Nathaniel Cable
CSE 308
Project 2 Report
08 March, 2023

Implementation Report

For this project, I have coded an algorithm in Java that reads in two nucleotide sequences from a fasta file, aligns them using affine gap scoring in a matrix and traversing that matrix backwards, then outputs the aligned sequences in another file called Alignments.aln. Along with these alignments, the total score for the alignments are also output. These scores are calculated by adding up the following scoring metrics: a match gives a +1, a mismatch gives a -4, opening a gap gives a -10, and continuing a gap gives a -0.5. With various nucleotide sequences used as inputs, the alignments can have many different types of scores. The higher the score is, the more similar the two nucleotide sequences are. With a low score, the aligned sequences likely have many gaps and mismatches. This means the two sequences are very different. Higher scores are often observed when the two sequences provided are derived from organisms of the same species, and represent something like a similar gene. High scores can also be found through gene comparisons across different species, likely due to similar evolutionary needs between these species. Lower scores suggest the sequences have little to nothing in common, and likely do not represent sequences from similar origins.

Nucleotide sequence alignments are a huge deal in the realm of genetics and evolutionary studies. Through the process of aligning sequences of known origin, many important patterns and findings can be drawn up. If the two nucleotide sequences that are aligned are representative of a population of organisms, rather than just being two random sequences, these patterns and findings can speak a lot towards the relationship between specific genome locations and reproductive fitness. Regions of the sequences that are highly conserved and rarely have any differences can likely be used to identify important subsequences that are responsible for vital functionality or structural features for the organism. These subsequences are likely to be important genes, key regulatory parts of the organism, or structural pieces that are under a lot of selective pressure. The fact that these conserved regions exist strongly suggests a large positive role in reproductive fitness, otherwise it would be nearly impossible for the same region to appear among multiple different sequences.

Differences between these nucleotides shown in the alignments, such as single nucleotide polymorphisms, insertions, and deletions, can also be used to draw some relationships between their specific genome locations and reproductive fitness. Many times, differences in nucleotides between organisms of the same species have little to no impact on that specific organism's reproductive fitness. However, by running these alignments across multiple sequences representative of organisms in the same population, the appearance of certain differences coupled with the number of times these differences appear within different organisms can give us an idea as to how they impact reproductive fitness. For example, if certain genes are found to contain a lot of genetic variations within the same population of organisms, it is highly probable that these genes have little to no impact on the organism’s reproductive fitness. If it did, these genes would likely fall into the highly conserved regions of the genome. If one were to align sequences between organisms across multiple generations, differences in these sequences could also give an

idea as to which genes became more impactive on reproductive fitness within the population, and which ones may have begun to not be as important. Of course, all of these conclusions can only be taken with a grain of salt, as there are many, many factors that go into this sort of thing, however, it does provide a good starting point for research.

If alignments are done across sequences that are not representative of a population of organisms and are instead just a set of sequences, a lot of valuable data can be lost. For example, it would become a lot harder to identify highly conserved regions that could provide important insight into the evolutionary traits that could be biologically relevant. It is also likely that some inaccurate conclusions could be drawn from these alignments if the sequences are derived from a specific subset of the population, and these conclusions may not be applicable to the entire population. These biases could cause many inaccuracies, highlighting the importance of representative data. At the end of the day, if the set of sequences being aligned are not representative of the organism's population, any of the conclusions being made from the alignments cannot be applied to the population as a whole and really have no relevance at all in terms of the population.

Now, gene alignment is only one step in understanding how a genome affects an entire biological system. Genomes also contain certain mechanisms for gene regulation that are able to have an effect on the overall system. For example, two mechanisms involved in gene regulation are promoters and enhancers. Each of these regulatory elements are DNA sequences that have an impact on transcription, where a segment of DNA is copied into a segment of RNA. Promoters are crucial segments of DNA typically found near the transcription start site of a gene. They work by acting as the binding site for RNA polymerase to begin transcription. These promoters are responsible for determining the starting point and direction the transcription will follow, as well as impacting the frequency that transcription at that site is initiated. Similarly, and working hand-in-hand with promoters, are enhancers. Enhancers are also DNA segments of the targeted gene, and as the name suggests, they are responsible for increasing the likelihood that that gene begins its transcription initiation. They work by acting as binding sites for activator proteins, and once they bind together, RNA polymerase is essentially coaxed towards the promoter region in order to begin the process.

In order to detect promoters and enhancers experimentally, I have found a process that utilizes chromatin immunoprecipitation, then follows that up with sequencing (known as ChIP-sequencing). It starts out by first treating the cells with formaldehyde to begin what is known as "crosslinking". Crosslinking is when DNA and DNA binding proteins are linked covalently to preserve the interactions they would have in the living cell. Then, using enzymes, chromatin is extracted and broken into small fragments around 500 nucleotide pairs in length. To these chromatin fragments, an antibody is added (for promoters, antibodies against RNA polymerase and for enhancers, antibodies for co-activators) that creates a large DNA-protein-antibody complex. The crosslinking is then reversed through heating, and then the DNA is purified from the complex leaving only the segment associated with the targeted protein. Thus, we are then left with evidence of promoters or enhancers, depending on which antibodies were used.

To detect promoters and enhancers through computational means, one possible solution would be to generate a machine learning algorithm trained on known promoter and enhancer data. One way to get some of this data would be to sequence and compare the genomes of many

different organisms, as promoters and enhancers serve the same function across all organisms. The pieces of the genome that are conserved across different species are likely to contain functionally significant elements like the regulatory ones that are promoters and enhancers. More data on these elements could be gathered based on the specific motifs they have to be recognized by transcription factors. By finding the overrepresented motifs within that set of sequences, you are left with a data set that likely contains promoters and enhancers. Through the use of a model such as a support vector machine (SVM) or random forest, and training it on the data collected by the methods described above, the algorithm could be fed genomic data to make a prediction on the location of these promoters and enhancers. In order to specify if it is a promoter or an enhancer, however, the model will probably need to be trained on more specific data tailored to the one you are specifically looking for.

There are many ways that genes can be recognized by transcription factors. Typically, with a larger population size, it is common for organisms to develop a wider variety of these recognition mechanisms. However, viruses go completely against the grain on this. With their immense population sizes, it would be expected that they could have developed an insanely wide range of these mechanisms, but that is not the case. In fact, most viruses use recognition mechanisms so similar that Glimmer and Genemark can identify their genes. I have found many reasons as to why this is the case.

One reason that viruses use such similar recognition mechanisms is likely due to their evolution and the constraints that are put on it. Obviously, being some of the smallest organisms on Earth, they have some of the smallest genomes on Earth. This size constraint somewhat limits the gene recognition mechanisms they can evolve. Virus genomes are also known to have constraints on evolution altogether, thanks to how compact they are. This compactness usually causes genes to overlap as well as the use of different reading frames. Due to all of these factors, it is likely that viruses had to evolve with gene recognition mechanisms that operate well under these conditions.

Another reason that viruses use such similar recognition mechanisms probably lies within the fact that viruses rely heavily on their hosts' cells for replication and gene expression. This means that their gene recognition mechanisms have had to evolve to match and interact with the host's regulatory proteins and transcription factors. This severely limits the diversity to which their gene recognition mechanisms can evolve. Some viruses can even take over the host cell's promoters and enhancers for their own genes, which also places a constraint on the extent to how diverse their gene recognition mechanisms can evolve.

Finally, as is the case with evolution in general, if something works as it is without causing a downside, it is likely to remain preserved and passed on. In viruses, these similar gene recognition mechanisms have likely been effective and passed on through "generations" of viruses. It is even likely that the similar mechanisms are a result of shared "ancestors" or convergent evolution. Viruses are also under strong selective pressure to have efficient gene selection as it is essential for their survival. If the gene recognition mechanism they have is optimally efficient, then of course it will be passed along in the genome.