Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Dr. Pedro Feijão

Algorithms in Genome Research

Winter 2015/2015

Exercises

Number 1, Discussion: 2015 November 13

1. Discuss the main experimental problems that make sequence assembly difficult.
2. 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.

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 and paired-end reads? How do they help simplifying 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, 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?