Bioinformatics I
WS 15/16
Tutor: Alexander Seitz

1	2	$\sum$ (10)

Jonas Ditz
& Benjamin Schroeder

Blatt 10

(Abgabe am 18. January 2016)

Theoretical Assignment - *Assembly using de Bruijn graph*

The de Bruijn graph for the reads provided on the assignment sheet would look like the graph in Figure 1, if one chooses $k = 4$. As one can see easily, there are not just one but four different Eulerian paths in the graph. Each path results in one of the following superstring:

$$S_1 = ACCGTTAACGTAAACGT$$
$$S_2 = ACCGTAAACGTTAACGT$$
$$S_3 = ACCGTTAAACGTAACGT$$
$$S_4 = ACCGTAACGTTAAACGT$$

This happens due to the fact that there are nodes with more than one outgoing an incoming edge. If one chooses $k = 5$ the resulting de Bruijn graph would look like the graph in Figure 2. It is obvious that the number of nodes is depended on k. With a bigger k there are more nodes in the resulting de Bruijn graph. The resulting graph for $k = 5$ is non-connected. So the superstring for this graph is not just one but three strings:

$$S_{part_1} = TAAACTG$$
$$S_{part_2} = CGTAACGTTAA$$
$$S_{part_3} = ACCGT$$

One can see that $S_{part_1} = f_5$ and $S_{part_3} = f_1$, while S_{part_2} is the result of an overlap alignment between f_2, f_3 and f_4. In general the second graph is more realistic. It is very unlikely to get a connected graph with real-life data. So it is possible that just f_2, f_3 and f_4 come from the same area in the target DNA. f_1 and f_5 could come from a different contig and, hence, result in this graph (Figure 2). Since this is a minimal example the first graph (Figure 1) is not wrong but one should not expect to get a connected graph all the time. In fact a connected graph is suspicious, normally.

Practical Assignment - *Assemble the human mitochondrial genome*

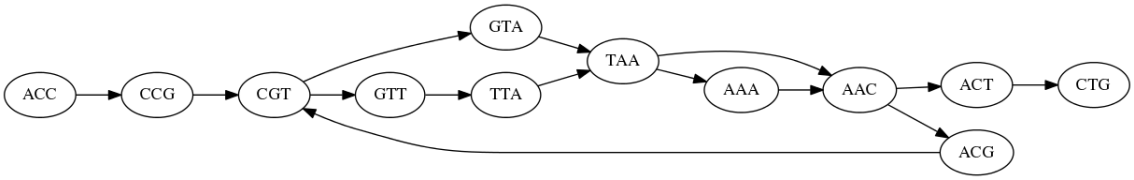

Figure 1: De Bruijn graph for the given reads and $k = 4$.

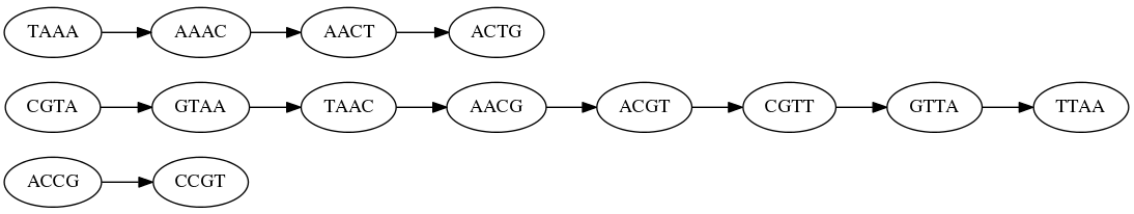

Figure 2: De Bruijn graph for the given reads and $k = 5$.