23.1 Genome Assembly Algorithms

1. Overview of Genome Sequencing and Assembly.
2. The Idury-Waterman Algorithm
3. The Human Genome

23.1.1 Background on Graph Theory

Recall the Eulerian cycle. For a graph $G = (V, E)$, an Eulerian cycle is a path that covers each edge exactly once. An Eulerian cycle exists if and only if $G$ is connected and every vertex has even degree.

A de Bruijn graph of order $k$ over some alphabet $\sigma$ has a vertices labeled with values in $\sigma^{k-1}$ and edges with values in $\sigma^k$. Generally, these edges are determined by $k$-merizing a set of strings and adding an edge for each $k$-mer. An example graph is constructed in Figure 23.1.

![Figure 23.1: De Bruijn Graph Construction with $k = 4$.]
23.1.2 Sequencing

Given a long DNA sequence, technological limitations of most DNA sequencing technologies requires us to break the sequence into short pieces and sequence each individual. The Human Genome was sequenced with Sanger sequencing, which allowed for relatively long reads at high cost, and today next generation sequencing technologies, such as Illumina sequencing are used, which allows a very large number of sequences of length around 100bp to be cheaply sequenced. Newer third-generation sequencing technologies are currently under development which aim to produce much longer contiguous reads at low cost.

Sequences are inexact, and mistakes are made. Different technologies have different biases, but no technology is exact.

23.1.3 Sequence Technology and Assembly

Sequencing technology confers limitations on assembly: clearly \( k \) can not be chosen to be longer than the sequence length, as then no \( k \)-mers would exist. However, \( k \) can not be too small, as if sequences of length \(< k\) are repeated in the genome, there wil be ambiguities in the de Bruijn graph. Specifically, if a repeat of length at least \( k \) exists, these ambiguities arise.

23.1.4 The Idury-Waterman Algorithm for Genome Assembly (A First Glance)

<table>
<thead>
<tr>
<th><strong>INPUT:</strong> ( N ): number of reads, ( k ): ( k )-mer size for de Bruijn Graph.</th>
<th><strong>OUTPUT:</strong> Assembly of the given reads.</th>
</tr>
</thead>
<tbody>
<tr>
<td>• All sequences and their reverse complements are ( k )-merized, producing a multiset of ( k )-mers.</td>
<td></td>
</tr>
<tr>
<td>• Construct the deBruijn graph with ((k - 1))-mers as nodes, and the ( k )-mers produced in step 1 as edges.</td>
<td></td>
</tr>
<tr>
<td>• Perform a variant of the Eulerian tour (cycle) algorithm and infer the sequence.</td>
<td></td>
</tr>
<tr>
<td>• Align the fragments to the inferred sequence from step 3.</td>
<td></td>
</tr>
</tbody>
</table>

Because biological data is rife with sequencing errors, and also due to the sequencing statistics, a straight Eulerian tour won’t necessarily exist, nor would it be a good idea to look for one.

23.1.5 Poisson Statistics

The Poisson distribution can be used to model the number of fragments that align with a given base. The Poisson distribution is defined as follows:

\[
P(\text{Poisson}(c) = k) = \frac{\lambda c^e}{k!}
\]

Here \( c \) represents the coverage of the readset, which is the average number of reads that overlap each base. In general when discussing the Poisson distribution, this parameter is referred to \( \lambda \), and it is also the expectation of the Poisson distribution.

These ideas are studied further in Lander-Waterman statistics, which analyze the number of sequences we need to draw in order to sufficiently cover the genome.