Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Dr. Marília Dias Vieira Braga

Algorithms in Genome Research

Winter 2023/2024

Exercises

Number 3, Discussion: 2023 November 24

1. The *Shortest Common Superstring* problem (SCP) is defined as follows. Given a set of fragments $\mathcal{F}$, find a string S of minimal length such that f or its reverse complementy $\overline{f}$ is a substring of S for all $f \in \mathcal{F}$. Use the overlap graph to solve SCP for the fragments below.

   Notes:

   - For taking reverse complements into account, each fragment can be represented by two vertices corresponding respectively to its head and to its tail.
   - For this dataset it suffices to consider only the edges corresponding to overlaps of length at least 3.

   $$\begin{array}{ll} f_1 = \texttt{ATAT} & f_4 = \texttt{TATA} \\ f_2 = \texttt{TATT} & f_5 = \texttt{TTAT} \\ f_3 = \texttt{TTAT} & f_6 = \texttt{AATA} \end{array}$$

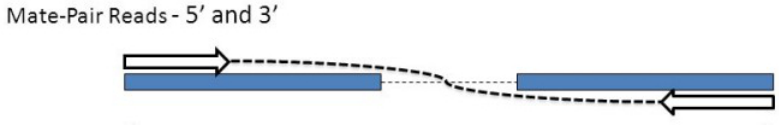
2. Consider the following set of reads, assuming that you know already that all of them originate from the same DNA strand.

   $$\begin{array}{ll} r_1 = \texttt{ATCCA} & r_6 = \texttt{GCAAG} \\ r_2 = \texttt{AGAGC} & r_7 = \texttt{AGATC} \\ r_3 = \texttt{AAGAT} & r_8 = \texttt{TAGAG} \\ r_4 = \texttt{GAGCA} & r_9 = \texttt{AGAGC} \\ r_5 = \texttt{CCATA} & r_{10} = \texttt{GAGCA} \end{array}$$

   (a) Build the corresponding overlap graph with a minimum overlap of 2.

   (b) Find a shortest common superstring for all reads. Is the coverage uniform? If not, find a layout with a more uniform coverage.

3. While in single-end sequencing, the sequencer reads a fragment from only one end to the other, in paired-end sequencing it reads the fragments from both ends. This gives a mate pair and a good estimation of the distance between them.



   How can mate pairs help in the assembling process?