Nathaniel Cable

Professor Brian Chen

CSE 308: Bioinformatics Issues and Algorithms

14 May, 2023

Project 3 Implementation Report

Typically, recidivism is a term used to describe the act and tendency for a convicted criminal to repeat their criminal activity even after experiencing the punishment associated with their crimes. When speaking about the treatment of HIV, however, we can use this term to describe the common experience among HIV positive individuals to be reinfected with the disease. This was common before the development of Highly Active Anti-Retroviral Therapy (HAART), and is still a relevant issue among medicinally non-compliant HIV positive individuals. This is common for multiple reasons. The first reason is that HIV weakens the immune system. Those that are positive for the virus have their immune systems weakened over time, making the risk for other infections a lot higher. Due to this, if an HIV positive person comes into contact with a different strain of the virus, it is highly likely that they will contract that disease. The second reason is that HIV is typically passed to someone through either unprotected sex or the sharing of needles, and it can be hard to realize you have the virus at first. When you are someone that regularly does one of these two things, especially with multiple different people, it is very likely that you could contract another strain of the virus this way. Finally, it is also likely that the strain of the virus a person initially contracted can mutate within their body. HIV is a virus with a very high mutation rate, and it is possible for a new strain to develop within someone while they are infected. This is bolstered by the fact that HIV can recombine when two different strains infect the same cell, creating a new strain out of the combined viral DNA.

We are now to imagine a hypothetical patient infected by one strain of genetically identical HIV capsids. The patient does not seek treatment for several years after the initial infection, and is never exposed to another strain of HIV. After these years are passed, and if we could sequence the RNA genome in every HIV capsid in the patient, what could we expect about the number of genetically identical HIV capsids? Clearly, we would expect a large amount of diversity within the HIV genome with few genetically identical capsids. This is due to some of the aforementioned characteristics of the HIV virus. To start, HIV has a high replication rate coupled with a high rate of error in replication, leading to a very high mutation rate. This leads to slightly different versions of the virus after every single replication, and after multiple replications, these mutations can add up to form what could be considered an entirely new strain of the virus. If this were to continue for many years without treatment, it is likely that the infected individual is a carrier of multiple different strains of the virus. As mentioned previously, HIV is capable of recombination, which occurs when two different strains infect the same cell. When this happens, and viral replication is happening within the cell, the DNA of the two strains become mixed together and can combine to form an entirely new strain of the virus. If this patient is going untreated for years without exposure to another strain, this is still a likely

occurrence due to the high mutation rate. Of course, there is also the matter of environmental and selective pressures such as the person's immune system, and in turn, natural selection. The immune system is always going to be fighting the virus, so it is advantageous for the strains to diversify, leading to more advantageous genes such as faster replication or overall resistance. So, again, if this situation were to occur, it would be certain that we will see a high amount of genetic diversity between the capsids, and therefore a low number of genetically identical ones.

Now we are to assume that it is the late 80s, and the patient is one of the fortunate few to receive a course of Zidovudine (I am assuming this is still after all of that time has passed). If we were to sequence the RNA genome in every HIV capsid in the patient, how would the number of strains afflicting this patient differ from the situation discussed previously? Zidovudine is an antiretroviral medication classified as a nucleotide reverse transcriptase inhibitor, or NRTI for short. NTRIs like Zidovudine work by targeting reverse transcriptase, which is an enzyme that retroviruses like HIV use to convert their RNA genome into DNA. These drugs then target that enzyme, inhibiting its functions. Zidovudine is then integrated into the DNA of the virus, as it is structurally similar to thyminide, a nucleoside used to build the DNA, and the virus uses it to construct its DNA. Once it is bound to the chain structure, it acts as a chain terminator, stopping the construction of the DNA. This stops the virus from successfully replicating, and reduces the overall spread and replication of the HIV in the body. Due to the fact that Zidovudine does not directly kill the virus, and that all of this time has passed, giving the virus enough time to already greatly diversify, we would probably expect a little less diversity, if any less, than the original situation. Since it slows down replication, it may give the immune system a chance to catch up by killing some of the weaker strains of the HIV virus, but most of the damage has already been done by this point. If, however, the patient took this drug immediately after being infected, we would expect a much lower diversity among the capsid genomes. This does not mean that there still would not be different strains that came through generations and generations of mutations, but the overall mutative effect on the diversity would be significantly slashed. Of course, if this is the only drug being used to treat the patient, it is still possible for the virus to mutate enough to find a resistance to the drug, rendering it useless and putting us back at square one.

An interesting topic would be to observe a phylogenetic tree of the HIV genomes within the individual both before and immediately after the treatment. If we looked at a tree directly before the person underwent their treatment, it would probably be a very wide tree with many different branches, where each branch represents a different strain. Since the virus had a lot of time to undergo mutations and recombinations, the tree is going to be very diverse. Compared to a phylogenetic tree made immediately after the treatment was finished, the tree from before is likely to be slightly more diverse. In this new tree, we are likely to see an increase in frequency of certain branches appearing, and a sharp decrease in the others. These higher frequency branches are all strains of the virus that have an advantageous mutation that makes them more resistant to the drug. This means they are still free to replicate, with the Zidovudine having less of an impact on their replicative abilities. This is the major reason that HAART applies several drugs in parallel. It is much more unlikely for HIV, despite its high mutation rate, to develop a resistance to multiple different drugs at once. This is also why they are given in parallel and not sequentially, as giving them sequentially could provide enough time for the virus to develop a resistance to each drug individually and in order, rendering the treatments useless. Even though HAART is effective, medicinal noncompliance is still a major problem and for multiple reasons.

First and foremost, these antiviral drugs tend to have a lot of side effects that can become a pain, especially since the virus never goes away and they will have to continue treatment for life. This also brings into light the cost of these different drugs and how accessible they are for patients. Some people cannot afford these medications, or just do not have access to all of the ones they need. On top of this, since HAART requires multiple drugs in parallel, sometimes it does not just come as a single pill, and instead requires a complex treatment regimen that can become annoying for the everyday person. All of these factors can turn patients away from the treatment, choosing to just live life as it is with the virus, making it even more dangerous for everybody else.

An interesting topic to discuss is the effect of incorporating protein shape into the alignment of protein sequences, especially for very different sequences, and how this could yield additional information for reconstructing evolutionary history. This is because proteins with very different sequences (e.g. 25% sequence identity) have been known to exhibit very similar "folds", or overall shape, while proteins with very different folds never exhibit very similar amino acid sequences. At the end of the day, the shape of a protein is a very important aspect of that protein's function. Similar to sequence alignment, there exists an alignment known as structural alignment that could provide more insight than we could get from just sequence alignment. One example of this is the existence of remote homologs. Remote homologs are proteins that have both similar structures and functions, but are hard to find similar nucleotide sequences between. These proteins would likely not be denoted as closely related through sequence aligners, but would be marked as strongly related through structural alignment. This occurs as a genome changes over time, and the protein's structure is preserved, but the nucleotide sequences are different. Structural alignment can also help us to understand the evolution of protein functions, as their structures have evolved to align more with other sets of proteins that are responsible for a set of specific actions. Another big thing this could help with is convergent evolution, where two unrelated species evolve similar features due to undergoing the same environmental pressures. This would explain similar protein shape and function, and a difference in nucleotide sequences within that protein. Nucleotide alignment would never be able to pick up on this, unlike structural alignment. The only difficulty with implementing these strategies is that they are extremely difficult to do computationally compared to sequence alignment, which is already quite difficult.

Let's say we have a vertical phylogenetic tree with the root at the top and the leaves at the bottom, and that if a horizontal line is drawn anywhere between the root and leaves that there is a certain set of amino acids that is conserved. This draws some interesting questions about those amino acids. First, what could a set that's conserved only in a specific subtree but not the others mean? These amino acids likely indicate functional adaptations that are only relevant for that specific line of the tree. These could be for a protein that's functions this lineage has evolved to need, such as a combatting a specific environmental pressure that they help to combat that no other lineage relies on, such as high intake of sunlight or cold weather. If these acids were found to be located near active sites, that means this protein is essential for survival. Active sites are where proteins function, and therefore, these amino acids are needed for the protein to function in the way it was needed to. Any change to these active sites could drastically alter the protein's function, and since this is not the case, it can be concluded that this protein is essential for this organism's survival.