George Stephenson Assignment #5 March 21st 2023

Dr. Alain Laederach and his colleagues have made a significant contribution to our understanding of how disease-associated mutations can affect the structure and function of RNA molecules. In their paper, Disease-Associated Mutations That Alter the RNA Structural Ensemble, they presented a large-scale analysis of how certain mutations associated with diseases can alter the overall RNA structure[1]. Through their research, they identified regions of the RNA molecule that are highly conserved and discovered that disease-associated mutations occurred in these regions. Furthermore, they found that these mutations could cause alterations in the overall structure of the RNA molecule, which could potentially lead to disease.

I found this relationship between single nucleotide polymorphisms (SNPs), and their conservation to individual loci across the UTR fascinating. Conservation typically entails some sort of evolutionary mechanism is involved, and since this paper notes that these can be found in smaller 5' UTR regions and even larger noncoding RNA sequences. I found this intriguing since these RNA sequences do not encode any amino acid coding information, and thus are not translated, yet they exhibit conservation as if they are highly involved in modification of the RNA ensemble. All of this information piqued my interest in investigating this phenomena on multiple SNPs in the 5' UTR of the FTL gene (199bp) and the ncRNA of the RMRP gene (265bp) respectively, to find any relationships between the two, sequence size, and SNP effect on RNA structure.

Choosing 3 individual SNPs for both RMRP and FTL respectively I composed a Matlab pipeline utilizing the bioinformatics, parallel processing, and machine learning toolboxes in order to produce 18 unique and telling plots for any RNA sequence. First Raw, minimum free energy, and all wildtype sequences are plotted with rnaplot(), with any SNP loci indicated in the plots. This script creates a mutation table of the given RNA sequence and determines the mutation at the lowest comparison score. It then mutates the given sequence with the nucleotide at the lowest score position and plots the circle and dot diagrams of the raw and mutated sequences, highlighting the mutation position in both. This script could be used to understand the impact of a single nucleotide mutation on the structure of an RNA molecule. It could also be used to compare the structure of an RNA molecule to its wild type, or to compare two different RNA molecules. This analysis was then followed by histogram, and kernel distribution plots for both sequences. I finally utilized MATLAB's parallel processing toolkit to streamline my second for loop, allowing me to accurately, and efficiently output a massive heatmap evaluating the structural similarity score for all possible mutations at every position in the sequence. For heat map generation for these sequences, I noticed that utilizing parfor() looping massively increased my processing speed on my workstation's 16-core processor, a powerful tool I most definitely will use in the future to parallelize my future MATLAB scripts and avoid MATLAB's natively sluggish linearized computation on larger batches. This allowed me to loop and score every possible mutation in these sequences in a fraction of the time.

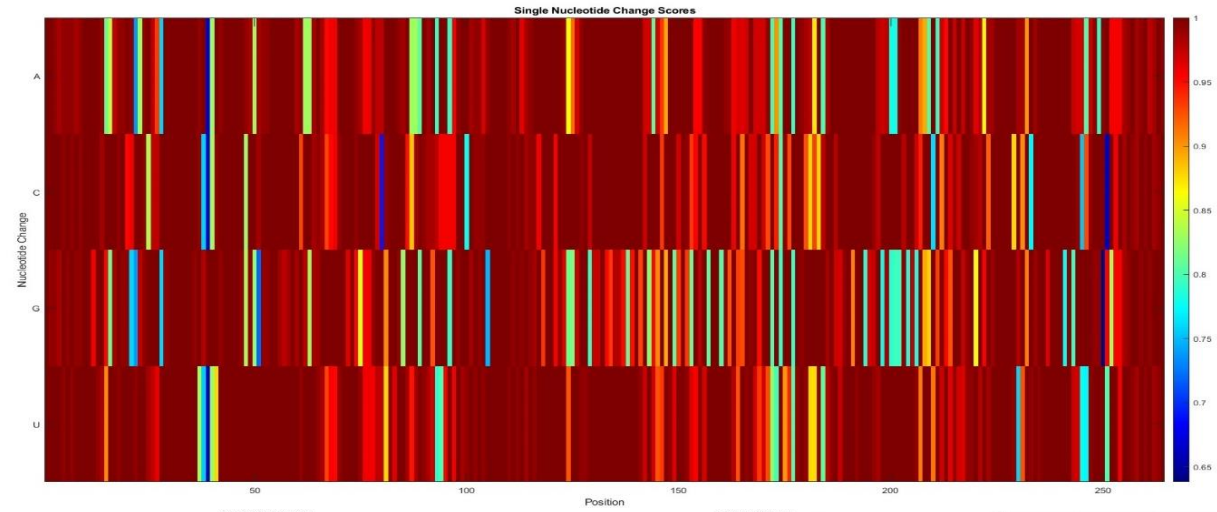
After analyzing the results of my analysis pipeline and cross referencing all 18 outputs, my results clearly supported the hypothesis' made in the research paper. First, from the Histogram of Mutation scores between the smaller 5' FTL UTR region and RMRP's ncRNA; if a change in pairing is caused by a SNP in a smaller sequence, the similarity score between its structure and its reference would be much larger than a similar mutation in a larger molecule. This hypothesis can again be confirmed through cross comparison of my Heatmaps for the FTL and RMRP sequences respectively, for there are much fewer possible mutations that will induce a structural change in larger sequences than in smaller sequences. This is clearly indicated by the fact that more of the cells in the FTL Heatmap are indicating lower similarity scores ('Cooler Colors') then in the RMRP Heatmap respectively, for more SNPs induce mutable change in RNA structure in the smaller sequence. This hypothesis is supported by the principles of hydrogen bonding and van der Waals forces. Hydrogen bonds are formed between two molecules when a hydrogen atom is covalently bonded to a highly electronegative atom, such as oxygen or nitrogen. When two molecules are in proximity, the hydrogen bond will form between them, and the strength of the bond will depend on the distance between the two molecules. Van der Waals forces are attractive intermolecular forces that form between molecules when their electron clouds overlap. These forces are weaker than hydrogen bonds but can still influence the structure and stability of molecules. When a mutation occurs in a smaller molecule, the distance between the molecules is either reduced or increased, resulting in a stronger hydrogen bond or weaker van der Waals forces, respectively. As a result, the structural differences induced by this change would be larger than the effects of a similar mutation in a larger molecule, for their magnitudes of interactions are different. I was also able to computationally determine the similarity score and change in pairing for any mutation in these sequences and found that often, mutations resulting in a change in pairing had lower similarity scores with the reference than those that do not. Again highlighting the differences in stability for these sequences of varying lengths. I was also able to confirm the paper's hypothesis on conservation of SNP loci and their respective magnitudes across all possible mutations at a given position in the reference sequences, from the production and analysis of my two heat maps. It is obvious that there are some 'bands' of a specific color that occur at similar mutation loci, with similar similarity scores. For example, at the head and tail ends of these sequences on the heat map, there are red bands indicating that at these regions, no matter the nucleotide change the SNP will never influence RNA structure, which makes sense, for these head and tail-end regions are rarely involved in structural pairing biologically. By performing sequence analysis in Matlab, I concluded that disease-causing mutations are present in highly conserved regions of the sequence: as these regions typically contain nucleotides that can undergo potentially pathogenic alterations in RNA structure upon mutation, they are typically associated with low similarity scores. This was then confirmed by plotting Wildtype structural similarity scores as points on both the histogram and kernel distribution for each sequence. In both cases, these SNP induced wildtypes had relatively low similarity scores, and thus more conserved regions, compared to all other possible mutations, again confirming the hypotheses made in the paper.

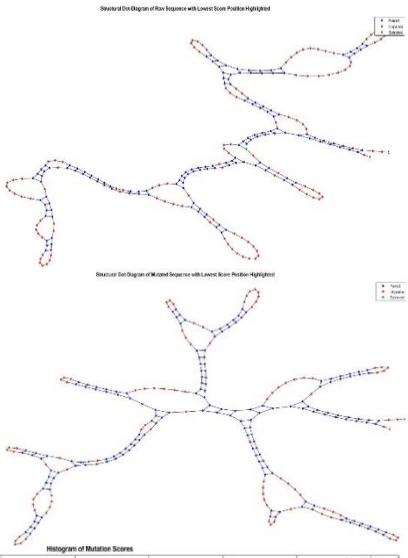
Dr. Alain Laederach and his colleagues have made a significant contribution to our understanding of how disease-associated mutations can affect the structure and function of RNA molecules. Through their publication, they identified regions of the RNA molecule that are highly conserved and found that disease-associated mutations occurred in these regions. This paper inspired me to investigate this phenomenon, for myself, on multiple SNPs in the 5' UTR of the FTL gene and the ncRNA of the RMRP gene, respectively (I ran this pipeline on several other RNAs with known SNPs but did not include them for the sake of my two page limit). After analyzing the results of my analysis pipeline, my results clearly supported the hypothesis' made in the research paper, as well as my own. They showed that if a change in pairing is caused by a SNP in a smaller RNA sequence, the similarity score between its structure and that of its reference would be much larger than a similar mutation in a larger sequence. Furthermore, I found that there is a higher frequency of mutations that can induce a structural change in smaller sequences than in larger ones, which is supported by the principles of hydrogen bonding and van der Waals forces. Overall, Dr. Laederach and his colleagues have provided valuable insight into how SNPs in RNA molecules impact their structure and function, and this has been further supported by my own research.

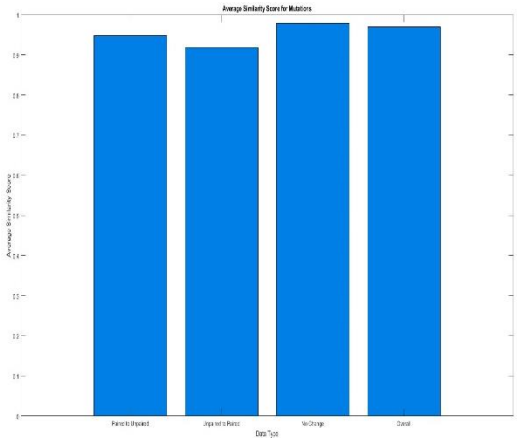
[1]Halvorsen M, Martin JS, Broadaway S, Laederach A. (2010) Disease-associated mutations that alter the RNA structural ensemble. PLoS Genetics, 2010, 6(8)

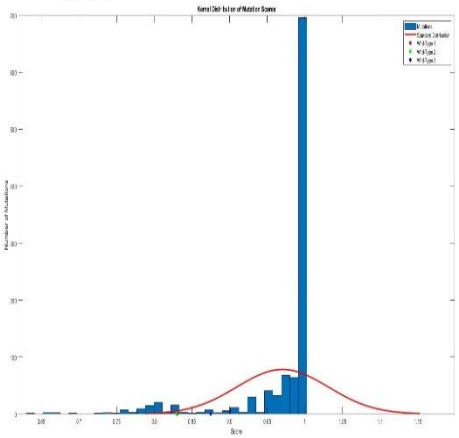
George Stephenson Assignment #5 March 21st 2023

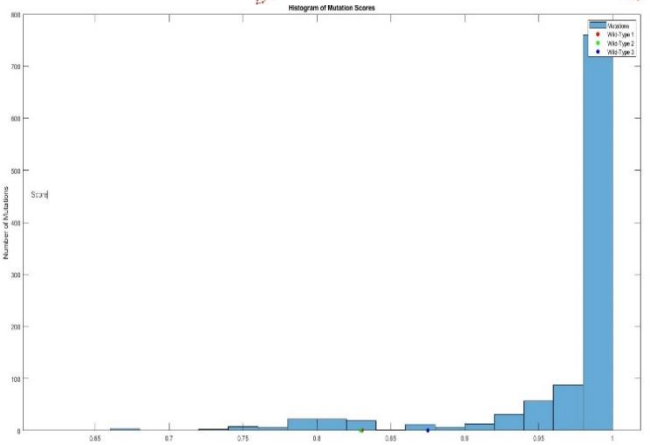
RMRP Figures:











FTL Figures: