Marcantonio Soda Jr
CSE408
Project 2 Report
2 April 2023

Section 1:

Sequence alignment offers bioinformatics researchers the ability to analyze and compare the genetic information coded in two or more sequences of nucleotides. By analyzing the alignment of sequence(s) of DNA or protein, researchers can observe similarities or differences in genetic information in an effort to gain insights into the evolution and function of different genes and regions of an organism(s) genome. This concept has implications in understanding the relationship between specific genome locations and reproductive fitness if the computed alignments are representative of a population of organisms. Any conservation or variation in specific genome regions can provide some insight into the relationship between said regions and their contribution to reproductive fitness in relation to the aforementioned population. Before unpacking the revelations that can be realized, it is important to have a good working understanding of what reproductive fitness actually is.

Reproductive fitness can be described as an organism's (or population’s) ability to reproduce in their environment, causing it to produce successful offspring that they, themselves are able to reproduce in the very same or similar environments. It can be thought of as a measure of the success of a population in passing its genes along to future generations. Reproductive fitness is determined by a variety of factors, genetic and environmental. Genetic factors include traits such as physical aspects of the organism’s appearance, how resistant it is to disease, the expression of behavior, etc that affect the organism’s future reproductive success. Environmental factors include the availability of food, risk of predation, overpopulation, etc that affect the same thing. Reproductive fitness is arguably the most important quality for a population to have because without it, the population will cease to exist unless it further adapts, and quickly. Knowing this, it stands to reason why researchers would be interested in understanding it further through sequence alignment.

Sequence alignment is able to provide a lot of insight into the relationship of different genome locations and reproductive fitness by successfully identifying the regions of the genome that have been conserved or have varied across different individuals or populations over time. Typically, the more conserved regions are more crucial in increasing a population’s reproductive fitness. This makes sense. A genome location that has not changed over time in one species or over multiple similar species must not have changed for a very good reason. It is likely that when that region was changed via mutation in the past, the organism did not survive long enough to reproduce. Hence, conserved regions contribute to reproductive fitness more than variable ones.

If, on the other hand, the sequences being aligned are not representative of a population of organisms, it becomes difficult to draw conclusions in regard to the relationship between genome locations and reproductive fitness. In such a case, any conservation or variation in different genome regions cannot be confidently tied to common selective pressures and may be due to other unrelated factors. I suppose that consistent conservation across many unrelated organisms could shed some light on genome locations that are necessary for sustaining life in general. It is therefore not completely meaningless to align sequences present in organisms of unrelated populations because it could allow researchers to make conclusions about genome locations and reproductive fitness on a much broader scale. Afterall, every organism on Earth is subjected to selective pressures that are similar to some degree.

Section 2:

Sequence alignment does provide some important insights regarding the function and structure of individual genes, but sequence alignment in and of itself does not paint the whole picture on how a genome influences a biological system. It is only one piece of the overarching puzzle. Gene regulation and gene expression play critical roles in larger, more complex biological systems. Experimental and computational approaches do exist to detect and classify these elements.

In particular, gene expression plays a critical role in determining how and when genes are expressed as well as how they interact with other genes in an effort to contribute to larger systems. Gene regulation refers to the different mechanisms that exist to control the expression of genes. You can think of it as the mechanism by which genes are turned on and off. Gene regulation is critical for the function of cells as it affects cells' ability to respond to various changes in its environment. Gene regulation is affected by a gene's promoters, enhancers, silencers, and transcription factors by affecting the binding of RNA polymerase and other proteins to the DNA. Transcription factors are proteins that bind to specific sequences which activate or inhibit the transcription of a gene. They bind to promoters, enhancers, or silencers to regulate gene expression at the lowest level. Promoters are specific sequences that act as binding sites for RNA polymerase and transcription factors. These promoters determine exactly where transcription starts. Enhancers are sequences that are a binding site for transcription factors which enhance the rate of transcription of a gene. Silencers are sequences that are also a binding site for transcription factors that inhibit the transcription of a gene. The presence of these regulatory elements are hard to detect with sequence alignment because they may not be conserved across different species, or even within the same species. They often interact with other cellular molecules and proteins in complex ways that cannot be realized through sequence data alone. Other approaches, experimental and computational, have been developed to identify them.

The most common experimental method by which to detect the aforementioned regulatory elements is chromatin immunoprecipitation followed by high-throughput sequencing (ChIPseq), which I discussed in a previous paper. ChIPseq is a technique that is used to determine the specific interaction between DNA and proteins. More specifically, it is used to determine what region in the DNA a protein binds to. This information is able to help scientists better understand the specific function of the protein in question. In the nucleus of a cell, many proteins are attached to the chromatin which surrounds the DNA. The proteins in question are constantly binding and unbinding, so a cross-linking agent needs to be applied to ensure that the proteins stay bound. The chromatin is then lysed and cut into small fragments. We are now left with many DNA fragments with proteins attached to them. An antibody must be applied which will separate the binding between the protein and the DNA of only the protein of interest. The still protein-bound DNA and lone proteins separated from the rest of the solution through immunoprecipitation, leaving only the DNA that was recently separated from our protein of interest. Now that the DNA of interest has been separated, the standard DNA microarray technique can be used to hybridize the DNA samples in the same method described above. The DNA fragments from the ChIPseq process are dyed and applied to the DNA microarray which is full of designed probes. The DNA Microarray allows us to see which specific part of the genome the DNA belongs to. Therefore Chromatin Immunoprecipitation combined with DNA microarrays can be used together to determine the interactions between specific proteins and DNA. This is the perfect experimental approach to detect regulatory elements because their elements are typically bound by proteins like RNA polymerase.

One of the most popular, but complex computational methods to detect regulatory elements is motif-finding. Motif-finding is used to identify short sequences that are often binding sites for transcription factors. First, a set of sequences that are known to contain regulatory elements of interest must be collected and fed to the computer. This is typically done using ChIPseq or a similar experimental technique. Then, regions of those sequences that are likely to contain the regulatory elements must be identified. This is done by looking for features such as histone modifications or open chromatin. The identified regions can then be scanned for short, conserved DNA sequences (motifs). The found motifs may be specific to one regulatory element or they may be shared across multiple. Next, the statistical significance of each motif must be calculated to determine which are likely to be functionally relevant. This is usually done through an algorithm called motif enrichment. The final output of motif-finding is a set of motifs along with their corresponding scores and locations within the input sequences.

Section 3:

Genes can be recognized by transcription factors in many ways, yet many of the most common viruses use similar mechanisms; similar enough, at least, that Glimmer and Genemark accurately identify their genes. This very interesting observation can be easily answered by understanding the “life”cycle of a virus.

A virus has a unique life cycle compared to that of similar constructs. They are not cells, nor are they alive, but they do contain genetic information. This information represents the instructions for creating new viruses. Viruses cannot reproduce on their own. They are intracellular parasites that infiltrate cells and hijack host-cell machinery in an effort to create more of itself. The viruses produced by the hijacked cell are expelled and go on their way to repeat this vicious cycle.

The fact that viruses depend on the host-cell machinery for reproduction illuminates why viruses have similar mechanisms compared to genes. They evolved recognition mechanisms that are similar to those of the host cell to maximize their ability to utilize the host resources. This allows viruses to easily take over a cell with minimal need to have developed its own machinery, allowing the virus to remain as small and simple as possible, which serves to its advantage.

Similar recognition mechanisms can also help the virus to evade detection by the immune system. When a virus infects a cell, an immune response should be triggered by the infected cell, causing the cell to self-destruct or the immune system to eliminate it. However, if the virus has similar recognition mechanisms, it may be able to evade detection and continue to replicate without contest.

Finally, by using similar recognition methods as the host cell, the virus can reduce the metabolic burden on the host-cell. If the cell would have to develop new recognition mechanisms, it would cost energy that could otherwise be spent creating more viruses. Put simple, it would be inefficient. It is no wonder viruses evolved this way.