Jonah Tuckman
Assignment 2
Modeling Molecular Evolution
April 29, 2020

Findings of Assignment 2 — Modeling Molecular Evolution

Introduction:

When putting together this project, our chief goal was to compare two levels of nature based alignment. The first being using a strategy of purely randomization, meaning taking the alignment of two codons and from there randomly calculating their alignment using a rate of 15% as our comparison rate. Our second strategy was then based on a higher level of understanding and knowledge of the goal at hand. We have an understanding that the third codon is more often misaligned than others may be through evolutionary changes and thus our comparison structure comprised of using the same comparison rate of 15% but solely using this when comparing the third codons of different sequences. This allowed us to take a further look into this question using not just the randomization in our sequence analyzation but rather using our background knowledge of evolutionary trends and thus making a more detailed model through this knowledge.

What I learned of the stages of molecular evolution through this modeling was that we do indeed see a higher rate of molecular evolution in third codons than we do elsewhere. We were able to support this prior understanding through a look at the data as we saw higher rates of evolution when looking at the third codons solely rather than when we analyzed all codons at a random level. This is important information that we were able to solidify as it may narrow our search and intensify our understanding when comparing the sequences of evolutionary beings, as we did with this software.

What I found most fascinating about this experiment was this finding of the third codons being the most commonly mutated as apposed to others in sequences of beings. Mutations within genes exist at a higher rate within the third codon and this then becomes something that sparks questions for me as a non-biologist, as someone more versed in the software side of this project rather than the underlying biological understandings. I wonder why the third codons find a higher mutation rate, I wonder if this is biologically beneficial or if it takes part in Darwins theory of survival of the fittest and whether we are able to track this.

When species adapt over time to a certain situation, this is the showing of a species adapting to certain genetic variances in order to survive. Those who are able to adapt in this evolutionary rate are those who survive. We would like to look at the genetic mutation in order to understand the evolutionary adaptation that species have taken and how that has affected their survival versus other species.

The goal of this experiment of modeling molecular evolution was to understand the existence of this trend of the third codon of a sequence being more often mutated than others. Our way of understanding this was to do what all good comparisons do (something we as computer scientists to a great deal with algorithm runtimes) and compare a randomized algorithm to one with the algorithm correctly crafted based on the underlying theory. We do this to show different speeds of algorithms when sorting

through data randomly vs with a specific algorithm. Here it was done to understand natural molecular evolution as either a random event, or one with a further background.

Results:

In our initial strategy of randomization we found there to be a mutation rate between the collected sequences of roughly 68%. Below is a collection of some of the results when running the experiment with this strategy 100 times.

```
Trial, % identity

1, 0.712050078247
2, 0.676056338028
3, 0.713615023474
4, 0.702660406886
5, 0.696400625978
6, 0.674491392801
7, 0.679186228482
8, 0.707355242567
9, 0.679186228482
10, 0.719874804382
11, 0.668231611894
12, 0.677621283255
13, 0.688575899844
14, 0.66510172144
15, 0.699530516432
```

When running the experiment with the same sequences as above, also 100 times, yet only comparing the third codons of sequences, we found there to be a higher rate of mutation than found through solely randomization. We found this mutation rate to be closer to 84%.

```
Trial, % identity

1, 0.862284820031
2, 0.835680751174
3, 0.849765258216
4, 0.865414710485
5, 0.846635367762
6, 0.857589984351
7, 0.85289514867
8, 0.857589984351
```

```
 9, 0.8779342723
10, 0.885758998435
11, 0.848200312989
12, 0.846635367762
13, 0.835680751174
14, 0.849765258216
15, 0.845070422535
```

Then when looking at the mutation rates in different codon sections, we do indeed see a higher rate in position 3 than in 1 and 2.

```
Average diversity by nucleotide position within codons:
                Codon position 1: 1.4
            Codon position 2: 1.26973684211
            Codon position 3: 2.20131578947
```

Findings:

As stated above briefly in my introduction, we were able to find a higher rate of mutation in the third codon than when looking randomly at the sequence as a whole. This is important information that we were able to solidify as it may narrow our search and intensify our understanding when comparing the sequences of evolutionary beings, as we did with this software. We see above that the average rate of diversity within the codons for 1 and 2 is similar, and then for position 3 is nearly double the prior found rates. This shows a significant difference between the two. This significant difference is exemplified through the data, and what it means is that we can be reasonably comfortable making the assertion that codons in position three have a higher rate of mutation than those in position one or position two.

If I were to criticize the results, I would worry about the scope of the data collected and the number of tests done. I would like to see this experiment done with a comparison of more than two sequences from different DNA collections. I would also like to see a more advanced average system in which the test is being run more times and averaging results.

Next Steps section:

When thinking of the SARS-CoV-2 virus and its adaption into a human host, there are a variety of things that we can pull from this experiment into that topic. One of these would be the idea of proteins evolving in terms of interacting with the hosts in question. Also, viral proteins that might change amino acid sequence if the virus was adapting to human hosts.

To track something of this sort as a bioinformatician we could think of models of possibly tracking genomes of viruses being shown at hospitals and such, we can track

whether this evolution is showing in order to indeed interact with human hosts at a higher rate. Evidence that we would find about DNA adapting would be an outcome and sign of natural selection. When species adapt over time to a certain situation, this is the showing of a species adapting to certain genetic variances in order to survive. Those who are able to adapt in this evolutionary rate are those who survive. All this begins with mutation of the DNA and from there, the tracking of the change of protein and thus the adaption of the species. I am excited to see these next steps taken in the bioinformatics world in order to track the fascinating idea of natural selection.