Eulerian walk and DNA sequence assembly

Graph Theory Project Proposal

Yifan Zhu

1 Introduction

Children like jigsaw puzzles, and the way to assemble the puzzle is by putting together pieces that match, one by one, till the puzzle is complete. DNA sequence assembly could be thought of as something similar: when a strand of DNA is passed into a particular "sequencing machine", it gives a large number of short reads of the DNA sequence. This type of technology is called shotgun sequencing. These reads form the jigsaw pieces in the puzzle and one must put together these pieces in an intelligent way to obtain the original sequence. There is, however, a catch here; apriori, we do not know what the original sequence or the"jigsaw big picture" looks like. Yet, the philosophy remains the same: to connect pieces or reads which are similar and hope that the "big picture" is reconstructed.

2 Methodology

- Consider a DNA sequence, say for example $AGTCGTCA$. This sequence is passed into a shotgun sequencer and reads are obtained. We must recover the original sequence from the reads.
- The read length depends on the sequencer. Here we assume that the reads are all the k-length substrings of the original sequence. The set of reads is called the k-mers of the sequence. In our example the set of $3-$mers is $\{AGT, GTC, TCG, CGT, GTC, TCA\}$. Of course, the reads are not obtained in the same order, they are all jumbled and that makes the recovery process difficult.
- The first step is to form the De-Bruin graph as follows:
  - The set of vertices are all the $k-1$-length sequences of the alphabet $\{A, C, G, T\}$. For a read of the sequence $r_1r_2...r_k$, draw a directed edge from $r_1r_2...r_{k-1}$ to $r_2r_3...r_k$ in the De-Bruin graph. For example, for the read AGT, we draw a directed edge from AG to GT as shown in the figure. We do this for all reads. This means that the number of edges in the De-Bruin graph is equal to the number of reads obtained. For our example sequence with the set of 3-mers are the reads, the graph looks as shown in the following figure.

- Clearly, with this representation, any feasible reconstruction of the original sequence must correspond to an Eulerian walk in the De-Bruin graph. Hence finding an Eulerian walk in the graph and finding the corresponding sequence is one way to solve the problem. Referring to the figure above, there exists exactly one Eulerian walk which is $AG \to GT \to TC \to CG \to GT \to TC \to CA$ which corresponds to the sequence $AGTCGTCA$, the original sequence that we started with.

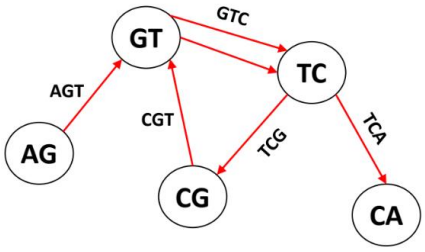

Figure 1: De-Bruin graph for our example. Note that not all vertices corresponding to sequences of length 2 are needed for this representation.

- Note that given a set of k-mers, it is entirely possible that there are multiple Eulerian walks, in which case, there are multiple feasible reconstructions given the reads. We will explore this as well in the homework.

3. Implementation

I would try to find and use some small gene data from "EcoGene" to do my simulation by using R or Matlab...(it depends)

- write a function that constructs the De-Bruin graph from a given list of sequence
- find a Eulerian walk from the De-Bruin graph, the Eulerian walk is a walk along the edges of a graph such that it passes through every edge once
- create the sequence corresponding the Eulerian walk we found before
- compare the reconstructed sequence with the original sequence

After finishing all of this I plan to try some huge real dataset

Reference

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001, August 14). An Eulerian path approach to DNA fragment assembly. https://www.pnas.org/content/98/17/9748