

**Algorithms in Genome Research
Winter 2020/2021**

Exercises

Number 3, Discussion: 2020 November 27

1. Discuss the reasons why the traditional assemblers fail to assemble short-read data.
2. The basic data structure used for short-read sequence assembly is the de-Bruijn graph. While it is conceptually easy, there are several challenges when you want to implement it in practice – name a few.
3. Draw the 4-dimensional de-Bruijn graph (i.e. where vertices correspond to 4-grams) for the following set of “reads”. Can you assemble the data set into a single contig? (There may be some “sequencing errors” that need to be corrected.)
AAATG, AATGA, AATGAC, AATGC, ACCAG, ACCAGA, ACCTG, ACGTT, AGACG, AGACGG, ATAAT, ATAATG, ATAATGC, ATGAC, ATGCA, ATGCAC, CACGG, CAGAC, CCAGA, CGTTA, CTGACGT, GACCA, GACCAGA, GACGTT, GCACG, GCACGG, GTTAAT, GTTAATG, TAATG, TAATGA, TACTA, TGACC, TGCAC, TTAAT.
4. In an assembly strategy that is based on long reads only, overlapping long reads have to be detected. What are the challenges, and how could a dynamic programming algorithm be designed that solves this problem?