Brad Broyles

Independent Study Project Summary

We are running a bioinformatics study, trying to better understand transcription activation domains (tAD). Our project could be split up into three sections, but first let me explain the context.

Transcription activation domains are a portion of activator molecules. Activators start the process of transcription, but the mechanism by which it operates has still not been agreed upon even after 30 years of focused research on the topic. Our goal is to test thousands of sequences for functionality as transcription activation domains and use this pool of sequences to further our understanding of activation domains.

The first section of this project is designing a pool of sequences to test for tAD functionality. Our sequences test specific hypotheses, but there are thousands of sequences, so we need to write code that systematically generates our sequences for us. The coding is done in R and hosted on a public GitHub repository named 2019-TAD-Project if anyone is curious. Once our protein sequences are coded for we write code to turn it into an analogous DNA sequence, which is a non-trivial step where codon usage by the organism our experiment operates in matters. Once we have a pool of DNA sequences we are ready to place an order for them and section one is complete.

The second section will be data processing. The way we get functionality data on our tAD sequences is through a high-throughput screen that ultimately results in next generation sequencing across different time points. Many steps are involved with taking raw sequence reads from next generation sequencing and turning them into protein sequences with functionality data. Aligning forward and reverse reads, discarding noise and unconfident reads, and generating functionality scores from multiple time points are just some of the data transformation we will have to perform.

The third section will be the analysis of our sequences and functionality data. We will use machine learning to try to separate the functional from nonfunctional sequences based on the primary protein sequence alone. Something the current field is very far off from doing. Features are numerical transformations of our primary sequence, such as number of aromatic amino acids or molecular weight. Machine learning uses these features to separate the two classes and will tell us which features it used and how predictive they were. By relating which features were used back to the relevant biological context we can infer further on the mechanism of transcription activation domains. And by developing new features we can generate more predictive models and further our understanding of tADs.