Pipeline Project Proposal - Ani Kesanapally
DNA Motif Alignment

DNA motifs are nucleotide sequence patterns that have biological significance, such as conserved regions of genes important for regulation or binding. However, even with their significance, much is still unknown and determining what is or what is not significant is of key interest in genetic research. One of the most crucial steps in this field of research is motif recognition, and the pipeline discussed in this project proposal seeks to approach motif recognition via alignment.

The initial dataset for this pipeline consists of eighteen *E. coli* genes that are of the same length and contain a known motif; this motif varies in sequence and in position, but the general motif can be approximated as TGTGA…TCACA, with the … referring to variable sequence in-between the motif's beginning and end. The steps of this pipeline involve:

1. Encode nucleotide sequences to a mathematical model called *wyk* encoding
    a. *wyk* encoding is discussed in the following paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3070529/.
2. Calculate the correlations between these now encoded sequences with an initial sequence that is also *wyk* encoded.
    a. This step is algorithmic, and is a work-in-progress. However, since these motifs are known, the algorithm drawbacks can be circumvented.
3. Align the *wyk* sequences based on the correlation results.
4. Sum the aligned *wyk* sequences together and divide by the total number of sequences (18, in this case).
5. Convert the "fractional" *wyk* into a nucleotide probability matrix, input this into a logo creator, and output a sequence that has the most plausible nucleotides per position, which ideally, should show the matrix with the highest probability.

The algorithmic-side of this pipeline will be addressed simultaneously but with the assumption that no improvement will be made in this area, the final results for this project will incorporate the result of a potential algorithm as well as the result of a "hard-coded" algorithm, where the correlation/alignment is determined since the motifs are known for this sequence. However, due to this dilemma, this project may also seek to assess the accuracy and precision of this pipeline via the incorporation of a few different parameters, intermediate results, and even a pipeline step such as:

1. The introduction of test/dummy datasets where the motif pattern is easily recognizable by viewers where these datasets seek to test the rigidity of this pipeline/algorithm.
    a. For example, a dataset with motifs with components separated farther away than that of the E. coli set or a dataset with A and T repeats.
2. Presentation of graph outputs detailing the correlations and the squared sums of sequences to detail more regarding the alignment mechanism and results.
3. Computationally parse or interpret results to a DNA logo maker
    a. Currently, a website is used to determine the results: http://www.benoslab.pitt.edu/cgi-bin/enologos/enologos.cgi.