Creating an Auxiliary Page for The GRNSight: Web App and Service for Visualizing Models of Gene Regulatory Networks

Mary S. Balducci[1], Nicole M. Kalcic[2], Arash A. Lari[3], Hayden T. Hinsch[4],

[1] Biology Department, Loyola Marymount University, Los Angeles, CA, United States

[2] Communications, Loyola Marymount University, Los Angeles, CA, United States

[3] Computer Science Department, Loyola Marymount University, Los Angeles, CA, United States

[4] Psychology Department, Loyola Marymount University, Los Angeles, CA, United States

BIOL 367 Biological Databases

15 December 2017

Introduction

Saccharomyces cerevisiae, also known as yeast, has been used for centuries by humans to make bread or alcoholic beverages. In more recent times, *Saccharomyces cerevisiae* has been used to study the functions of genes. *Saccharomyces cerevisiae* has served as a model for the eukaryotic system. In particular, the organism's response to stressors. DNA microarrays have been utilized in order to find quantitative data to describe the cellular processes. These DNA microarrays have been used to describe the response of *Saccharomyces cerevisiae* to cold shock and heat shock. As a result of this cold shock and heat shock *Saccharomyces cerevisiae* change their gene expression in order to cope with these stressors.

Students in the BIOL/CMSI 367-01 class performed a statistical analysis in order to find which genes changed expression significantly. This produced a list of genes which were clustered based on how they were expressed. Dr. Dahlquist's experiment produced two distinct patterns: groups of genes that were upregulated in reaction to cold shock and downregulated in reaction to heat shock as well as genes that were downregulated in reaction to cold shock and upregulated in reaction to heat shock. We then needed to determine the function of these genes by utilizing the Gene Ontology, which explains the function of the genes. There was a Gene Ontology Term related to each cluster which communicated the function of the genes. Next, we needed to determine which transcription factors were causing the changes of the genes. Our Data Analyst utilized YEASTRACT to determine what the transcription factors were turning on and off. The results of this analysis spawned a network of transcription factors. These YEASTRACT results were required to determine which networks were causing the gene expression to be

upregulated or downregulated. These networks were then put into a mathematical model in order to see whether or not they are turning on or off a gene. This network could then be visualized on GRNSight.

As members of the Page Design Team, we were tasked with the designing and coding of an auxiliary page for GRNSight. An auxiliary page that would display relevant information of the transcription factors in the network displayed on GRNSight. The students of BIOL/CMSI 367-01 determined that the relevant information that needs to be displayed on the auxiliary page is as follows: corresponding gene ID's of the transcription factors from each database, the description and function of the transcription factor from Ensembl, the transcription factor DNA sequence from Ensembl, the transcription factor protein sequence from UniProt, the locus tag of the transcription factor from NCBI, alternative names of the transcription factor from NCBI, the regulation of the gene from SGD, the interaction of the transcription factor from SGD, similar proteins from Uniprot, the gene ontology of the transcription factor from SGD, and the matrix ID, class heat shock factors, family HSF factors, sequence logo,frequency matrix and the consensus sequence from JASPAR. This auxiliary page creates a single location where biologists can go to find relevant information of transcription factors.

Designer Materials, Methods, Results & Discussion

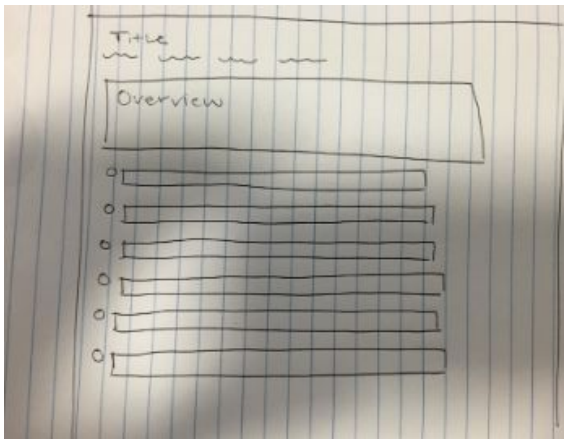
I had the opportunity to be the designer for our project, which let me simultaneously tap into my artistic background and learn what it takes to put together code for a webpage. To begin

my part of the assignment, I went back to the "Class Journal Week 9" page on our LMU BioDB 2017 wiki. The class journal assignment for Week 9 was to reflect on the "My Favorite Gene" project and identify two personal favorite pages (as well as why the chosen favorite pages were particularly appealing). I made a simple tally to showcase the popularity of each gene page. The ASP1 page was mentioned ten times by students, making it the most liked page out of the different projects. The SPT15 and ACT1 pages tied, with five votes each. They were followed with CLN1 in fourth place-- four votes.

While there was an overwhelming response towards SPT15, I didn't feel that it was the best page to model our project after. Instead, I tried to pull together a consensus on qualities that the class liked in the gene pages. I found that everyone wanted something that was organized and clear. Pages were mentioned because they didn't have too much going on, meaning that anything too fancy or distracting was going to be a miss. Yet, I did note that people wanted something that looked professional and slightly stylized. Anything that appeared to be an information dump on a white background was not going to satisfy the class.

I decided to base our page off of the ACT1 page. This may appear to be a biased decision, because I worked on this page in the middle of the semester with Eddie Bachoura. However, I mostly acquired the biology research component of our ACT1 project, so the base of the design was truly created by him. I think that the drop down bars keep the page clean but interesting. The subtle hints of color via the links at the top add something different, without taking away from the academic nature of the website.

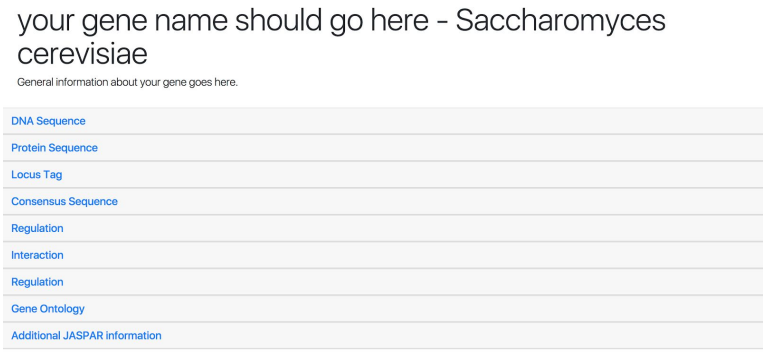
I quickly sketched out the design so that our coder, Arash, would have a copy on paper. It looks as follows:



Once Arash was able to do the setup for the page and explain it to me, he asked me to develop a logo for the top layout on Adobe Illustrator. I set up Illustrator with a 0.5 by 6 inch graphic template. I knew that it needed to be simple. Arash mentioned the idea of an eye for the GRNSight logo. After deciding on a font for the title, I used the shaper tool to make a transparent eye to layer over the green background. I added in the website description and the small detail of the white bar, and then I layered the graphic with a transparent light green rectangle. This would probably be a fairly easy task had I known how to use Illustrator beforehand, but it was brand new to me. Luckily, it turned out well and I emailed the saved SVG (Scalable Vector Graphic--so that there would not be a loss of image quality) to Arash. He notified me that he received it and that it worked. He then implemented it. Below is an image of the logo I created, implemented on the page.

GRNSIGHT Web app and service for visualizing models of gene regulatory networks

Coder Materials, Methods, Results & Discussion

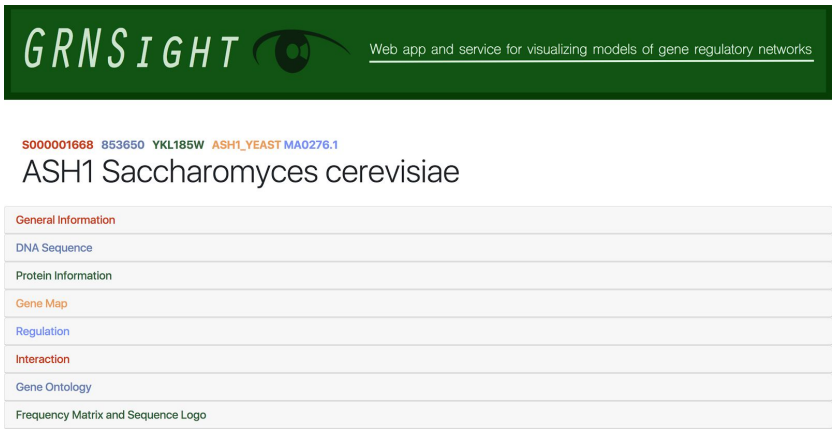


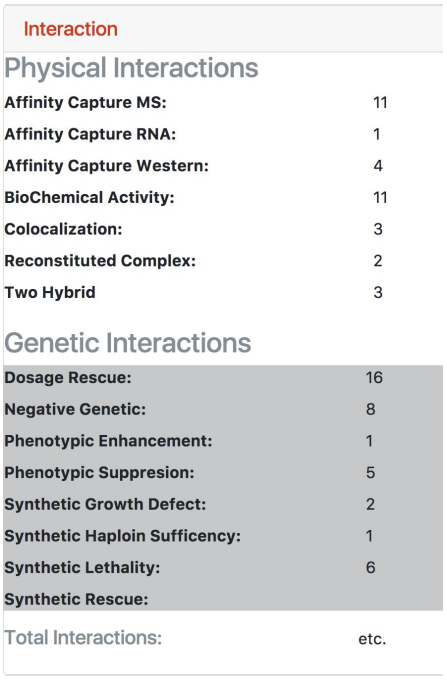
As the Coder for this assignment, I was tasked with providing a template page that hosts all the information pulled by the API teams and provide selectors for the Integration team to know where to place the information on the page. The first part of this process involved looking at all the previous web pages made by our fellow classmates and determine that the ACT1 page designed by Eddie Bachoura was the most aesthetically pleasing and informative design. Based on the preliminary design our designer, Nicole, made on paper. I began writing a basic html document based on her design which served as the foundation for the rest of the page.

After that was laid out and uploaded, I was in constant back and forth with the integration team to ensure that I had all the right selectors for them to be able to place all the relevant information. Furthermore I spent time with fellow student and Biology major Katie Wright and was informed on how the page layout could be further optimized to host information that biologists want to see. As someone who is not very well versed in biology, this was very helpful. Afterwards, I had to make the page look less plain and more effective to the user, as previously most of the information was just laid out in html paragraph tags. The way the data was laid out

before being stylized using bootstrap worked, but did not make for a very enjoyable reading experience.

The final design eventually had distinct colors on each tab, which stayed within the theme of the gene ID's that were colored since the beginning. I realized that for much of the page, a description list would be the most effective way of laying out the information as generally that's what they were, a list of descriptions.






```
<div class= "geneticInteractions">
  <h3 class = "text-muted"> Genetic Interactions </h3>
  <dl class = "row table table-dark">  <!-- Made gray for style purpo
  <dt class = "col-sm-3">Dosage Rescue: </dt>
    <dd class="dosageRescue col-sm-9"> </dd> <!-- Return Just the Val
    <dt class = "col-sm-3">Negative Genetic: </dt>
    <dd class="negativeGenetic col-sm-9"> </dd>
    <dt class = "col-sm-3">Phenotypic Enhancement: </dt>
    <dd class="phenotypicEnhancement col-sm-9"> </dd>
    <dt class = "col-sm-3">Phenotypic Suppresion: </dt>
    <dd class="phenotypicSuppression col-sm-9"> </dd>
    <dt class = "col-sm-3">Synthetic Growth Defect: </dt>
    <dd class="syntheticGrowthDefect col-sm-9"> </dd>
    <dt class = "col-sm-3">Synthetic Haploin Sufficency: </dt>
    <dd class="syntheticHaploinsufficiency col-sm-9"> </dd>
    <dt class = "col-sm-3">Synthetic Lethality: </dt>
    <dd class="syntheticLethality col-sm-9"> </dd>
    <dt class = "col-sm-3">Synthetic Rescue: </dt>
    <dd class="syntheticRescue col-sm-9"> </dd>
  </dl>
</div>
```

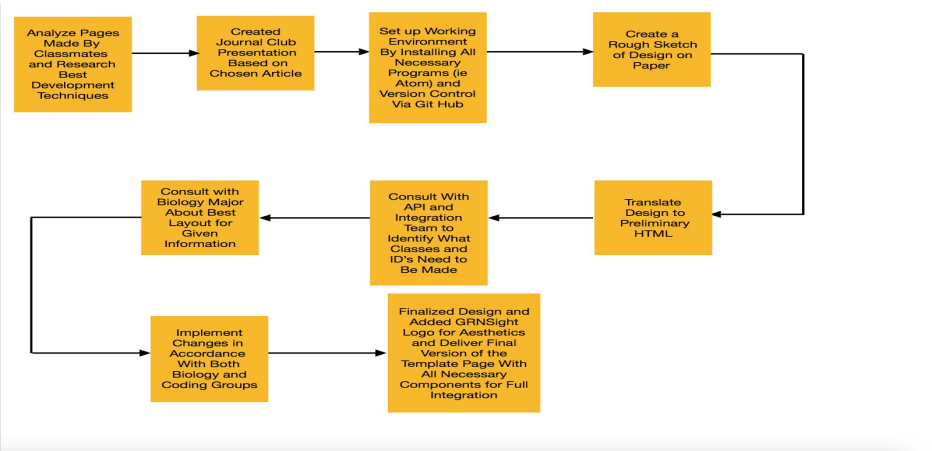
Above, you can see the final result and some of the corresponding code. I used html comments to clarify anything that would potentially be unclear for others who wanted to read the

page. One of the last things I had to do was let Blair and Zach (the coders from the integration team) know that they had to remove all the strings that their objects returned as I decided it would look better if I placed the strings on the page myself. I felt bad doing this because I had previously requested they do the exact opposite, but thankfully it was not a very long process. Below you can see the code for the element that the interaction team placed on our template page. As you can see, there is no string being returned in the final version; before there was a ““Affinity Capture RNA “ +” within the text parenthesis, this was ultimately the best design choice.

```
var sgdAffinityCaptureRNA = gene.sgd.affinityCaptureRNA;
$(".affinityCaptureRNA").text(sgdAffinityCaptureRNA).attr({ href: sgdHrefTemplate + sgdAffinityCaptureRNA });
```

There were similar stylistic additions and changes throughout the page for all the elements that successfully got returned by the API calls. The very last element I added was at the recommendation of Blair, who designed the ACT 1 page, and that was to add a fluid-container tag to the entire body so the page wouldn’t be stuck to the margins, and this really helped make the page more pleasing to browse.

Workflow for the Coder/Designer:



Data Analyst Materials, Methods, Results and Discussion

I analyzed data from a DNA microarray, gathered by Dr. Dahlquist's lab. I started with data for which the red/green fluorescence ratios had already been normalized. My first step was to download the Excel spreadsheet with the data. I was focused solely on the strain with a CIN5 deletion, so I deleted the data from the other strains. After that I found p-values for the number of genes affected by the temperature change in that strain. The p-values had different parameters and each of them resulted in different percentages of the geneset which were affected by the change, the specific processes I followed are located in my Week 8 page.

After finding the p-values, I loaded the data into the Short-Time series Expression Miner (STEM) software. This program created clusters of genes which changed expression similarly to each other. It also labelled the seven most significant clusters (called "profiles") in color. I chose profile #22 to analyze further. This profile came with a Gene Ontology term list, containing terms related to the genes in the profile and their expressions. The processes I followed for running STEM are outlined in my Week 10 page.

From the STEM results of profile #22 I found a list of transcription factors which were connected to each other. I did this by using YEASTRACT, specifically their "Rank by TF" feature. Once I had the list, I used the YEASTRACT Gene Regulation Matrix feature to create a mathematical model of how the transcription factors related to each other. This model was then made readable by GRNSight. I uploaded it into GRNSight to transform the mathematical data

into a visual map. Next, I used MatLab to add weights to the map so that the magnitude and direction of each relationship would be shown on the GRNSight map via different colors and sizes of arrows between the transcription factors. The processes of using YEASTRACT, GRNSight, and MatLab are all in my Week 14 page.

The first part of my process involved using Excel to find p-values for the dCIN5 dataset. I found the number and percentage of genes at multiple p-value limits, as well as corrected p-values. I then compared that data to the results from the wild-type strain and compared the two. In both, the Bonferroni-corrected p-value gave the least number and percentage of significantly affected genes. This is because the Bonferroni-corrected p-value has the most stringent criteria for significance of any of the types. The comparison between the two strains is shown in Table 1.

ANOVA	WT	*dCIN5*
$p < 0.05$	2528 (40.85%)	2290 (37%)
$p < 0.01$	1652 (26.70%)	1380 (22%)
$p < 0.001$	919 (14.85%)	691 (11%)
$p < 0.0001$	496 (8.01%)	358 (6%)
Benjamini & Hochberg-corrected $p < 0.05$	1822 (29.44%)	1453 (23%)
Bonferroni-corrected $p < 0.05$	248 (4.01%)	151 (2%)

Table 1: Shows the number and percentage of total genes that significantly changed expression according to various p-values, for both the wild-type (WT) and dCIN5 strains.

After finding those results, I ran the data through the STEM software. This created clusters of genes based on how they changed expression. The seven most significant clusters, called profiles were put in color. This is shown in Figure 1. I chose one of those seven most significant profiles to analyze further. I chose profile #22, which is shown in green in Figure 1, and more close-up in

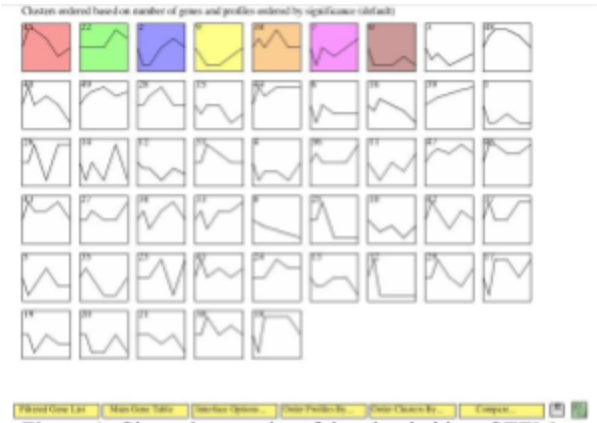


Figure 1: Clustering results of data loaded into STEM. Most significant profiles are shown in color at the top.

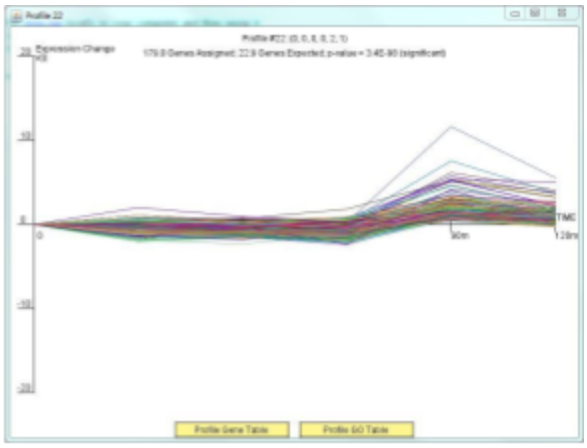


Figure 2: An up-close view of Profile #22, indicating the expression changes for the genes. They seem to increase in expression at about 90 minutes, during the recovery time from the cold shock.

Figure 2. I chose profile #22 for a few reasons. For one, this profile contained many genes, which I thought would give me more accurate results as well as Gene Ontology terms which could be used to determine how those genes work in relationship to each other. Another reason I chose this one was because of the interesting expression pattern it showed. The genes in this profile seem to only change in expression at the 90 minute mark, during the recovery period after the cold shock, at which point they show an increase in expression and then a decrease by the 120 minute mark. The STEM software also came up with a Gene Ontology (GO) term list for each of the significant profiles. This contains terms related to the genes in the profile and their expressions. A portion of the GO list from profile #22 is shown in Figure 3. Some of the gene ontology terms from profile #22 included "response to stress" and "cellular response to toxic substance". These seem to indicate that the genes in this profile did change expression in response to some sort of stress, mostly likely the stress from the cold shock.

Category ID	Category Name	#Genes Category	#Genes Assigned	#Genes Expected	#Genes Enriched	p-value	Corrected p-value	Fold
GO:0005737	cytoplasm	1013	150.0	124.0	+26.0	1.9E-6	<0.001	1.2
GO:0006979	response to oxidative stress	38	16.0	4.7	+11.3	2.6E-6	<0.001	3.4
GO:0003779	actin binding	6	6.0	0.7	+5.3	3.1E-6	<0.001	8.2
GO:0034599	cellular response to oxidative stress	35	15.0	4.3	+10.7	4.2E-6	<0.001	3.5
GO:0006950	response to stress	172	40.0	21.1	+18.9	1.2E-5	0.004	1.9
GO:0030863	cortical cytoskeleton	15	9.0	1.8	+7.2	1.3E-5	0.004	4.9
GO:0097237	cellular response to toxic substance	15	9.0	1.8	+7.2	1.3E-5	0.004	4.9

Figure 3: Shows a piece of the GO list from Profile #22. Significant terms include "response to oxidative stress", "response to stress", and "cellular response to toxic substance", indicating that this cluster of genes did change expression in response to the temperature change.

The next part of my process was to load the information from profile #22 into YEASTRACT, in order to get a list of significant transcription factors in this profile. This resulted in a list of transcription factors, part of which is shown in Table 2. Three of those transcription factors, CIN5, ZAP1, and GLN3 were kept in the list because those were analyzed by the other Data Analysts in our Biological Databases class. I used the list of transcription factors to create a mathematical gene regulation matrix in YEASTRACT.

Transcription Factor	% in User Set	% in YEASTRACT	p-value
CIN5p	0.4068	0.033	0.001495075
ZAP1p	0.3955	0.0455	0.000000012628217
GLN3p	0.1921	0.0286	0.152594064461196

Table 2: Shows a portion of the "Rank by TF" results from YEASTRACT. The data from these transcription factors was chosen because these were the strains that the other data analysts and I analyzed.

This matrix was read by GRNSight and transformed into an understandable map. However, this first map did not show the magnitudes or directions of the interactions between transcription factors. It only showed thin gray arrows to indicate that there was an interaction, shown in Figure 4. This issue was corrected with the use of MatLab to add weights and colors to those lines. The final result was a map with colored lines of varying thicknesses, shown in Figure 5. The different

Figure 4: Unweighted map of the relationships between significant transcription factors from GRNSight.

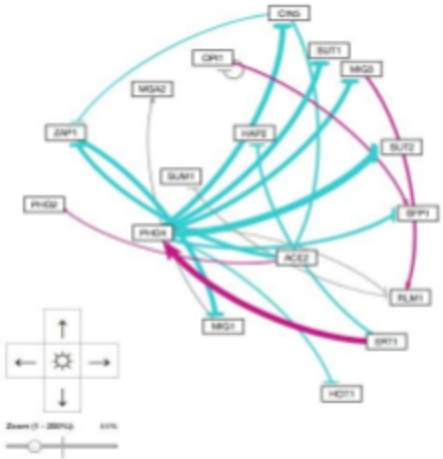


Figure 5: Weighted map of significant transcription factors from GRNSight. Shows the expression changes and relationships between transcription factors. Blue represents a downregulation and pink represents an upregulation. Thicker lines represent a large change in expression.

colors of the lines represent up (pink) or down (blue) regulation of the genes. The thicknesses of the lines represents the magnitude of the expression change.

The final graph from GRNSight is useful in seeing how these transcription factors affect each other. It is interesting to see that some of the transcription factors have large blue arrows indicating they are being repressed. They also have large pink arrows going away from them, indicating they are inducing a different gene, while being repressed themselves. In addition, most of the lines are blue, indicating that most of these transcription factors were repressed during the experiment. In the future, it may be possible to see more information about these genes and how they interact with each other by clicking on the transcription factors or the lines themselves.

Conclusion

The journal club presentation that the designer and coder were tasked with was a presentation about Data-Driven Design. The article provided went in depth about all the methods that could potentially be used by web developers for figuring out the best layout for their pages. Most of the material the designer and coder covered was interesting and potentially useful in their prospective future careers as a computer scientist and studio artist, but not immediately applicable to the current project simply due to a lack of resources and time. Having said that, the coder and designer did do some trial and error with design choices by talking to the biologists of the class and watching how they used the page and listened to the feedback they provided, so in that sense we did follow data-driven design. As more elements are brought onto the page, there are going to be more design choices future developers working on this site can add, given that they have a good idea of what biologists want out of the page.

In regard to the Data Analyst and Quality Assurance/Project Manager, the journal club article proved to have some differences and similarities to the results the Data Analyst found. The journal club article we reviewed also utilized Gene Ontology terms in order to determine which helped them find the transcriptional factors which were responsible for the increased fitness of the wine strain of yeast they used in their experiment. Gene Ontology terms were also utilized in our analysis to find which functions were related to corresponding clusters of genes. Pizarro et al. also found two distinct patterns within the clusters of their genes: genes that were upregulated when the temperature was decreased and downregulated when it was increased as well as genes that were downregulated when the temperature was decreased and upregulated when it was increased. This suggests that our analysis is consistent with previous studies. However, Pizarro et al. performed an experiment where there was an initial heat shock rather than a delayed heat shock. Our work further expands on the findings of the article because we have created a single page where biologists can go to find information regarding transcription factors of genes related to stress such as heat and cold shock. This creates a potential backbone for the industry and makes gene regulatory networks to be easily accessible to every scientist interested. We've laid grounds for a tool that could be used all over the world; a tool that will aid in further research and potential findings.

Acknowledgements

We would like to acknowledge Dr. Dahlquist and Dr. Dionisio for teaching this class. We would also like to acknowledge the other teams from class who worked with us to create an auxiliary page for GRNSight and analyze data.

References

GRNsight. (2017). Home. Retrieved November 30, 2017 from http://dondi.github.io/GRNsight/

Liikkanen, L. (2017). The Data-Driven Design Era in Professional Web Design. Accessed November 13, 2017, from http://delivery.acm.org/10.1145/3130000/3121355/p52-liikkanen.pdf?ip=157.242.223.254&id=3121355&acc=ACTIVE%20SERVICE&key=80B0E63637265656%2E958C566062FD1C18%2E4D4702B0C3E38B35%2E4D4702B0C3E38B35&CFID=829067610&CFTOKEN=20249523&__acm__=1510614806_1e8b314e23f771f9b139d99596801365

LMU BioDB 2017. (2017). Week 8 Journal Retrieved Decmber 14, 2017 from, https://xmlpipedb.cs.lmu.edu/biodb/fall2017/index.php/Mbalducc_Week_8

LMU BioDB 2017. (2017). Week 10 Journal Retrieved Decmber 14, 2017 from, https://xmlpipedb.cs.lmu.edu/biodb/fall2017/index.php/Mbalducc_Week_10

LMU BioDB 2017. (2017). Week 14 Journal Retrieved Decmber 14, 2017 from, https://xmlpipedb.cs.lmu.edu/biodb/fall2017/index.php/Mbalducc_Week_14

Pizzaro, Jewett, Nielson, Agosin. (2008) Growth Temperature Exerts Differential Physiological and Transcriptional Responses in Laboratory and Wine Strains of Saccharomyces Cerevisiae. "Applied and Environmental Microbiology", 74, 6358-6368, doi:10.1128/AEM.00602-08

YEASTRACT. Generate Regulation Matrix. Retrieved November 30, 2017, from http://www.yeastract.com/formgenerateregulationmatrix.php

YEASTRACT. Rank by TF. Retrieved November 30, 2017, from http://www.yeastract.com/formrankbytf.php

Github. (2017). bhamilton18/GRNsight, Retrieved December 15, 2017, from https://github.com/bhamilton18/GRNsight/tree/page-design