Exercise sheet

High Throughput Sequencing Methods
Bioinformatics Leipzig

WS 13/14

1   Bowtie

1. What does Bowtie do? (input/output)

2. Which methods are used and why?

1.1   Suffix array

Construct the suffix array for the string **Bioinformatik$**.
Construct the BWT table.

1.2   Burrows Wheeler Transformation

Given the following column of the Burrows-Wheeler table as in the lecture, find the original string:
$LTRAHGOUIMS

1.3   FM-index

Construct the suffix array, the BWT table and an array **C** and a matrix **O** for the string
GCGAATATCTGAAATGCTTA.
Find the positions of the following substrings in the suffix array. Do the substrings exist in the suffix array? If yes, how often?

1. **AAT**

2. **CAA**

3. **TAT**

2   Segemehl

1. What does Segemehl do? (input/output)

2. Which methods are used and why?

2.1   Suffix tree

Construct a suffix tree for the string **banana$**.
Construct a suffix tree for the string **Mississippi$**.

2.2   Longest common prefix

Construct the longest common prefix arrays for prefixes of the lengths 1,2,3 and (if possible) 4 for the strings **Mississippi$** and **banana$**.

3   Sequence Assembly

3.1   De Bruijn Graphs

1. Create the 4-mers for the following fragments:

$$F = \{ACGAACG, TGCTGAC, ACTGCT, AACGG, CTGACGA\}$$

2. Create the De-Bruijn-Graph for the set of 4-mers.

3. Find an Euler path through the graph and the resulting sequence.