Joyce Sun
BIOL266– Professor Weir
Maya Milrod, Annika Velez
9 May 2022

Yeast Gene Protein Expression Exploration via Information Theory

Introduction

Molecular biology has greatly benefited from the use of information theory (IT), which was originally developed for the analysis of communication systems, but has been useful in genome global analysis, classification of motifs, and specifically to this exploration: the prediction of transcription factor binding sites and sequence characterization based on local sequences. Information theory conveys information measured in bits about an event via a message. The message in bits provides a probability distribution of the likelihood of an event occurring. In this exploration, an event refers to the probability of a nucleotide being in a certain position, and bits of information are calculated by taking sets of aligned sequences and tabulating the likelihood of nucleotides to appear at certain positions. Bits of information are interpreted as such: events with greater likelihood of occurrence carry less bits of information than those with a lesser likelihood of occurrence carry more bits of information.

In this exploration, information theory is used to analyze different parts and features of the sequences surrounding the annotated (known) start site of yeast genes from the dev_Yest350 database, and focused on examining aspects of sequences surrounding the start site of genes in the top 10% of protein expression, the middle 10%, from 45% to 55%, of protein expression, and the bottom 10% of protein expression. A general exploration of information content was conducted on sequences 10 nucleotides before (-10) the known start sequence (annAUG) and 10 nucleotides after the annAUG (+10). In addition, further investigations of these data were conducted, which includes using information content to predict consensus sequences of nucleotides -6 to -1 relative to the annAUG, comparing the frequency of G content to the frequency of GNN codons, and determining and comparing information content of sequences with upstream ATGs in the 5’ UTR and sequences without upstream ATGs.

Methods

Data for these explorations were selected from the dev_Yeast350 relational database using SQL to select and data pertinent only to the investigations. cDNA sequences of genes and their respective names in the yeast genome were sorted by protein expression, and the genes in the top 10%, middle 10%, and bottom 10% of protein expression were collected. The files were then formatted into a text file (.txt) via a formatting python program (Appendix 5) so that the data could be processed by the functions calculating information content. All files used for the preliminary investigation and further explorations were formatted using the aforementioned formatting program (Appendix 5) to splice sequences specific to each investigation from the original cDNA sequences. The code (Appendix 5) has been annotated to indicate what commands select which parts of the original sequence, and the selected parts of the sequence

(typically without gene name as there was no need to name specific genes) are outputted with one gene per line into a new text file with a descriptive name.

For the preliminary investigation (Appendix 1) , four functions were created to work in conjunction with each other to output either an information content score for each line in the file, or the average information content of each gene in the file. To calculate the information content scores across a singular gene or an entire file of genes, weights for each position in the sequence was calculated using equation $w(x,j) = 2+\log_2(f_{x,j})$, where w refers to weight of the nucleotide, j refers to the nucleotide's position, and f refers to the frequency of a certain base at position j. These weights are then totaled for an entire sequence to output the information content score, using equation $\text{infoscore}(t_i) = \text{sum } \{w(t_{ij}, j) \mid 1 \leq j \leq m \}$, where t_i refers to a sequence, and the $\text{infoscore}(t_i)$ is just the sum of the bits contributed by the nucleotides found at each of the positions in the sequence t_i. These info scores for each sequence were then averaged to output the overall average info score per file (Sources 4 and 5).

To determine a consensus sequence for yeast in positions -6 to -1 relative to the annAUG for the 3 sets of data, information content for each position in the sequences for each set of genes was calculated using functions from the preliminary investigation (Appendix 1). From this output, the bases at each position with the highest information content was selected for the consensus sequence. However, when constructing consensus sequences, the threshold for selecting bases must also be considered based on $w(a, j) \geq \tau$ and $w(a, j) \leq \tau$, where a refers to the base and j refers to the position of a, and τ refers to the threshold. Because the sequences being investigated are aligned with a lesser degree of conservation, a smaller τ is necessary to compensate for the conservation. Thus, in creating the consensus sequence, τ was set to 0.5, so that bases at selected positions were only selected for the consensus sequence if the base with the greatest weight exceeded or was equal to 0.5. For positions in which no base was greater than or equal to 0.5, a dash was placed as a placeholder to indicate that not enough information was present to predict a base at said position. In addition, a consensus sequence for a supplemental partial Drosophila gene dataset from Cavener 1987 (also sequenced from -6 to -1) was predicted using the same threshold. Code for this function was specifically altered to adhere to the 0.5 threshold, but can be altered as need be for any consensus sequence (Appendix 3).

To compare frequencies of G nucleotides and GNN codons in yeast ORFs (open reading frames), +1069 nucleotides (average yeast ORF length) of each gene for each data set were examined. The percentage of G nucleotides relative to other nucleotides for each set of protein expression data was calculated, as well as the number of codons beginning with G relative to all other codons for each set of protein expression data.

The final investigation conducted was to compare the information content of genes with ATGs prior to the known translation start site to genes without prior ATGs. Data used for these analyses were conducted with sequences -135 relative (excluding ATG itself due to repetitiveness) to the annAUG (average 5' UTR length for yeast(Source 1)) for each data set. To analyze information scores for each sequence with and without upstream ATGs, each data set was first separated into 2 separate text files, one containing genes with upstream ATGs, and one

without upstream ATGs. Then, using the functions from the preliminary investigation (Appendix 1), average information scores for each separated file was calculated.

In analyses done in class, and in interpreting results of these analyses, it was sometimes difficult to piece together what all the outputted information scores really meant. So, to supplement the understanding of results of the conducted analyses, python code was written using the matplotlib library to create histograms and/or bar charts for pertinent outputs to provide a visual representation of the findings.

Results

Preliminary Investigation:

To validate and ensure that the results outputted from the code in Appendix 1 were of the same magnitude, information content outputs for individual genes (from -10 to +10 relative to and excluding the annAUG) from Appendix 1 were compared to the individual information scores calculated for the same genes via Resource F–Wesleyan Information Theoretic Analysis Tool(Source 6). Shown below are the first 33 results of the output of the information_by_line function from Appendix 1 when asked to compute the information by line for 'bottom10%_nogene.txt'(Figure 1) and the output from Resource F of the same genes (Figure 2). It can be seen that information scores calculated by functions in Appendix 1 (to the right of the colon, before the comma) are accurate and correspond to the indInfo column in the output from Resource F. Thus, these functions can be relied on for calculating information content for the remainder of the investigation.

```
{0: 2.7, 1: 5.02, 2: -3.85, 3: 3.48, 4: 2.63, 5: 3.33, 6: 0.68, 7: -0.26, 8: 2.67, 9: -0.12, 10: -1.2
9, 11: 4.29, 12: 6.24, 13: 1.33, 14: 1.29, 15: 1.1, 16: 4.21, 17: 3.29, 18: 3.26, 19: 3.05, 20: 1.81,
21: -0.83, 22: 1.35, 23: 1.21, 24: 1.58, 25: 1.95, 26: 3.92, 27: 2.12, 28: 4.45, 29: 1.87, 30: -1.79,
```

Figure 1: First 30 results from output for information_by_line('bottom10%_nogene.txt). Appendix 1.

Individual Information Scores

Using alignment positions: 1 through 19
Sorted by: seqNum

seqNum	seqName	indInfo	indRelEntropy	indCorrectedInfo	indCorrectedRelEntropy	seqData
1	sequence 1	2.69615922003567	2.69615922003567	2.58795709196899	2.58795709196899	GATCTTATATCCAACCCAA
2	sequence 2	5.02026579246748	5.02026579246748	4.91206366440081	4.91206366440081	CTAAAAAATGTCTTGATAA
3	sequence 3	-3.85255914259322	-3.85255914259322	-3.96076127065989	-3.96076127065989	CCATCTCCTATGAGTCCCG
4	sequence 4	3.48157483655639	3.48157483655639	3.37337270848972	3.37337270848972	CAATACATATTAAAGTTAA
5	sequence 5	2.62571211811315	2.62571211811315	2.51750999004648	2.51750999004648	CAGGACAAGAACAATGATG
6	sequence 6	3.32938068719106	3.32938068719106	3.22117855912439	3.22117855912439	AGAAGTATTACTTTTACAT
7	sequence 7	0.682602453917109	0.682602453917109	0.574400325850437	0.574400325850437	CATTGGTAACTAAAATCCG
8	sequence 8	-0.261180434832131	-0.261180434832131	-0.369382562898804	-0.369382562898804	GCTAAATGGATTTCGACTA
9	sequence 9	2.67200699727912	2.67200699727912	2.56380486921244	2.56380486921244	GCTACAACTGTTAATAATA
10	sequence 10	-0.120581520149001	-0.120581520149001	-0.228783648215673	-0.228783648215673	TGCACTGGAAGCATTGCAT
11	sequence 11	-1.28729358905621	-1.28729358905621	-1.39549571712288	-1.39549571712288	GCAACGCAGTCTCGAGTTG
12	sequence 12	4.28954123002107	4.28954123002107	4.1813391019544	4.1813391019544	TTTTTCAAATCAATACGTC
13	sequence 13	6.2358270317353	6.2358270317353	6.12762490366863	6.12762490366863	AAAAGCAAATATACAAAAC
14	sequence 14	1.32609182891007	1.32609182891007	1.2178897008434	1.2178897008434	GTACTCGAAAGCTCTAGAG
15	sequence 15	1.29105721669901	1.29105721669901	1.18285508863234	1.18285508863234	TATCCCTTAAACACATCAT

Figure 2: First 15 results from output of individual information scores through Resource F: Wesleyan Information Theoretic Analysis Tool). Because the first item in python corresponds to positions 0, sequence 1 from Resource F corresponds to 0 in the python output. Source F.

After confirming the validity of results obtained from functions in Appendix 1, these functions were used to then create visual representations of the exploratory analyses. In order to understand how information is generally distributed across sequences with varying degrees of protein expression, average information content for genes in the top 10%, middle 10% and bottom 10% of protein expression was calculated and displayed in a histogram (Figure 3).

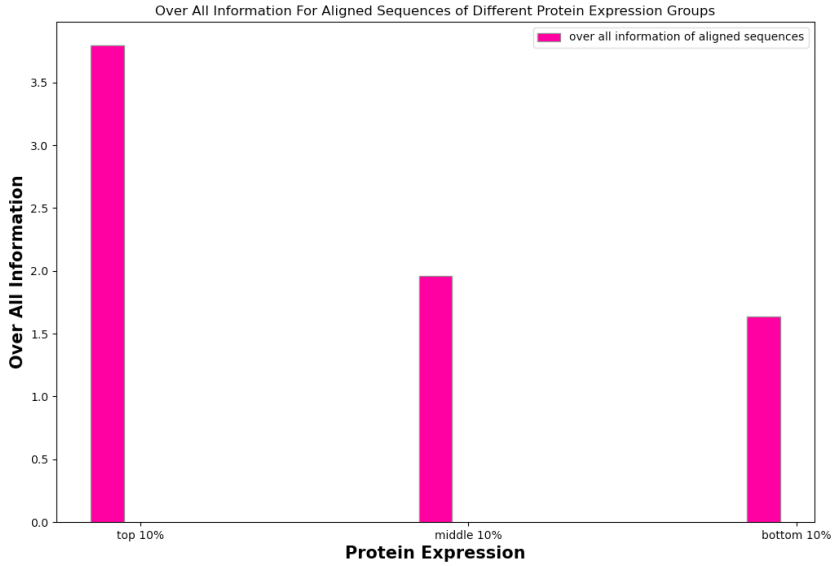

Figure 3: Histogram showing the averaged information scores for each dataset of genes with varying protein expression (top 10%, middle 10%, and bottom 10%). Appendix 1.

This figure shows that genes present in the top 10% of protein expression tend to display what looks to be significantly more information(~3.75 bits) on average than those in the middle 10% (~2.0 bits) and the bottom 10% (~1.8 bits), with those in the middle expressing slightly more information on average than genes in the bottom 10% of protein expression.

More specific histograms display the information content per line for each gene in each data set of varying protein expression. Figure 4 shows the distribution of information scores for genes in the top 10% of protein expression. This histogram appears to be bimodal, though the distribution is not wide, with peaks at ~0.0 bits of information and ~5.0 bits of information, telling us in the sequences of 10 nucleotides before and after the annAUG in the top 10% of protein expression most likely express either ~0.0 or ~5.0 bits of information.

Figure 5 shows the distribution of information scores for genes in the top 10% of protein expression. This histogram appears to be unimodal, though there is a large jump from very few information scores less than 0.0 bits to more than 0.0 bits. This histogram tells us in the sequences of 10 nucleotides before and after the annAUG in the middle 10% of protein expression most likely expresses between ~0.0 and ~2.0 bits of information.

Figure 6 shows the distribution of information scores for genes in the top 10% of protein expression. This histogram appears to be pretty much unimodal and skewed slightly right, and again there is a large jump from very few information scores less than 0.0 bits to more than 0.0

bits. This histogram tells us in the sequences of 10 nucleotides before and after the annAUG in the bottom 10% of protein expression most likely expresses between ~0.0 and ~2.0 bits of information.

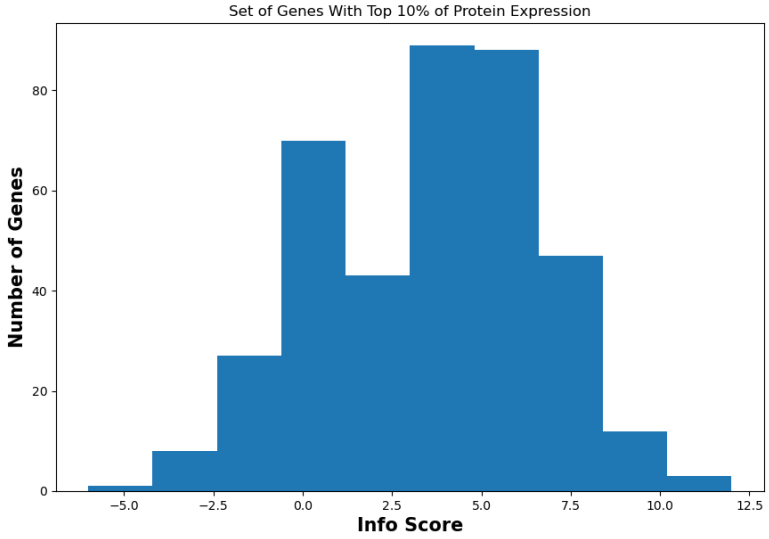

Figure 4: Histogram showing the distribution of information scores for genes in the top 10% of protein expression. Appendix 1.

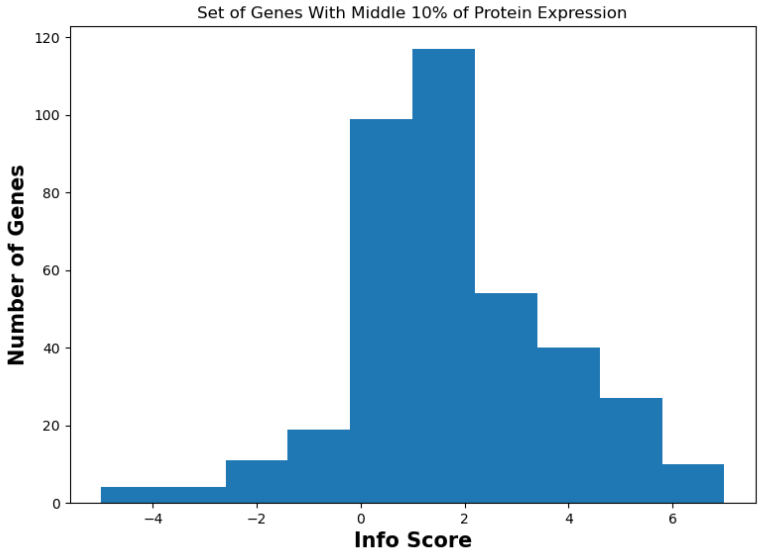

Figure 5: Histogram showing the distribution of information scores for genes in the middle 10% of protein expression. Appendix 1.

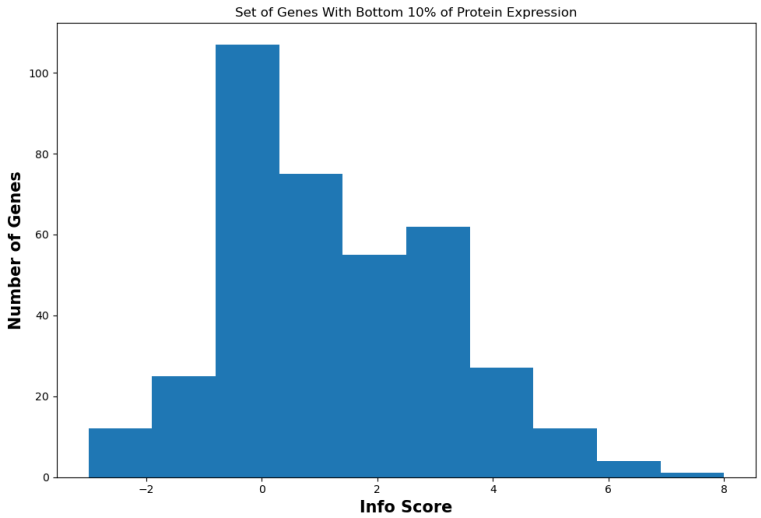

Figure 6: Histogram showing the distribution of information scores for genes in the bottom 10% of protein expression. Appendix 1.

Results from these histograms show that as protein expression decreases, information content also decreases, and that genes with the greatest protein expression also are more likely to have higher information scores.

Predicting Consensus Sequences:

Code from Appendix 3 was used to predict consensus sequences from positions -6 to -1 relative to the translation start site, for genes in all 3 protein expression groups as well as the partial Drosophila data set from Cavener. Running the code outputted the following results:

Data set	Predicted Consensus Sequence
Cavener Drosophila Partial Data set	G–CAAA
Top 10% Protein Expression Yeast	AAAAAA
Middle 10% Protein Expression Yeast	AA–AAA
Bottom 10% Protein Expression Yeast	A– –AAA

Table 1: Shows predicted consensus sequences calculated based on information scores and calculated position weights at each position for the specified dataset. Appendix 3.

From the results of this investigation, it seems that the nucleotides before the translation start site consist mostly of adenines. As we move down the percentage of protein expression, we see that the number of predicted nucleotides decreases as expected since genes with greater protein expression carry more information. Specifically, the predicted consensus sequence from -3 to -1 (the last three characters) match across all data sets.

The results of these analyses are then compared to those of Kozak and Cavener as a baseline. Cavener's Drosophila sequences validate the code used to predict consensus sequences (therefore confirming the outputs of the yeast datasets). The Drosophila consensus is reported as (C|A)AA(A|C) (-4…-1) from Cavener 1987(Source 11). Based on this reported consensus, the results from the above analyses (Table1) seem to match up from positions -4 to -1 (CAAA), confirming that the consensus sequence code is valid, though some disparities must be expected due to the consensus sequence only being predicted from a very small data set (50 sequences/genes).

However, for the yeast datasets there seems to be some disparity in the yeast consensus sequences. For eukaryotes (yeast is a eukaryote), Kozak found the consensus sequence to be CCACC (-5…-1). Disparity seems to lie surrounding the -3 position, which Kozak states to be the most important nucleotide in translation initiation, and the nucleotides surrounding the -3 position have a lesser effect on translation initiation. In the results of the above analyses (Table 1), at -3, an adenine (A) is always present. The disparity between surrounding sequences can be attributed to many things, for example the lack of conservation leading to some missing nucleotide predictions in the consensus sequence, or results from Kozak and Cavener's paper being ~35 years old.

G and GNN Frequencies:

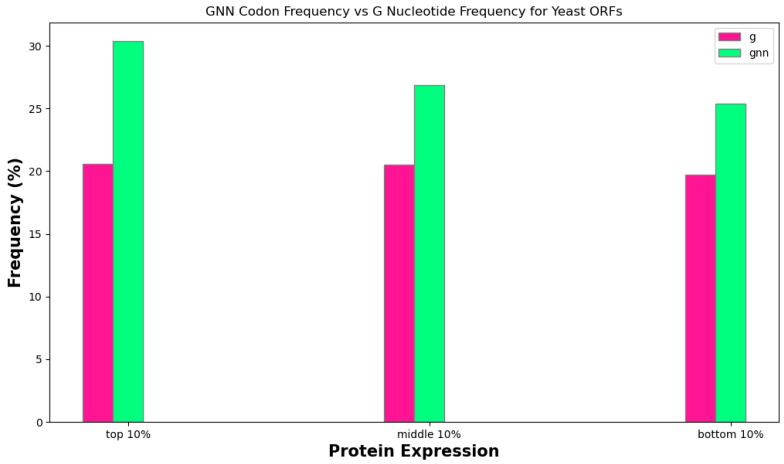

Figure 7: Histogram showing the percentage of G nucleotides across all sequences in varying protein expression data sets (pink), and percentage of codons beginning with G across all sequences in varying protein expression data sets (green). Appendix 2.

In comparing the percentages of G nucleotides in all sequences in a protein expression data set and comparing the number of codons starting with G, the above histogram was yielded (Figure 7). In this histogram, we can see that across all levels of protein expression, the G nucleotide percentage in the 1069 nucleotides upstream the translation start site are all very similar (~22%). However, it appears that in genes with greater protein expression, though the

frequency of G nucleotides does not change, the amount of codons beginning with G (GNN) is greater than in genes with less protein expression.

Based on the data from this analysis, one may infer that the presence of codons beginning with G may have a potential effect on the protein expression of the gene. As protein expression increases, the frequency of codons beginning with G also increases. Though this relationship is not confirmed from our analyses, Gutierrez et al. 1996 find a similar relationship in E. coli genes between translational efficiency and a preference for codons beginning with G.

Infoscores of Sequences with Upstream ATGs and Sequences Without Upstream ATGs:

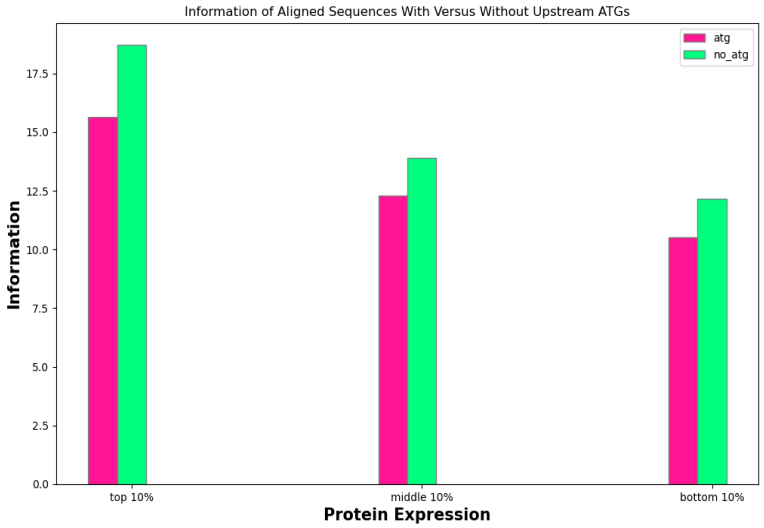

Figure 8: Histogram showing the information content of genes with upstream AUGs in varying protein expression data sets (pink), and the information content of genes without upstream AUGs in varying protein expression data sets(green). Appendix 6.

In the exploration of differences in information content 135 nucleotides upstream the annAUG between genes with upstream AUGs and those without upstream AUGs across varying levels of protein expression, the above histogram was obtained (Figure 8). The results show that genes without upstream AUGs (green) tend to contain more information than those with upstream AUGs (pink) across all levels of protein expression. For genes in the top 10% of protein expression those with upstream AUGs had an info score of ~15.5 bits, and those without had an info score of ~18.5 bits. For genes in the middle 10% of protein expression, genes with upstream AUGs had ~12.5 bits of information, and those without had ~14.0 bits of information. For genes in the bottom 10% of protein expression, genes with upstream AUGs had ~11.5 bits of information, and those without had ~12.5 bits of information.

In considering conservation, the results uphold the concept itself: Genes containing upstream AUGs have been conserved, and thus are expected to be there, resulting in lower information scores. However, it seems perplexing that as protein expression decreases, information scores for genes with upstream AUGs also decrease. One would expect that for

genes with greater protein expression (if the assumption that genes with greater protein expression yield more crucial proteins than genes with lesser protein expression is true), nucleotides upstream from translation would be more conserved as to preserve more important genes, which one would then expect to have a lower information score. However, the assumption about protein expression level may be wrong and further research is imperative to determine why the info scores of genes with AUG sequences decrease across protein expression levels.

Conclusion

This comprehensive study of the sequences surrounding known translation start sites in genes in the top 10%, middle 10%, and bottom 10% of protein expression in the yeast genome via information theory has shown the importance of computational studies to comprehending and understanding molecular biology. Using information theory, this study evaluated information scores from -10 to +10 (without ATG) for individual sequences in a data set and average information scores for data sets of varying gene expression, predicted -6 to -1 consensus sequences for all yeast data sets as well as a supplemental partial Drosophila data set, compared frequencies of G nucleotides and GNN codons across 3 yeast data sets with varying protein expression, and comparing the information content of sequences with and without upstream AUGs across 3 yeast data sets. Upon interpreting the results from all of these analyses, it is clear that information theory is vital in aiding the understanding of biological processes. For example, in the prediction of consensus sequences, it was noted that across all data sets, at whatever protein expression level, even across Drosophila and yeast sequences, which leads us to predict that these 3 nucleotides are conserved, and thus must be important. While this assertion was not completely correct, upon reading Cavener 1987, showed that it was partially correct. Information theory has the potential to revolutionize the study of molecular biology, and can aid in research and understanding of various biological systems and mechanisms.

Role Distribution

A bulk of the project, creating the preliminary investigation functions, was done collaboratively, though some of us focused on varying parts of creating this functioning code. I specialized in the retrieval of data using SQL and the formatting of that data and other specific datasets using python. Maya specialized in the usage of Matplotlib to visualize data. Annika specialized in making sure the code was functional. Upon completion of the preliminary investigation, I focused on some of part a and c, Annika focused on part d, and Maya focused on part e.