Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Michel T. Henrichs

Algorithms in Genome Research
Winter 2020/2021

Exercises

Number 2, Discussion: 2020 November 20

1. For “reads” 1-8 from the following list, build the overlap graphs (a) with a minimum overlap of 2, and (b) with a minimum overlap of 3.

1	ATCCA
2	AGAGC
3	AAGAT
4	GAGCA
5	CCATA
6	GCAAG
7	AGATC
8	TAGAG
9	AGAGC
10	GAGCA

2. Find a shortest common superstring (error-free) for “reads” 1-10 of Exercise 1 above. Is the coverage uniform? If not, find a layout with a more uniform coverage.
3. Discuss the main experimental problems that make sequence assembly difficult in practice.
4. What are mate pairs and paired-end reads? What can be done with long reads that can not be done with paired-end reads?
5. What are the main differences between “traditional” (*de-novo*) genome assembly and comparative assembly?
6. What are the major steps in the comparative assembly strategy?
7. Let the following DNA sequence be a “reference genome”:

   AATGAGGTCATCCTTGCTGGACTCTAGCAC

   The following three sets of “reads” (a), (b), (c) originate from three “target” genomes that are closely related to the reference.

   Consider the following conditions:

   - There are no sequencing errors.
   - Each target genome differs from the reference by a single structural variation (rearrangement).
   - A read may come from any of the two complementary DNA strands.

Reconstruct the three target genomes by mapping the reads to the reference and identify the rearrangements.

(a)

1	AATGAGGTCA
2	AGGTCATCGAC
3	AGTCGATGAC
4	CATCGACTCT
5	CTAGAGTCGAT
6	GTGCTAGAGT

(b)

1	ACTCTAGCAC
2	AGTCCTGTACAG
3	CCTTGCTGTA
4	GCTGTACAGGAC
5	GGTCATCCTT
6	TGACCTCATT

(c)

1	AATGACAAGG
2	ACCCTGGACTCT
3	GGATGACCCTG
4	GTCATCCTTG
5	GTGCTAGAGT
6	TCCAGGGTCA