Exercise sheet

Sequence Alignments and Phylogeny
Bioinformatics Leipzig

WS 13/14

1 Biological Background

1.1 Which of the following are pyrimidines?

1. Adenine and Thymine
2. Cytosine and Thymine
3. Cysteine and Guanine
4. Cytosine and Guanine

1.2 Which of the following contain phosphorus atoms?

1. Proteins, DNA and RNA
2. DNA and RNA
3. DNA only
4. Proteins only

1.3 Which of the following contain sulfur atoms?

1. Histidine
2. Lysin
3. Methionine
4. all of the above

1.4 Which of the following is not a valid amino acid sequence?

1. EINSTEIN
2. CRICK
3. FARADAY
4. WATSON

1.5 Which of the following 'one-letter' amino acid sequence corresponds to the sequence Tyr-Phe-Lys-Thr-Glu-Gly?

1. YFKTEG
2. WPKTEG
3. WFLTGY
4. YFLTDG

1.6 Consider the following DNA oligomers. Which to are complementary to one another? All are written in the 5' to 3' direction (i.TTAGGC ii.CGGATT iii.AATCCG iv.CCGAAT)

1. i and ii
2. ii and iii
3. i and iii
4. ii and iv

2 Pairwise Alignments

2.1 Needleman-Wunsch Algorithm

Given the alphabet $B = \{A, C, G, T\}$, the sequences $s = ACGCA$ and $p = ACCG$ and the following scoring matrix D:

	A	C	T	G	-
A	3	-1	-1	-1	-1
C	-1	3	-1	-1	-1
T	-1	-1	3	-1	-1
G	-1	-1	-1	3	-1
-	-1	-1	-1	-1	0

1. What kind of scoring function is given by the matrix D, similarity or distance score?
2. Use the Needleman-Wunsch algorithm to compute the pairwise alignment of s and p.
3. Do the backtracking step and write down the resulting optimal global alignment of the two sequences.

3 MSA and Profile Alignments

3.1 Given the following multiple sequence alignment (without gaps!)

A	A	C	T	C	G	A	C	T	T	C	G
A	T	G	A	T	G	A	C	T	T	G	G
A	T	G	A	T	C	A	G	T	G	C	C
G	T	C	A	T	C	A	T	G	G	G	C
G	A	C	A	T	C	A	C	C	T	C	C

1. Construct the profile P for the MSA.
2. Construct the consensus sequence K.
3. When aligning as many sequences, why is working with profiles more suitable than using the consensus sequence?
4. Given the sequence S = GAGACGACTGGG. Calculate the score for the alignment of K and S (using the edit distance) and the score for the alignment of S and P (using the score below and the edit distance).

   Edit distance (for columns i,j):

$$d(x_i, y_j) = \begin{cases} 1 & \text{if } x_i \neq y_j \\ 0 & \text{if } x_i = y_j \end{cases}$$

Profile scoring (for columns i,j):

$$B = \{A, T, C, G\}$$
$$sc(s_i, p_j) = \sum_{b \in B} d(b, s_i) * p_{b,j}$$

Hints:
To get the overall score for the whole sequence compute the sum over the scores for each column i in the alignment.
Since there are no gaps in the alignment, the column indices i and j are the same (i=j).

4 Phylogeny

1. What does UPGMA do? (input/output)
2. What is the difference between UPGMA and WPGMA? (formula)

4.1 Construct a) the UPGMA tree and b) the WPGMA tree for the following distance matrix. Where are the differences? Write the resulting distances at the inner nodes of the tree.

	A	B	C	D
A	0	9	7	5
B	9	0	8	10
C	7	8	0	8
D	5	10	8	0

4.2 Compute the WPGMA tree for the following distance matrix. Write the resulting distances at the edges of the tree.

	A	B	C	D	E	F
A	0					
B	2	0				
C	4	4	0			
D	6	6	6	0		
E	6	6	6	4	0	
F	8	8	8	8	8	0

5 Bowtie

1. What does Bowtie do? (input/output)
2. Which methods are used and why?

5.1 Suffix array

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

5.2 Burrows Wheeler Transformation

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

5.3 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**

6 Segemehl

1. What does Segemehl do? (input/output)
2. Which methods are used and why?

6.1 Suffix tree

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

6.2 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$** and **banana$**.

7 Sequence Assembly

7.1 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.