

**Algorithms in Genome Research**  
**Winter 2017/2018**

**Exercises**

**Number 2, Discussion: 2017 November 03**

1. Find the shortest common superstring of the following sequences:

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

Is the coverage uniform? If not, find a layout with a more uniform coverage.

2. Discuss the main experimental problems that make sequence assembly difficult in practice.
3. Discuss the reasons why the traditional assemblers fail to assemble short-read data.
4. 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.)

GTTAAT, AGACG, ACGTT, CACGG, ACTAGG, CTGACG, TACTAG, GACCAGA, TAATG, AATGC, TGCAC, GCACG, ATGCA, GTTAATG, AAATG, TGCAC, GCACG, CACGG, TAATGA, AATGAC, CAGAC, AGACG, ACCAGA, ATAATG, TAATG, AATGA, GCACGG, ACTAG, TTAATG, TAATG, TGACC, ATAAT, CCAGA, ATGCA, ATAAT, ACCTGA, ATGCAC, TGCAC, CGTTA, CGTTA, TTAATG, GACCA, ACCAG, CCAGA, CAGAC, ATGAC, GACGTT, ATGGA, ACGTT.

5. What are the major steps in the comparative assembly strategy?