Reflection

By Wei Jun Tan

In Spring 2022, I joined the Statistics Directed Reading Program under the track "Introduction to Computational Biology". I and Iris Zhou are matched up to work with two Statistics Ph.D. students Anna Neufeld and Alan Min.

For the first week, we read the textbook Computational Genome Analysis to grasp a basic foundation in biology, such as cells, mutations, genes, nucleotides, codons, etc. Then, we have some revisions on probability, independence, and expected values.

Next, we dived into learning Markov chain and Hidden Markov Model to model CpG islands in human genome. We learned about the classic example of occasionally dishonest casino too. Moreover, we presented the concept of Viterbi algorithm, which aims to find the most probable state path, to Anna and Alan and implemented and visualized the algorithm in Python as a programming exercise.

Afterwards, we learn about more advanced algorithms such as forward-backward algorithm, Viterbi Training, and Baum Welch algorithm to estimate parameter values for transition and emission matrix for hidden Markov model. We spent a few weeks on these and implemented the forward-backward algorithm and Baum Welch algorithm.

Then, we started brainstorming the final project. We decide to use hidden Markov model to model the gene sequences in human chromosome 20 and 21. I parsed the source files into training-ready csv. I estimate the model parameters using the annotated genome of chromosome 21 and test the model on human chromosome 20. I made some visualization and performance evaluation, such as accuracy, precision, recall, and baseline accuracy of the model.

Nonetheless, my model does not work as expected as we achieve higher accuracy in testing set than training set, despite that testing set has greater size. I believe that the missing data is the biggest problem as the genome turns out to have about 50% of the time are genes, which are not reasonable in real world. It is reasonable to think that the missing data in the source files are all non-gene region; as they are simply ignored by the current model, out model turns out to have a boost in gene frequency. I also proposed other potential improvement, such as adding more emissions and hidden states to the model to avoid underfitting. Eventually, I shared and presented my project on Tuesday 6/7/2022 with other mentors and mentees.

My experience in DRP is amazing. I learned a lot about computational biology in genome analysis. I am comfortable and happy in the process of learning and building project with Iris, Anna, and Alan. Thank you all very much! This is a wonderful spring learning experience, and I would definitely recommend this program to my friend!