Zhenyu Wu
Prof. Brain Chen
CSE-308
3th April 2023

Section 1): a) If the sequence alignments I have computed are representative of a population of organisms, this implies that the sequences I have analyzed are from a group of organisms that share a common ancestry. Therefore, any differences in the sequences between organisms can be attributed to mutations that have arisen over time since the ancestral organism.

By comparing the sequences of different organisms, I can identify which regions of the genome are conserved and which are not. Conserved regions are those that have remained the same over time and are likely to be important for the organism's survival or reproduction. Conversely, non-conserved regions are those that have accumulated mutations and are less likely to be essential for the organism's survival.

Thus, if the sequences I have analyzed are representative of a population of organisms, I can infer which genome locations are important for reproductive fitness by looking for conserved regions across the different organisms. In other words, if a particular region of the genome is highly conserved across many organisms, this suggests that it is important for the organism's survival or reproduction. On the other hand, if a region of the genome is highly variable, this suggests that it is less important for reproductive fitness.

For example, if we are studying a population of bacteria that live in a harsh, acidic environment, and we sequence the genomes of several different strains of this bacteria and align them to compare the differences and similarities between the sequences. Upon aligning the sequences, some certain regions are highly conserved, for example, genes responsible for the bacteria's ability to pump out excess acid from their cells, while some are highly variable. Therefore, we can suggest that the conserved regions of the genome are important for the bacteria's survival in the harsh, acidic environment. Mutations in these regions could potentially disrupt the bacteria's ability to pump out excess acid and thus reduce their ability to survive in the environment and affect their reproductive fitness.

b) However, if the set of sequences I have analyzed is not representative, and just a set of sequences, then it is difficult to make meaningful inferences about the relationship between specific genome locations and reproductive fitness. For example, if I were to compare sequences from a group of organisms that were not closely related, I would likely find many differences between the sequences that are not necessarily related to reproductive fitness. In such a case, it is difficult to distinguish between regions of the genome that are conserved because they are important for reproductive fitness, and regions that are conserved simply because they have not accumulated mutations by chance.

Section 2): a) In addition to the primary sequence of the gene, other parts of the gene can also have a significant impact on the larger biological system. Only one percent of DNA is made up of protein-coding genes; the other 99 percent is noncoding. Some of the non-coding regions play

a significant role in cell function, particularly the control of gene activity. The most common examples are those non-coding regions containing sequences which can control when to turn the gene on or off, acting as regulatory elements. Such elements provide sites for specialized proteins, transcription factors, to attach and either activate or repress the process by which the information from genes is turned into proteins through transcription. Some common types of regulatory elements include promoter, enhancer, silencer and insulator. Promoters which can be found ahead of the gene on the DNA strand provide bind sites for protein machinery that carries out transcription. Enhancers that can be found on the DNA strand before or after the gene they control can control the activation of transcription by providing binding sites for proteins that activate transcription. Silencers located before or after the gene they control can slow down the transcription by providing binding sites for proteins that repress the transcription. Insulators can provide binding sites for proteins that can affect transcription in many ways. For instance,some insulators prevent enhancers from aiding in transcription; meanwhile others prevent structural changes in the DNA that repress gene activity. Thus, regulatory elements located upstream or downstream of the gene can interact with transcription factors and other regulatory proteins to activate or repress transcription.

Non-coding regions such as introns can also play a role in gene regulation and function. Introns,located within protein-coding genes, are transcribed along with exons, but they are spliced out during RNA processing before the mature mRNA is translated into a protein. However, introns can also contain regulatory elements that affect gene regulation. In the article *Introns as Gene Regulators: A Brick on the Accelerator*, scientists revealed that some introns possess the ability to strongly stimulate mRNA accumulation from several hundred nucleotides downstream of the transcription start site, even when the promoter has been deleted. In the experiment, they found that certain introns located in transcribed sequences near the 5′ end of a gene have a large effect on mRNA accumulation, which they named “intron-mediated enhancement” or IME. Through deletion analysis, which has been used to locate important promoter and enhancer sequences, and experiments revealing the location of an expression-stimulating intron, scientists have came up with a strong evidence that mRNA-increasing introns represent a new kind of regulatory element is that their properties are different from the characteristics of enhancers and promoters. Furthermore, through the IMEter Algorithm, they found a strong correlation between the IMEter score of an intron and its ability to increase mRNA accumulation, which supported their idea that intron can increase expression in the absence of a promoter. For example, introns can regulate transcript initiation.

b) One of the experimental approaches to detect the regulatory elements is the reporter gene assays, which are typically used to measure the regulatory ability of an unknown DNA-sequence. In the reporter gene assays, we firstly construct an assay consisting of one of the regulatory elements that we’re interested in and clone it into a vector with a reporter gene which encodes genes with easily measurable activities, such as luciferase. By transfecting them into cells, mRNA is transcripted with the aid of regulator elements, like promoters , and mRNA will be translated to active protein. The amount of luciferase will be dependent on the activity of regulatory elements, which can suggest the presence of such regulatory elements.

For the computational approach, the gene prediction algorithm can be useful to detect regulatory regions. For example the Exon/Intron Identification. The algorithm analyzes the potential gene locations to identify the boundaries of exons and introns, which are the coding and non-coding regions of a gene, respectively. One of the techniques for this algorithm is the splice site prediction, which looks for specific sequence patterns that mark the beginning and end of introns, the splice sites. The algorithm then searches for acceptor splice sites that indicate the beginning and end of introns, respectively.

Section 3) There could be multiple reasons why many common viruses use similar mechanisms for gene recognition by transcription factors despite the potential for a wide range of recognition mechanisms to evolve in viruses. From the biological perspective, there might be three possible reasons.

One possible reason is that there may be selective pressure for viruses to maintain certain conserved features in their genomes to ensure successful infection and replication within host cells. To infect the host cell, a virus needs to express its genes at the right time and in the right amounts. This requires the recognition of specific DNA sequences by transcription factors, which regulate gene expression. If virus initially evolves a wide range of recognition mechanisms for transcription factors. Some of these mechanisms may work well in some host cells, while others may not be as effective. However, over time, there may be selective pressure for the virus to maintain certain conserved features in its genome that allow for successful infection and replication within a broad range of host cells. If the virus were to evolve recognition mechanisms that were too specific to certain cell types, it may not be able to successfully infect and replicate within a broad range of host cells. These conserved features could include sequences for transcription factors that are necessary for proper gene expression during infection. Therefore, despite the potential for a wide range of recognition mechanisms to evolve, viruses may converge on similar recognition sequences over time due to this selective pressure.

Another reason could be that viruses may be limited in their ability to evolve complex mechanisms due to the relatively small size of their genomes compared to their host cells. As a result, viruses may need to rely on a smaller set of sequences that are already present in their genome or that can be easily acquired through mutation or recombination. This could result in a convergence towards similar mechanisms among different viruses, even though they have the potential to evolve a wider range of mechanisms.

From the computational perspective, Glimmer and Genemark are gene prediction tools that use a combination of statistical models and heuristic algorithms to predict the locations of genes in genomic data, which can recognize a wide range of gene features and adapt to different genomic contexts, including viral genomes. The reason that these tools are designed to be adaptable to different genomic contexts is by training their Markov model on a diverse set of genomic data. For example, Glimmer uses a training set of known genes to learn patterns in the genome that are indicative of gene location. This training set is derived from a diverse range of organisms, including bacteria, archaea, prokaryotes, and eukaryotes, which allows Glimmer to recognize a wide range of gene features and adapt to different genomic contexts. Since Virii has immense

populations, it also provides a great amount of testing dara and training data to the model. In addition, the smaller size of virus genome provides greater opportunities that the Glimmer and Genemark can make more accurate predictions against those with larger size of genome.

Reference

Khan, W. by F. (2021, September 22). *The luciferase reporter assay: How it works & why you should use it.* Bitesize Bio. Retrieved April 4, 2023, from https://bitesizebio.com/10774/the-luciferase-reporter-assay-how-it-works/

Rose, A. B. (2018, December 4). *Introns as gene regulators: A brick on the accelerator*. Frontiers. Retrieved April 4, 2023, from https://www.frontiersin.org/articles/10.3389/fgene.2018.00672/full

Singh, N., Nath, R., & Singh, D. B. (2022, May 26). *Splice-site identification for exon prediction using bidirectional LSTM-RNN approach*. Biochemistry and biophysics reports. Retrieved April 4, 2023, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9157471/

U.S. National Library of Medicine. (n.d.). *What is noncoding DNA?: Medlineplus Genetics. MedlinePlus*. Retrieved April 4, 2023, from https://medlineplus.gov/genetics/understanding/basics/noncodingdna/

"*the virus is under increasing selection pressure*". Max-Planck-Gesellschaft. (2021, February 1). Retrieved April 4, 2023, from https://www.mpg.de/16371358/coronavirus-variants#:~:text=Under%20what%20conditions%20can%20such,a%20longer%20period%20of%20time