4 columns and that was inputted in the generic function. Next, the observed difference of CV of interacting pairs was simulated 1000 times. After the histogram of the CV was permuted, which gave a p-value. The observed difference calculated in Positive GIN and Negative GIN both generated figures and p-values

**Results for GSE33276**

The average signals for all the probes in the same ORFs was used to first calculate the total expression signal for every experiment. Observing the signals some have strong signals and some have weak signals which indicates either too much RNA or less RNA **Figs 1-3.**



**Fig. 1** After normalization, the first experiment was summed up to the maximal signal distribution



**Fig. 2** coefficient of variation



**Fig. 3.** Distribution of signals after log transformation.

The coefficient of variation calculated by row for every gene, provided the variants for every gene in the data set. Next the quantity calculated to indicated the extent of deviation for a group as a whole gave the standard deviation (stddev). This was then incorporated in the data frame, outputted into a csv file, and tested **Figs 4-6**.



**Fig. 4** Frequency of signals from the coefficient of variation



**Fig. 5** Frequency of signals from the standard deviation



**Fig. 6** Frequency of signals from the calculated mean value

**The merged data from the Replicative lifespan (RLS) table and the coefficient of variation table (CV) Figs 7-8.**

Standard Deviation and RLS

Multiple R-squared: 0.0004902, Adjusted R-squared: -0.001529

p-value: 0.6224

Calculated Mean and RLS

Multiple R-squared: 0.0004052, Adjusted R-squared: -0.001614

p-value: 0.6544

RLS and CV regression data

Multiple R-squared: 2.823e-05; p-value: 0.9059

Long RLS and CV regression data

Multiple R-squared: 0.0001108; p-value: 0.8149

plot( RLS.tb2$RLS\_Del\_alpha ~ RLS.tb2$myCV )

abline(m, col='red')

**Fig. 7**  The correlation of expression CV (standard deviation or mean) to RLS**.** The absolute value line for the merged data of the RLS and CV

The merged data from the coefficient of variation (CV) table and the fitness (YPD)

CV and YPD data

Multiple R-squared: 0.02634; p-value: < 2.2e-16

plot( YPD ~ myCV, data=fit.tb2)

abline(m, col="red")



**Fig. 8** The correlation of expression CV (standard deviation or mean) to fitness**.** The absolute value line for the merged data of the fitness and CV

Merge GIN, PIN to RLS

After the histogram of the CV was permuted, which gave a p-value. The observed difference calculated in Positive GIN and Negative GIN both generated figures and p-values **Figs. 9-13**.

Expression CV, stddev, mean ~ interaction positive genetic interaction degree

Multiple R-squared: 0.004723; p-value: 0.3566

Multiple R-squared: 0.01195; p-value: 0.09967

Expression CV, stddev, mean ~ interaction degree of negative genetic interaction degree

Multiple R-squared: 0.003069; p-value: 0.4051

Permutation of protein intereacting pairs (PIN).

hist(permutated.diff.CV)



**Fig. 9.** calculated the oberserved difference in variation of coeffictients of interacting pairs (simulated 1000 times)

p-value: .001

#generate a figure

mylim = c(min(c(permutated.diff.CV, diff.CV.obs))\*0.95, max(c(permutated.diff.CV, diff.CV.obs))\*1.05 )

hist( permutated.diff.CV, xlim=c(mylim ), br=20, main="PIN" );

arrows( diff.CV.obs, 50, diff.CV.obs, 2, col="red" );

text( diff.CV.obs, 50.5, "obs");



**Fig. 10.** Histogram of the protein interaction network (PIN): minimal of the simulated and observational value (scaled down 5%) and the maximum value of it (scale up to 5%)

Permutation of positive genetic intereacting pairs (Positive GIN)

permutated.diff.CV = call.diff.Value.in.permuted.pairs( PositiveGIN.pairs.tb, inData, 1000)

hist(permutated.diff.CV)



**Fig. 11.**. calculated the oberserved difference in variation of coeffictients of interacting pairs (simulated 1000 times)

p-value: 0.054

Histogram Positive genetic interaction network (Positive GIN )

mylim = c(min(c(permutated.diff.CV, diff.CV.obs))\*0.95, max(c(permutated.diff.CV, diff.CV.obs))\*1.05 )

hist( permutated.diff.CV, xlim=c(mylim ), br=20, main="PositiveGIN" );

arrows( diff.CV.obs, 50, diff.CV.obs, 2, col="red" );

text( diff.CV.obs, 50.5, "obs");



**Fig 11.** minimal of the simulated and observational value (scaled down 5%) and the maximum value of it (scale up to 5%)

p-value: .062

Permutation of negative genetic intereacting pairs (NegativeGIN)

permutated.diff.CV = call.diff.Value.in.permuted.pairs( NegativeGIN.pairs.tb, inData, 1000)

hist(permutated.diff.CV)



**Fig 12.** calculated the oberserved difference in variation of coeffictients of interacting pairs (simulated 1000 times)

Histogram of the negative genetic interaction network (NegativeGIN):



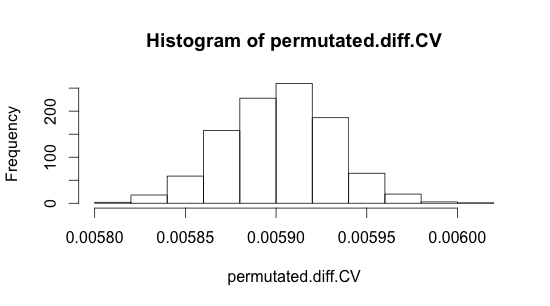
**Fig 13** minimal of the simulated and observational value (scaled down 5%) and the maximum value of it (scale up to 5%)

**Conclusion for GSE33276**

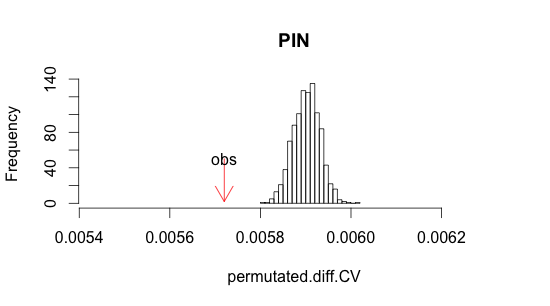
In this lab yeast gene expression data sets provided patterns of biological functions of proteins while adaptating to stressful environments1. This expression of robustness, measured in CV, showed the correlation in RLS, fitness, interaction degree of protein and genetic networks and this also gave multiple regressions and other results. The resulting signals in **Figs 1-3** for all the probes in the same ORFs illustrated the maximal signal distribution for the total expression. The strong signals gave evidence to the large amount of RNA expressed in the database, and the weak signals indicate the small amount of RNA. In **Figs 4 and 5** the frequency of the signals remained at a high height in the beginning region of the CV and stddev test. This means there was a large quantity of robustness that compared to the phenotypic measurements. For the mean in **Fig 6**, there was a large abundance of high frequency signals in the central area of the test. For **Figs 7** the R2 and p-values showed significant correlations for the merged data of RLS and CV regression data. There was also a significant relationship between the merged data of the fitness (YPD**)** the CV in **Fig 8**. In the expression of CV, stddev and mean with the interaction of positive and negative GIN there was significant p-value and R2. In **Fig. 9** the calculated p-value for the observed difference in variation of coefficients of interacting pairs was not significant, and that is why in **Fig 10**  the arrow for the observed value did not fall within the nominal distribution signals of the PIN. For the permutation of positive and negative genetic interacting pairs the p-values were both significant, indicated by the red arrows in **Figs 11 and 13**.

**Conclusion for GSE18334**

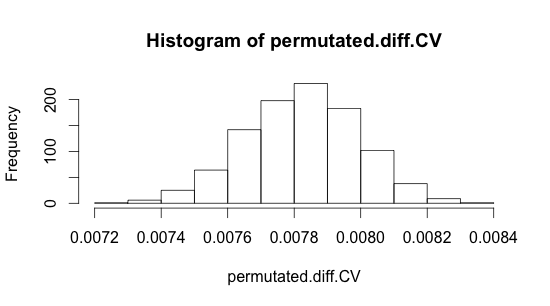
This experiment aims to examine the profiling oxidative stress in Saccharomyces cerevisiae by oxidation of transcription factor Swi6p. R coding was used for this procedure. Graphs were formulated to help create a better understanding of these results. For this particular part of the experiment, R studio and generic functions were used to do permutation tests on interacting gene pairs. It was possible to read into the protein interaction network and the genetic interaction network in pairs. The genetic interaction contains a an epsilon which shows a positive or negative interaction. A generic function was written to calculate the difference of the value in the interacting pairs. A sample function was used to permutated the pairs into two columns. The pairs were simulated one thousand times to create the following graph:



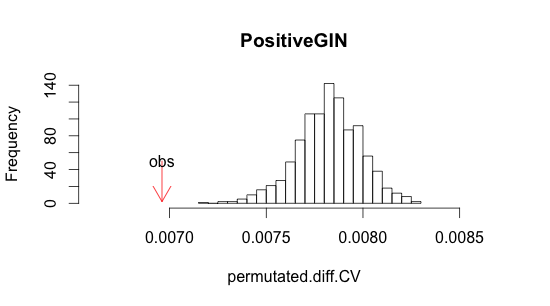
This graph is a histogram of permutated CV. Based on information on my workspace (which states that diff.CV.obs is equal to 0.0058 and p-value is 0.436), my data contains a highly significant p-value. By scaling our values, we can create a second graph which is the following:



This graph is a histogram showing information for the protein interaction network and the significance of the p-value. It shows the observation of delta CV. We can do use the same methods to create graphs specifically for the positive interaction network. Another permutation is done for the positive interaction pairs which produces this graph:



As before we can formulate a second graph that will show p-value and other pertinent information.



This graph is a histogram showing information for the positive gene interaction network. The p-value and its significance is also seen in this graph.

**Discussion**

In this experiment the gene expression profiles from S. cerevisiae under heat stress GSE33276 and GSE18334 under oxidation was analyzed. To determine the robustness effect of every yeast gene in this data set, the standard deviation (or coefficient of variation) for a gene expression data set was calculated. Then the correlation among the expression robustness measured against the RLS, fitness, protein interaction network, genetic interaction network, evolution3. Next using the calculated expression CV (or standard deviation/ mean) data, the expression CV or standard deviation/ mean) were correlated to RLS, fitness, interaction degree of protein and genetic networks2. Finally, permutation is used to analyze distribution of CV (or stddev) in the protein networks and genetic networks. What was investigated in this study will lead future researchers to gain insights into the physiological effects of each of the stresses and its correlation to other biological functions.

References

1Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown􏰍 *Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes.* Molecular Biology of the Cell Vol. 11, 4241– 4257

2Sukhraj Pal Singh Dhami, Nitish Mittal, Sarath Chandra Janga and Nilanjan Roy. *Comparative analysis of gene expression and regulation of replicative aging associated genes in S. cerevisiae*

3Zotenko E, Mestre J, O’Leary DP, Przytycka TM (2008) Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential: Reexamining the Connection between the Network Topology and Essentiality. PLoS Comput Biol 4(8):

4Hunter B. Fraser, Aaron E. Hirsh, Lars M. Steinmetz, Curt Scharfe, Marcus W. Feldman. *Evolutionary Rate in the Protein Interaction Network*