Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye

Algorithms in Genome Research
Winter 2009/2010

Exercises

Number 7, Discussion: 2009 December 11

1. De-novo assembly of short-read data:

   (a) Discuss the reasons why the traditional assemblers fail to assemble short-read data.

   (b) The basic data structure used for short-read sequence assembly is the de-Bruijn graph. What are the great challenges when applying this data structure in practice?

   (c) Draw the 4-dimensional de-Bruijn graph for the following set of reads. Can you assemble the data set into a single contig?

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

2. Comparative genome assembly:

   (a) What are the main differences to “traditional” genome assembly?

   (b) What are the major steps in the assembly strategy?

   (c) Develop the details of a simple read mapping algorithm that uses both the q-gram lemma and the pigeonhole principle.

   (d) What is “layout refinement” and how is it performed?