Summer Undergraduate Research Program –

Clustering DNA Strands for DNA-based Data Storage

PhD Student: Puru Sharma

PhD Supervisor: Weng-Fai Wong

Contact email: puru@u.nus.edu

Introduction:

In the ever-evolving landscape of information storage, DNA-based data storage has emerged as a groundbreaking solution that holds immense promise for the future. Unlike traditional methods reliant on silicon-based technologies, DNA data storage leverages the remarkable information-carrying capacity inherent in the genetic code, allowing for unprecedented data density and durability. As the volume of digital data continues to grow exponentially, conventional storage systems face significant challenges in terms of scalability and long-term stability. In response to this burgeoning need, DNA-based data storage has emerged as a transformative technology, offering the potential to store vast amounts of information in a compact, environmentally sustainable, and remarkably stable form.

To quickly describe the DNA-storage pipeline, the first step in DNA data storage is to encode digital information into DNA sequences. This typically involves converting binary data (0s and 1s) into a DNA sequence made up of {A, C, G, T} nucleotide bases. Once the information is encoded, the next step is to synthesize the actual DNA strands in the laboratory. The synthesized DNA strands are stored together in a test tube. To retrieve the stored data, the synthesized DNA needs to be read or *sequenced*. DNA sequencing determines the order of nucleotide bases in a DNA strand.

In the next step, the sequenced DNA data is grouped or **clustered** based on certain characteristics. This is often necessary because errors can occur during the synthesis or sequencing processes. By identifying and clustering similar DNA sequences, it becomes possible to mitigate errors and enhance the accuracy of data retrieval. Clustering algorithms play a crucial role in this step, grouping sequences that likely represent the same original information.

After clustering, the trace reconstruction step involves recovering the original digital information from the clustered DNA sequences. This process requires identifying and correcting errors, which may have occurred during synthesis or sequencing. Sophisticated algorithms and error-correction techniques are employed to ensure the accuracy of the reconstructed data.

In this project, we will be focusing on understanding and improving the **clustering** step of the pipeline, which consists of using machine learning algorithms to cluster similar DNA sequences together. DNA-storage research is very interdisciplinary, so expect to read papers ranging from journals such as Nature to AI conferences such as NeurIPS. While we are looking at this problem from a DNA storage perspective, similar problems exist in many other pipelines such as protein folding.

Objective:

In this project, we will first survey the existing literature to fully understand the state-of-the-art (SOTA) algorithms being used currently. Then, we will attempt to design our own DNA clustering algorithm. Our goal will be to significantly improve the accuracy, speed, or both, of the clustering algorithms we encounter in our survey.

Prerequisites:

This is a CS project, so you do not need any knowledge of Biology or Chemistry for it. However, interest and enthusiasm to explore something new is a requirement. For programming, there is no fixed language requirement. Any one of Python, C++ or Java is sufficient.

Since the problem we are tackling is essentially a Machine Learning problem, there is a lot of free material available to catch up on the required basics.

Expected Outcome:

Students will be introduced to a very exciting interdisciplinary field of study. Depending on the number of students participating in this project, we will try to produce between one and three of the following artifacts:

1) A survey of the literature covering the DNA clustering problem.
2) An open repository of our own improved algorithm built upon our own reproduction of the current SOTA.
3) A paper/poster covering our own algorithm. The goal would be to eventually submit it to a reputed conference or journal.

Depending on the quality of the work produced, we can utilize the DNA Storage wetlab within the School of Computing where we can perform practical in-situ experiments to evaluate our algorithm.

If you have any questions about the project before signing up, you can email me (Puru, puru@u.nus.edu) during this semester, and I will get back to you in a timely manner.