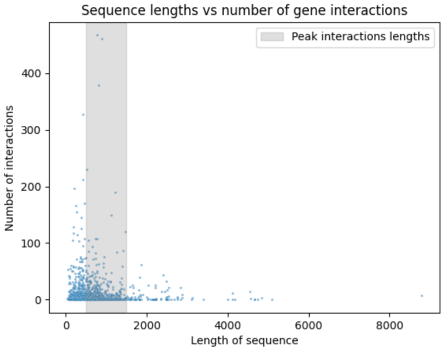
RBIF 100 Final Project Summary Report

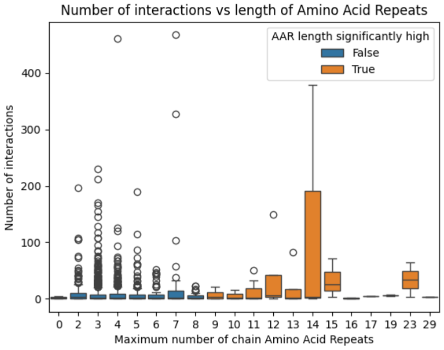
Emily Gelfand

In my analysis, I've investigated the human Insulin protein and its protein-protein interactions (PPI), a critical point to understand about any given protein. My first hypothesis was that the sequence length would impact the number of PPIs a given protein would have – specifically that proteins with longer sequences would interact with more proteins than those with shorter sequences. Additionally, Amino Acid repeats (AARs) can impact protein interactions in different ways depending on how and when they occur, including the common simple repeat pattern generated by DNA slippage[1]. I hypothesized that proteins that with longer chains of simple AARs, or single amino acids repeated multiple times within a sequence, would have more interactions than proteins without these mutations, as these proteins tend to have abnormal functions which may increase their interactions.

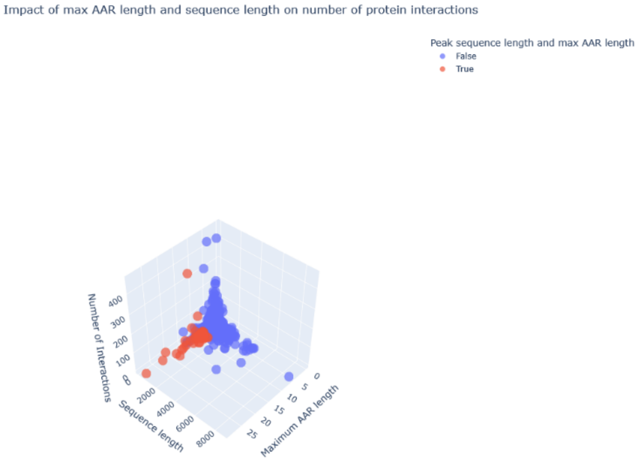
To investigate these hypotheses, I first created a table of all insulin proteins and their PPIs from the Uniprot database using their REST API. Then for each protein I collected its full sequence and gene name from Uniprot. To begin processing the data, I first calculated the sequence lengths and the number of interactions for each sequence. I then plotted a scatterplot of sequence lengths vs number of interactions for each sequence (Fig 1) and found that the distribution was highly left shifted – there was a peak of interactions for sequences with lengths between 500 and 1500 amino acids.



Next, I hypothesized that sequences with longer chains of simple AARs would have more interactions than those with fewer. To calculate this, for each sequence in the table, I checked for sequences with chains of repeated amino acids and noted the maximum repeated sequence length. I also tagged the proteins with AAR lengths greater than 2 standard deviations above the mean maximum simple AAR length. I then generated boxplots of the number of interactions vs maximum AAR lengths for each protein and noted that proteins with significantly higher AAR lengths tended to have slightly higher mean numbers of interactions than proteins with lower AAR lengths.



Finally, to determine whether there was potentially a co-correlation between both the lengths of the sequences and the peak AAR lengths, I created an interactive plotly 3d scatterplot showing all three components in a single place. I first labeled the subset of proteins which both had sequences that fell within the band of lengths with the greatest number of interactions as well as significant max simple AAR lengths; I felt that these had the greatest chance of being proteins with peak numbers of interactions. However, as the figure below shows, that wasn't the case; therefore, something other than those two components is likely governing the peak number of interactions across these interacting proteins.



Citations:

1: Luo H, Nijveen H. Understanding and identifying amino acid repeats. Brief Bioinform. 2014 Jul;15(4):582-91. doi: 10.1093/bib/bbt003. PMID: 23418055; PMCID: PMC4103538.