Universität Bielefeld
Technische Fakultät

Dr. Alexander Sczyrba
Linda Sundermann

Algorithms in Genome Research
Winter 2012/2013

Exercises

Number 3, Discussion: 09. November 2012

1. Construct the overlap graph for the following set of reads, assuming no sequencing errors, i.e. only exact prefix-suffix matches are allowed, and considering only overlaps of size three or more. (Note that the orientation of the reads is unknown.)

   ```
   1 TCCCA
   2 GGTAAT
   3 TCTTAGT
   4 ACCGAG
   5 CCAGT
   6 GGATTG
   7 AATCT
   ```

   (a) Compute a layout. How many contigs do you get?

   (b) Assume that the first two reads `TCCCA` and `GGTAAT` from above form a mate pair in opposite relative direction, originating from a "clone" with approximate length 25bp. What do you learn about the relative location of the contigs?

2. Discuss the reasons why the traditional assemblers fail to assemble short-read data.

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?

   `GAACCT,TGGGT, AACCG,CTGGG,CAACC, TAACTG,ACTGGG,TCGAACC,AACCGG, CTAACT,GAACCT, GGCTCAA, GTCAAC, GGGTCA, CAACCG, CCTAACT, ATCGAA, GGGTC, AACTGG, ATCGA,ACCTA, ACCTAA,TGGGTC, AACCTA, CGAAT CTGGCT,CGAACC,CACCCG,CCTAA, TCAACCGG,GGTCAAC, CTGGGTC, GGTCAA,ACTGG`