Differences in Codon Expression and tRNA Usage in Prokaryotes Due to Stressors

Are there differences in codon expression and tRNA usage in prokaryotes due to stress and what does this mean?

Aishwarya Basude
16 March 2022

Abstract: Relatively recent research shows that tRNAs react to environmental stressors. However, there is minimal tRNA sequencing data available to analyze the shift of tRNA usage in response to stress. Therefore, mRNA-seq data can be leveraged to predict shifts in tRNA as this data gives insight into codon expression in the organism. By analyzing changes in weighted codon counts and gene expression in Streptococcus pneumoniae under control and mild heat-shock conditions, we gain insight into how tRNA expression shifts under stress in prokaryotes.

Introduction to Problem

Organisms respond to environmental stressors by regulating gene expression. While the effect of heat-shock, and other stressors may be well documented at a transcriptome level, studies of tRNA mediated mechanisms are more sparse. The function of tRNAs is to translate codons, or nucleotide triplets that code for a specific amino acid, by delivering corresponding amino acids to ribosome complexes during translation to create proteins. It is logical then that tRNAs would contribute to regulation of proteins and be affected by shifts in codon demand due to stress. In fact, the NCBI finds that tRNAs additionally function as stress sensors and are key for initiating stress response. For example, in eukaryotic cells tRNA have been found to act as signaling molecules, mediating inhibition of protein synthesis due to nutrient deprivation (Huber et. al). Therefore, our goal is to analyze the effect of stress on tRNA usage in prokaryotes.

Although this is an interesting field of research, there is not a wealth of tRNA sequencing data available, which limits our ability to broadly study tRNAs through sequencing. However, there is an abundance of mRNA-seq data, which correlates to tRNA usage. This is because through mRNA expression, we are able to analyze codon usage. As tRNAs use codons as templates, gene and codon expression corresponds to tRNA usage. Therefore, by observing changes of weighted codon counts in prokaryotes in stressed and unstressed conditions, we are able to hypothesize that there will be a change in tRNA usage in these conditions.

Methods

The methods used for the project are separated into 3 major sections: obtaining data (green), passing the data files through pre-existing software packages (green), and parsing and analyzing the data through newly created program scripts (blue). The flowchart in **Figure 1** below illustrates the workflow which was used to obtain results.

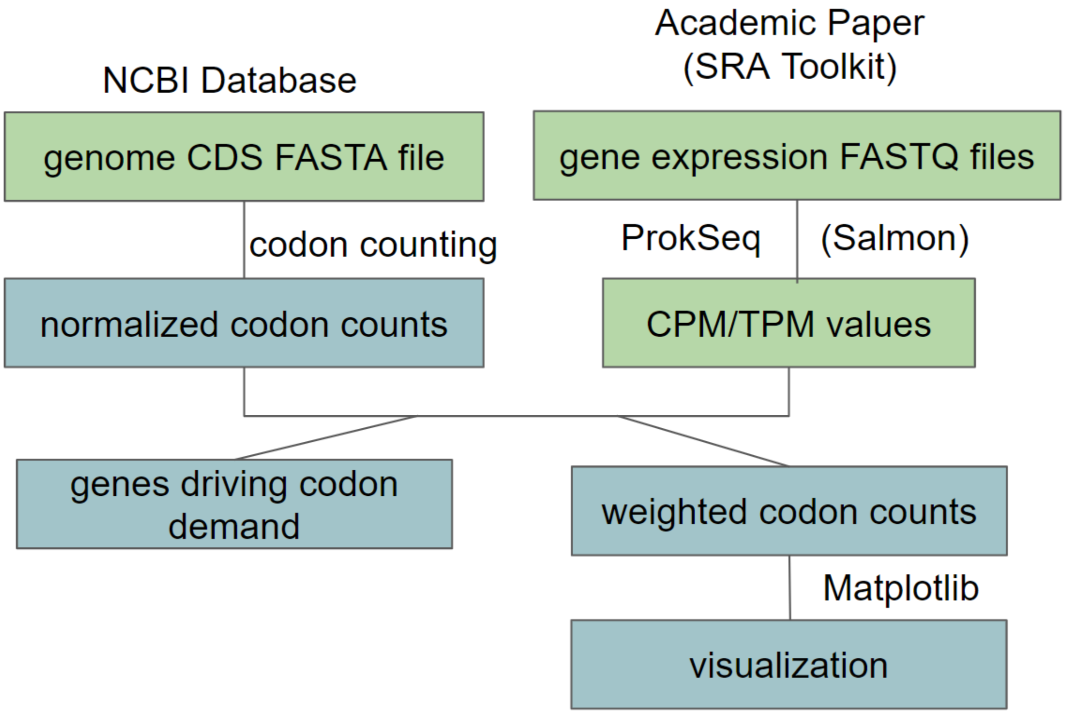


Figure 1: WorkFlow Flowchart: Green: obtained from outside software; Blue: obtained from created python script codonMetrics.py

The results are dependent on two sources of data, the first being FASTQ files on gene expression taken from an academic paper. As we are focusing on the effect of stress on prokaryotic codon expression, we found that the data from the Veening Lab titled "High-resolution analysis of the pneumococcal transcriptome under a wide range of infection-relevant conditions," published by Oxford was suitable to our needs. In their research, the Veening Lab performs quantitative RNA-seq studies on the prokaryote Streptococcus pneumoniae, an opportunistic pathogen. The lab creates pneumococcal cultures under a variety of conditions, including fever, which is subjecting the culture to 40 degree celsius temperature for five minutes when growing in cerebrospinal fluid-mimicking conditions (CSFMC medium). This was what we decided to use as our stressed condition for the prokaryotic data, as opposed to our control, a base culture grown in laboratory conditions (C + Y medium) (Aprianto et al.). This data was obtained through use of fastq-dump under the NCBI's Sequence Read Archive (SRA). Through this software, we were able to obtain the Veening Lab's FASTQ files for the stressed and unstressed Prokaryote data by importing the NCBI accessions that should be downloaded. These files could then be passed to ProkSeq for further analysis.

ProkSeq is an automated RNA-seq data analysis package for Prokaryotic organisms which performs RNA-seq data analysis. It includes a variety of options for output including quantification, normalization, DE analysis, and visualization (Mahmu et al.). This is helpful in more efficiently identifying and quantifying differences in gene expression across multiple samples of bacteria exposed to different conditions. Unfortunately, we were unable to run ProkSeq. Therefore, we decided to use Salmon, a tool that ProkSeq is dependent on, to create a file of gene expression data. Salmon is a tool that runs fast transcript quantification on RNA-seq data, using up to date methods to estimate transcript abundance. It uses a realistic model of RNA-Seq data that takes into account experimental attributes as well as biases commonly observed in RNA-Seq data (Mahmu et al.). To run Salmon, the program requires a FASTA file containing reference transcripts, which is the CDS file, to create a reference file which is used in indexing. It additionally requires a set of FASTQ files that contain reads, one for the prokaryote in unstressed conditions, and one in stressed conditions ("Salmon"). This is used in the quantification process. From Salmon, we obtained one TPM file for Streptococcus pneumoniae under mild heat-shock stress, and one under unstressed conditions.

What we are interested in from ProkSeq is the counts per million (CPM), obtained by dividing counts by the library counts sum and dividing by a million, and Transcripts Per Million (TPM), obtained by dividing the reads per kilobases (RPK) values by the sum of all RPK in a sample multiplied by a million, data obtained from a single output file ("Difference Between"). These values both signify the normalized read counts for each gene, which is helpful in ascertaining if there are any differences in gene expression between stressed and unstressed prokaryotes. We decided to use TPM in our analysis as it uses length normalized read counts, whereas CPM uses depth normalized read counts, and the length of a given transcript affects the number of reads produced ("Difference Between"). If we were to use CPM, longer genes may appear more highly expressed. In any case, in Salmon only the TPM values are output.

The second source of data the analysis is dependent on is coding DNA sequence (CDS) data in the form of a FASTA file. The TPM data from Salmon gives us expression of each gene, whereas the CDS data gives us the sequence for each gene. This FASTA file from the NCBI assembly database,, along with the TPM text files are passed into our python script, codonMetrics.py, to create a grouped bar graph of weighted codon counts, and tables of genes that contribute disproportionately to the shift in codon use. However, rather than genes, the files found for Streptococcus pneumoniae contain which protein product accession number is created from each gene. This is not an issue, as the NCBI includes an annotation table that links each product to its corresponding gene and function. Additionally, each product can be searched manually through NCBI. In **Figure 1**, the steps in blue are those done in the codonMetrics.py script, and the steps in blue are those done previously through the NCBI database, ProkSeq, or Salmon.

The first class in codonMetrics.py is FileReader, which contains the function readFasta() to parse the genomic CDS file, and the function readCpmTpmProkSeq() to parse the text file with CPM and TPM data for the treatment and control. The FASTA file is input into the script through standard input, and the CPM/TPM file is input through the command line with the flag ‘-ps’ to show that you are using ProkSeq data, with ‘-ctp’ or '--cmpTmpFileProkseq', as handled by the Commandline class based on David Bernick’s BME 160 code. The function readFasta() returns the dictionary genomeDict, which holds the gene names as keys and the sequences as values. For example, it could look like {'YPK_0001': ATCATC, 'YPK_0002': CTGATC, 'YPK_0005': ATCCTGCTG}, keeping in mind that the first DNA triplet will always be ATG, which codes for the start codon methionine, and the last triplet will code for a stop codon. Although these triplets are in truth the DNA coding values, we will refer to them as codons from here on out. The readCpmTpm() returns a function that returns the dictionary expressiondict. An example of the dictionary is {'YPK_0001': ['82.1296', '40.1565'], 'YPK_0002': ['141.2988', '197.3905']}, with the format {gene: [sample TPM, control TPM]}. However, with TPM files from Salmon, the user must input one control TPM file with the flag '-tc' or '--tmpControl,' and one treatment TPM file with the flag '-tt'' or --tmpTreatment.' These will be passed to the readCpmTpmSalmon() method, which creates a dictionary with the same format as above. These dictionaries are then passed to the CodonMetrics class to do analysis and stored as parameters in init.

The first function of CodonMetrics is makeCodonDictDict, which creates a dictionary of dictionaries, where the outer key is a gene and each value is a dictionary for a codon key and normalized codon count value. The iterates through each sequence in groups of 3, adding the codon as a new key or adding one to the innermost value as it goes. Then, it normalizes the values by dividing by the number of codons in each sequence so as to not give bias to genes that have longer sequences. An example of this dictionary is {'YPK_0001': {'ATC': 1.0}, 'YPK_0002': {'CTG': 0.5, 'ATC': 0.5}, 'YPK_0005': {'ATC': 0.3333333333333333, 'CTG': 0.6666666666666666}}.

The weightGenes() function is where the dictionaries from the two files come together. As the TPM values are the expression of the gene, and we have now created the normalized codon values for each gene, we find the weighted codon counts by simply multiplying them. To keep the values separate for the prokaryotes under stress and unstressed conditions, we multiply once by the treatment TPM value, and once by the control TPM value and store in another dictionary of dictionaries called weightedCodonDict. An example of this dictionary would be {'YPK_0001': {'ATC': [82.1296, 40.1565]}, 'YPK_0002': {'CTG': [70.6494, 98.69525], 'ATC': [70.6494, 98.69525]}}, with the format {gene: {amino acid: [sample tpm weighted codon count, control tpm weighted codon count]}}.

From these methods, we created a method groupedBar(), which creates a grouped bar graph comparing expressions for stressed and unstressed for each codon in prokaryotes using Matplotlib. It can be called in the command line with the flag ‘-gb’ or ‘--groupedBar’. In order to do this, the method makes use of another method called makeCodonDict, which takes in a dictionary of dictionaries, in this case being the weightedCodonDict, and returns a dictionary of codon expression for stressed and unstressed prokaryote across all genes, codonDict. The codonDict is in the form {'ATC': [223.4284, 237.54700000000003], 'CTG': [141.2988, 197.3905]}, where {codon: [weighted TPM across all genes for treatment, weighted TPM across all genes control]}. The graphing method then plots these values with the codons on the x- axis and the weighted codon counts on the y-axis. The codons on the x-axis are color coded by the one letter code of their corresponding amino acid, which was done by use of a dictionary to convert DNA triplets to amino acids, and a dictionary relating amino acids to colors. The graph can be seen in the results section as **Figure 2**.

The genes that contribute disproportionately to codon use are also of interest. The topTPMDiff() method is the first approach to finding these genes. It can be called in the command line with the flag '-t' or '--topTPMdiff'. The method simply takes in the expressionDict created from the TPM and CPM file, and finds the difference in TPM values, storing them in a list of tuples with the gene and the difference. Then it orders the tuple list by difference value from greatest to least, and prints out a table of genes and differences of the genes with the biggest TPM disparity. The number of genes printed can be changed by changing the range in the print statement for the table. The table for Streptococcus pneumoniae can be seen in the results section with further analysis.

The geneCodonDisparity() method further adds to gene analysis by finding the genes for each codon that have a greatest difference from the mean unweighted normalized codon usage. It can be called in the command line with the flag '-gdc' or '--geneCodonDisparity'. There are two major steps to this method. The first is creating an averageCodonDict in the form {'ATC': 0.6111111111111111, 'CTG': 0.38888888888888884}, where {'codon': "average normalized codon use across all genes"}. This is done by first adding the unweighted codon counts across the genes, and then dividing by the number of genes. Then, a geneCodonDisparityDict is created in the form {'ATC': ['YPK_0001', 0.38888888888888895], 'CTG': ['YPK_0005', 0.2777777777777778]}, where {codon: [gene with highest difference from avg for codon, difference]}. This is done by finding the difference between the average in the averageCodonDict and the true value of the normalized codon count for each gene, saving the gene which has the highest difference for each codon. Therefore, we are left with the gene that contributes most to the disparity of codon usage for each codon. The table printing in this class is further discussed in the results section.

Results and Discussion

The grouped bar graph in **Figure 2** illustrates the differences in weighted codon counts between the Streptococcus pneumoniae that was subjected to fever conditions, or a mild heat shock, and the control. This data is important as we can gauge which codons have a larger than normal difference between the conditions. This allows us to predict that there may be more demand for tRNA to decode those codons in certain conditions. Overall, the amino acids with the biggest differences between the conditions seem to be more downregulated due to the heat-shock than upregulated. This is most likely because when cells are under stress, they may need to shut down cellular machinery to conserve energy. Other codons may be upregulated to compensate for lack of another codon.

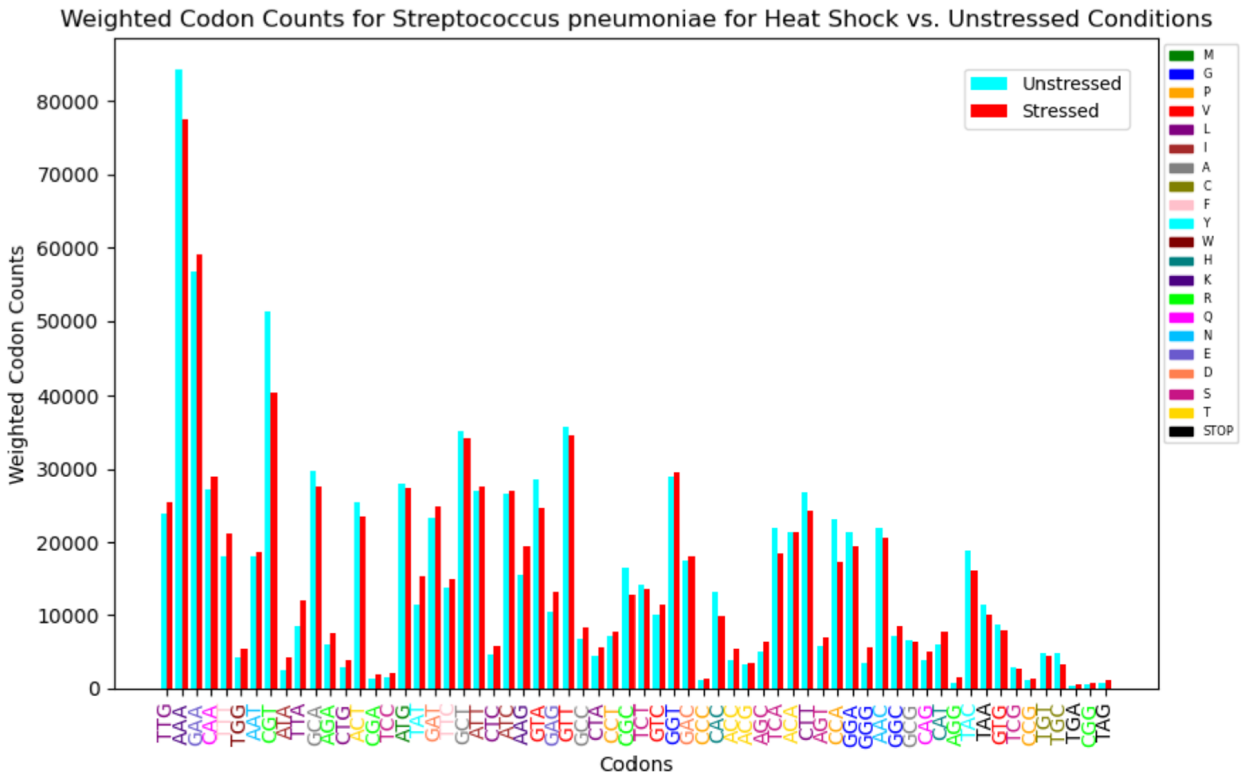


Figure 2: Grouped Bar Graph: Codons Are Color Coded by Single Letter Amino Acid

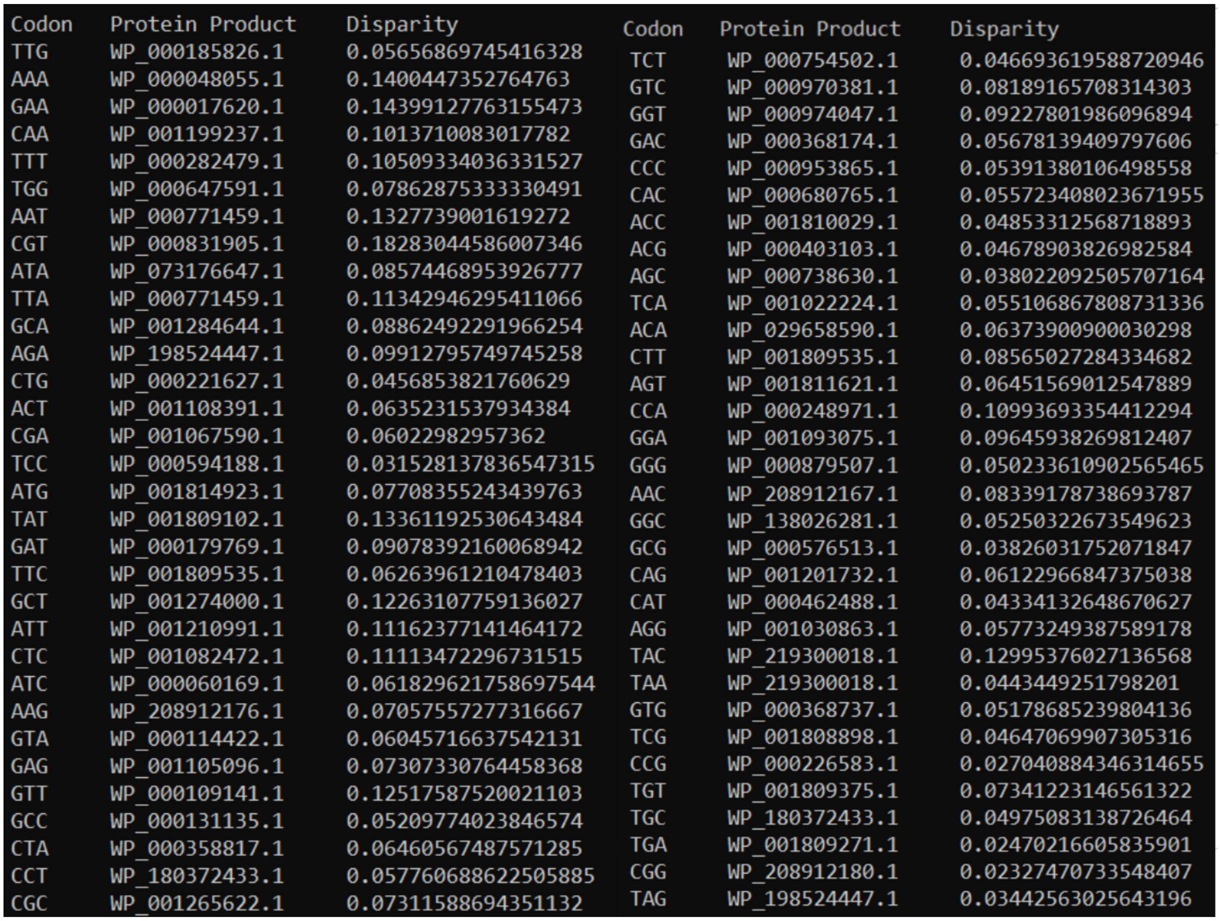
From the **Figure 7**, the immediately outstanding differences are the codons CGT, coding for the amino acid arginine, and AAA, coding for the amino acid lysine. These codons are both highly expressed compared to the others in the bar graph. This may have made them better candidates for downregulation in stress as there is more of the codon present to compensate for decrease in production. An additional point of interest is that arginine and lysine are both positively charged amino acids. In fact, in the amino acids with smaller weighted codon counts, the big differences occur in CGC, which is also arginine, and CAC, which is histidine, another positively charged amino acid. CCA, or the uncharged amino acid proline, additionally has a big gap between the conditions, but this is neither here nor there as uncharged amino acids seem to vary in upregulation and downregulation. This suggests that there is reduced demand for positively changed amino acids when Streptococcus pneumoniae is under stress, thus decreased expression of certain tRNA that decode them. From this, there is little information of the functional role of the regulation, which is why we continued on to analyze which genes play a part in regulation, and the specific changes.

Figure 3 is a table of the five protein products in Streptococcus pneumoniae with the greatest disparity in normalized gene expression between the culture that went through heat shock and normal conditions. These protein products correspond to genes, and give us insight into what functions may be regulated through tRNA. Based on NCBI's feature table for the prokaryote assembly file, the first two products with the biggest difference are 50S ribosomal proteins, meaning it is a protein that binds to an rRNA to form a complex. This complex is especially important to peptide bond formation and protein folding. As heat often denatures proteins, reduction of this protein may contribute to reduced function of proteins. Additionally, as ribosomes are the base of tRNA function, this may affect usage of tRNA apart from those that create the protein. This protein, alongside the second and third proteins with the greatest disparity are downregulated, which coincides with the data in **Figure 2**, as the biggest differences are in reduced codon usage. Meanwhile, WP_000799689, which is greatly upregulated with the stressed data, is a competence-stimulating peptide for Streptococcus pneumoniae. This adds evidence to our hypothesis that genes that are upregulated in stressful conditions are those that compensate for loss of other functions. This peptide is key to a communication system that is critical to the infectivity of pneumococci (Yang). This means that in pathogenic prokaryotes, stress conditions serve to upregulate those genes and subsequently increase use of those tRNA that lead to a higher effectiveness of survival.

Protein Product	TPM Difference	Direction
WP_001808836.1	37056.822076	downregulated
WP_001265622.1	26886.684837	downregulated
WP_000290417.1	25163.807926	downregulated
WP_000570244.1	21286.706203	upregulated
WP_000799689.1	19706.97324	upregulated

Figure 3: Top 5 Products With Biggest TPM Disparity

Although through looking at disparities in TPM, we can analyze the genes which have a great disparity in normalized gene expression as a whole, this does not tell us which genes, or codon products, are creating the greatest shift in specific codon use. Therefore, we created **Figure 4**, a table that illustrates the protein product that causes the biggest disparity for each codon. This is done by finding the mean unweighted normalized codon usage for each protein and sees which have the highest distance from the mean for the codons across all genes. The weightage of proteins from the TPM is not taken into account. This means that this analysis is based completely on the codon count in each protein and which tRNAs are needed to produce those specific proteins. We hope to see a connection with those proteins creating a high disparity in codons and proteins that have a high disparity of expression as a whole between the conditions.



Codon	Protein Product	Disparity	Codon	Protein Product	Disparity
TTG	WP_000185826.1	0.05656869745416328	TCT	WP_000754502.1	0.046693619588720946
AAA	WP_000048055.1	0.1400447352764763	GTC	WP_000970381.1	0.08189165708314303
GAA	WP_000017620.1	0.14399127763155473	GGT	WP_000974047.1	0.09227801986096894
CAA	WP_001199237.1	0.1013710083017782	GAC	WP_000368174.1	0.05678139409797606
TTT	WP_000282479.1	0.10509334036331527	CCC	WP_000953865.1	0.05391380106498558
TGG	WP_000647591.1	0.07862875333330491	CAC	WP_000680765.1	0.055723408023671955
AAT	WP_000771459.1	0.1327739001619272	ACC	WP_001810029.1	0.04853312568718893
CGT	WP_000831905.1	0.18283044586007346	ACG	WP_000403103.1	0.04678903826982584
ATA	WP_073176647.1	0.08574468953926777	AGC	WP_000738630.1	0.038022092505707164
TTA	WP_000771459.1	0.11342946295411066	TCA	WP_001022224.1	0.055106867808731336
GCA	WP_001284644.1	0.08862492291966254	ACA	WP_029658590.1	0.06373900900030298
AGA	WP_198524447.1	0.09912795749745258	CTT	WP_001809535.1	0.08565027284334682
CTG	WP_000221627.1	0.0456853821760629	AGT	WP_001811621.1	0.06451569012547889
ACT	WP_001108391.1	0.0635231537934384	CCA	WP_000248971.1	0.10993693354412294
CGA	WP_001067590.1	0.06022982957362	GGA	WP_001093075.1	0.09645938269812407
TCC	WP_000594188.1	0.031528137836547315	GGG	WP_000879507.1	0.050233610902565465
ATG	WP_001814923.1	0.07708355243439763	AAC	WP_208912167.1	0.08339178738693787
TAT	WP_001809102.1	0.13361192530643484	GGC	WP_138026281.1	0.05250322673549623
GAT	WP_000179769.1	0.09078392160068942	GCG	WP_000576513.1	0.03826031752071847
TTC	WP_001809535.1	0.06263961210478403	CAG	WP_001201732.1	0.06122966847375038
GCT	WP_001274000.1	0.12263107759136027	CAT	WP_000462488.1	0.04334132648670627
ATT	WP_001210991.1	0.11162377141464172	AGG	WP_001030863.1	0.05773249387589178
CTC	WP_001082472.1	0.11113472296731515	TAC	WP_219300018.1	0.12995376027136568
ATC	WP_000060169.1	0.061829621758697544	TAA	WP_219300018.1	0.0443449251798201
AAG	WP_208912176.1	0.07057557277316667	GTG	WP_000368737.1	0.05178685239804136
GTA	WP_000114422.1	0.06045716637542131	TCG	WP_001808898.1	0.04647069907305316
GAG	WP_001105096.1	0.07307330764458368	CCG	WP_000226583.1	0.027040884346314655
GTT	WP_000109141.1	0.12517587520021103	TGT	WP_001809375.1	0.07341223146561322
GCC	WP_000131135.1	0.05209774023846574	TGC	WP_180372433.1	0.04975083138726464
CTA	WP_000358817.1	0.06460567487571285	TGA	WP_001809271.1	0.02470216605835901
CCT	WP_180372433.1	0.057760688622505885	CGG	WP_208912180.1	0.02327470733548407
CGC	WP_001265622.1	0.07311588694351132	TAG	WP_198524447.1	0.03442563025643196

Figure 4: Genes/ Products That Cause Biggest Disparity for Each Codon (DNA Triplet)

From **Figure 4**, we can analyze the proteins that contribute to the perceived change in codon abundance in **Figure 2**. For example, the codon CGT is most changed from the stress condition in Streptococcus pneumoniae in the protein WP_000831905.1. According to NCBI's database, this is another 50S ribosomal protein. Additionally, the codon AAA has the biggest shift due to WP_000048055.1, which is a 30S ribosomal protein that provides the binding site for mRNA and is responsible for monitoring base-pairing between the codon on mRNA and the anticodon on tRNA ("The Structure"). This further provides evidence that in Streptococcus pneumoniae, stress causes an increase in ribosomal complexes, and in turn, tRNA usage. Compared to the other disparities in the table, the magnitude for these disparities are large, meaning that these protein products have an even greater effect on arginine and lysine, and the tRNA that code for them. Although we expected to see some overlap between genes that have the biggest disparity in TPM and genes that have the biggest effect on weighted codon counts of specific codons, they did not for the top five genes in **Figure 3**. However the proteins of interest in both figures had similar function, and therefore, most likely use similar codons and tRNAs.

Conclusion

Overall, we found that there was a change in weighted codon count between Streptococcus pneumoniae that has gone through mild heat shock versus the bacteria grown in normal lab conditions. Stress causes the organism to change its gene expression as genes code for proteins that may be in higher or lower demand due to environmental changes. This change in gene expression affects codon expression, and therefore has an impact on the tRNAs being used. The proteins that the tRNA are used to create additionally may have an effect on future tRNA use for the organism, as proteins contribute to the RNA complexes where tRNA function.

In the case of Streptococcus pneumoniae, a pathogenic bacteria prokaryote, the biggest shifts in codon expression due to heat-shock stress were down regulation for those codons coding for positively charged amino acids. These were especially linked with a decreased expression of proteins functioning in ribosome complexes. This suggests a shift in demand for tRNAs because of a change in the amino acid need that the tRNAs translate, and reduced functioning ribosomes where tRNAs work. Additionally, there is an upregulation in proteins that create higher effective virality of the pathogenic prokaryote, highlighting that tRNA use in stress conditions may shift to compensate for reduction of certain proteins. However, this is simply for one prokaryote in one condition. There is a blatant difference in weighted codon usage in the prokaryote between conditions and patterns in which proteins and codons are upregulated and downregulated. The fact that specific codons are upregulated and downregulated, rather than the all of codons shifting in one direction as a whole in the stressed condition, means that there is promise in further research as different tRNA are reacting differently to the stressors.

We can push forward now that we have a broad exploration of how tRNA usage shifts under stress. Something to look into is using clustering methods, such as KMeans from the sklearn toolkit, to group genes or proteins based on codon usage. With this, we can observe if genes are coregulated, and if pools of tRNA shift alongside the genes. Similarly, we can expand our tables to look at multiple genes. This way we can see the top genes that create a shift in codon usage and may contribute to shifts in particular tRNA. Although there is not much data on tRNA itself, we can predict overall shifts by focusing on the change of codon usage in different conditions.

Acknowledgements

Thank you to Henry Moore and Professor Todd Lowe in mentoring us throughout this project and providing us with an interesting problem to work through.

Works Cited

Aprianto, Rieza, et al. “High-Resolution Analysis of the Pneumococcal Transcriptome Under a Wide Range of Infection-Relevant Conditions.” *Oxford Academic,* vol. 46, no. 19, 2 Nov 2018, pp 9900-10006, academic.oup.com/nar/article/46/19/9990/5079695?login=false.

“Difference Between CPM and TPM and Which One for Downstream Analysis?” *StackExhange,* bioinformatics.stackexchange.com/questions/2298/difference-between-cpm-and-tpm-and-which-one-for-downstream-analysis.

Huber, Sabrina M, et al. “The Versatile Roles of the tRNA Epitranscriptome During Cellular Responses to Toxic Exposures and Environmental Stress.” *NCBI*, 25 Mar. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6468425/#:~:text=tRNAs%20also%20function%20as%20stress,tRNAs%20acting%20as%20signaling%20molecules.

Mahmu, A. K. M. Firoj, et al. “ProkSeq for Complete Analysis of RNA-Seq Data From Prokaryote.” *Oxford Academic*, pubmed.ncbi.nlm.nih.gov/33367516/.

“ProkSeq’s Documentation!” *Read the Docs*, prokseqv20.readthedocs.io/en/latest/.

“Salmon.” *Read the Docs,* salmon.readthedocs.io/en/latest/salmon.html#salmon.

“The Structure of the 30s Ribosome Subunit.” *ESRF*, www.esrf.fr/UsersAndScience/Publications/Highlights/2000/life-sci/LS6.html.

Yang, Yifang, and Yftah Tal-Gan. “Exploring the Competence Stimulating Peptide (CSP) N-Terminal Requirements for Effective ComD Receptor Activation in Group1 Streptococcus pneumoniae.” *Science Direct,* www.sciencedirect.com/science/article/abs/pii/S0045206819303438.

Questions

1. What are the differences between processing prokaryotic and eukaryotic transcripts?
    a. Most rna seq packages assume the input data reflect eukaryotic gene structures.
    b. Prokaryotic transcripts do not have introns and are not alternatively spliced; therefore, using an aligner developed to consider splice junctions often increases falsely assigned reads in the genome

2. What differences would you see in the outputs of prokaryotes and eukaryotes?
    a. Unlike in eukaryotes, under specific stresses, the expression of almost all prokaryotic genes can change
    b. Therefore when doing RNA seq analysis on prokaryotes it's especially important to normalize the data so results are not skewed

3. What exactly is RNA-seq?
    a. Deep sequencing of rna
    b. Shows the presence and quantity of RNA in a sample at a given time

4. What do you expect to find with KMeans clustering and expanding to look at multiple genes?
    a. As we saw patterns in what was upregulated and downregulated, with things such as downregulation of positively charged amino acids, we also expect to see shifts in groups of genes with similar function being upregulated and downregulated.
    b. In turn I would expect to see similar codons and trna groups being expressed similar to each other

5. What is KMeans clustering?
    a. A tool from sklearn where observations belong to cluster with the nearest mean
    b. Minimizes within-cluster variances (squared Euclidean distances)