Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye

Algorithms in Genome Research

Winter 2006/2007

Exercises

Number 2, Discussion: 2006 November 10

1. Discuss the main experimental problems that make sequence assembly difficult.
2. Find the shortest common superstring of the following sequences:

   ```
   ATAGCC
   ATATAT
   ATATCG
   CGGGAC
   GACATA
   GACTAT
   GCCGGT
   GGTATA
   TATATA
   TATCGG
   ```

   Is the coverage uniform? If not, find a layout with a more uniform coverage.
3. In the overlap phase, prefix-suffix "local alignments" are seeked.
   (a) Work out the details of a dynamic programming algorithm.
   (b) What are the time and space complexities of the seed-based algorithm mentioned in class?
4. What are mate pairs? Do they simplify the assembly problem?
5. Construct the overlap graph for the following set of reads, assuming no sequencing errors, i.e. only exact prefix-suffix matches are allowed. (Note that the orientation of the reads is unknown.)

   ```
   TCCCA
   GGTAAT
   CTTAGT
   CCGAG
   CCAGT
   GATTG
   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?