

**Algorithms in Genome Research**  
**Winter 2015/2015**

**Exercises**

**Number 2, Discussion: 2015 November 20**

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.
  - (a) While it is conceptually easy, there are several challenges when you want to implement it in practice – name a few.
  - (b) Give an efficient implementation of a de-Bruijn graph as a Java class.
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?  
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.
4. What are the major steps in the comparative assembly strategy?
5. Devise a strategy to use comparative genome data to order contigs resulting from a *de-novo* sequencing assembly procedure.