Introduction to Computational Biology

Mentee: Iris Zhou
Mentors: Anna Neufeld & Alan Min

In spring quarter of 2022, I worked with my mentors Anna Neufeld and Alan Min, and another student, Wei Jun Tan, to explore computational biology.

We started by reading textbook chapters from *Computational Genome Analysis* for a background on what genes, nucleotides, and codons are, and how genetic information is stored and transmitted. We then reviewed basic concepts of probability, and learned about how probability distributions, conditional probability, and markov chains can help model and simulate strings of DNA.

After becoming more interested in Markov chains, we pivoted into exploring how hidden markov models can be used to model and predict hidden states in genetic sequences, using Chapter 3 of *Biological Sequence Analysis* as a reference. Each week, we dove into a different set of algorithms, including the forward/backward algorithms and the Viterbi algorithm, dynamic programming algorithms which help to predict the most probably underlying states, and the Baum-Welch algorithm, an EM algorithm that iteratively improves parameter estimates. After studying an algorithm for the week, we presented what we'd learned to Alan and Anna, going through the formulas involved and demonstrating the steps of the algorithm. We also drew from an assignment in UW's Computational Biology course to implement the algorithms and predict hidden states in the genome sequence of *Methanocaldococcus jannaschii*.

In the last few weeks of the quarter, we went back to the Viterbi algorithm and used it to build a model to predict locations of genes in Chromosome 21. We used counts of nucleotides and dinucleotides to create the initial parameter estimates, trained the model based on previously discovered genes in the chromosome, and then plotted the predicted genes against the known genes to visualize the model's accuracy. The model I created had several limitations, including its low overall efficiency as well as how it handled undefined sections of the genetic sequence. I hope dive back into this project and improve the model's performance in the future. At the end of the quarter, I presented my project to a few other DRP mentors and mentees.

Overall, I enjoyed diving deeper into topics that I had seen appear in other courses (programming, statistics, probability, markov chains) and combining those topics with areas that were new to me (computational biology and genome analysis). I loved being able to learn in a more unstructured environment, and really enjoyed working with Wei Jun and getting guidance from Anna and Alan to explore these ideas!

References

Deonier, R., Waterman, M., & Tavaré, S. (2005). *Computational Genome Analysis*. Springer New York, NY. doi:10.1007/0-387-28807-4

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511790492