Bryce Hendrickson, Carson Sloan, Molly Meadows

Terence Soule

CS 415 Computational Biology

6 March 2023

**Subproject 2a**

Our experiment will analyze diverging populations that are exposed to the same environmental pressure. Individuals will each have a string of amino acids that is analyzed by a function to determine their fitness.The fitness function will determine fitness by a count of distinct sets of 3 amino acids with a max fitness of 64 from a string of 192 characters representing one individual’s genetic sequence. In order to simulate a diverging population, the initial population will be split (or cloned) into two parts, and individuals will only be allowed to ‘reproduce’ with others in the same subpopulation. After a large number of generations, we expect the groups to have optimized their sequences in obviously different directions. The focus of the experiment will be determining the minimum number of generations after which individuals from each population can be correctly classified. We will perform this by selecting a random subset of each population after specific generation intervals, and running our classification algorithm on this combined subset. The accuracy of the algorithm will be recorded each time.

The classification algorithm will use global sequence alignment, and sort each sequence into one of two groups. Initially, it will take in the amino acid sequences for the first two individuals, and record the alignment score for the two strings. This comparison against the first individual will be repeated for every other individual in the test set, and the four others with the highest scores (meaning their strings are most similar) will have their indices added to a list. These comparisons will then be repeated for every other individual. We will then pass the list of similar individuals to the grouping algorithm, which will return the prediction (0 or 1) for each individual.

We will likely need to make our own scoring matrix to implement this type of alignment. We expect that the alignment algorithm could be very accurate after ten generations, but it depends on the variation in each population, the scoring matrix we choose, and the difference between sequence layouts for each subpopulation. We have control over all three factors, however, by manipulating the fitness, alignment and mutation functions.