Project 2b: Data Generation Description

Dan Blanchette, Taylor Martin, Jordan Reed

March 20, 2023

1 Genetic Algorithm Description

1.1 Encoding

Each individual has a genome sequence comprised of nucleic acids, i.e. 'TCGA'.

1.2 Fitness Function

Our fitness function is simple: it looks at a single character of the genome sequence at a time. If the character is a 'T,' one point is awarded. If the character is an 'A,' 2 points are awarded. If the character is a 'G,' 3 points are awarded, and 4 points are awarded to a 'C.' This means that the maximum fitness score for an individual, based on our genome size of 50, is 200.

1.3 Mutation Rate

We chose to vary our mutation rates based on a high(99%), low(2%), and combined mutation that would randomly alternate from high to low with a 25% chance of switching the mutation rate.

1.4 Population Size

We chose a population size of fifty for each mutation rate (a total of 3). This was to provide a sample size of 50 based on the selection of individuals for comparison in our BLOSUM50 global alignment algorithm.

2 Data Description

The data sample that we ended up putting together is comprised of 50 individuals that were randomly selected from each of the 3 populations described above. We have a file that contains only the individual genome sequences and another file that contains which population each individual came from. This allows us to measure how well our global alignment performed accurately.