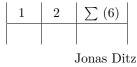
### Bioinformatics I

WS 15/16

Tutor: Alexander Seitz



& Benjamin Schroeder

## Assignment 6

(Handed in 23. November 2015)

# Theoretical Assignment - Sequence-profile alignment and expected patterns in sequences

(a)

The profile as a PSWM for the given MSA would look like table 1.

Table 1: PSWM of the given MSA

			$p_3$	$p_4$	$p_5$
$\mathbf{A}$	$0.\overline{3}$	0	$0.\overline{3}$	0	0
$\mathbf{C}$	0			1	0
$\mathbf{G}$	0	0	0	0	1
${f T}$	0	1	$0.\overline{6}$	0	0
-	$0.\overline{6}$	0	0	0	0

Using this PSWM we can now compute an optimal semiglobal alignment of our profile with the sequence A = CATTCCGTTC. First we calculate the scoring matrix using as a scoring function s(a,b) = -1, s(a,a) = 3 and d = 2:

	$b_1$	$b_2$	$b_3$	$b_4$	$b_5$	$b_6$	$b_7$	$b_8$	$b_9$	$b_{10}$
	$egin{array}{c} b_1 \ \mathrm{C} \end{array}$	A	${ m T}$	${\bf T}$	$\mathbf{C}$	$\mathbf{C}$	G	${\bf T}$	${ m T}$	$\mathbf{C}$
$p_1$	$ \begin{array}{c c} -1.\overline{6} \\ -1 \\ -1 \\ 3 \end{array} $	$-0.\overline{3}$	$-1.\overline{6}$							
$p_2$	-1	-1	3	3	-1	-1	-1	3	3	-1
$p_3$	-1	$0.\overline{3}$	$1.\overline{6}$	$1.\overline{6}$	-1	-1	-1	$1.\overline{6}$	$1.\overline{6}$	-1
$p_4$	3	-1	-1	-1	3	3	-1	-1	-1	3
$p_5$	-1	-1	-1	-1	-1	-1	3	-1	-1	-1

Now we fill the DP matrix using that scoring matrix:

	0	$\mathbf{C}$	A	${ m T}$	${\bf T}$	$\mathbf{C}$	$\mathbf{C}$	G	${ m T}$	${ m T}$	$\mathbf{C}$
0	0	0	0	0	0	0	0	0	0	0	0
$p_1$	-2	$-1.\overline{6}$	$-0.\overline{3}$	$-1.\overline{6}$							
$p_2$	-4	-3	$-2.\overline{6}$	$2.\overline{6}$	$1.\overline{3}$	$-0.\overline{6}$	$-2.\overline{6}$	$-2.\overline{6}$	$1.\overline{3}$	$1.\overline{3}$	$-0.\overline{6}$
$p_3$	-6	-5	$-2.\overline{6}$	$0.\overline{6}$	$4.\overline{3}$	$2.\overline{3}$	$0.\overline{3}$	$-1.\overline{6}$	$-0.\overline{6}$	3	1
$p_4$	-8	-3	$-4.\overline{6}$	$-1.\overline{3}$	$2.\overline{3}$	$7.\overline{3}$	$5.\overline{3}$	$3.\overline{3}$	$1.\overline{3}$	1	6
$p_5$	-10	-9	-4	$-3.\overline{3}$	$0.\overline{3}$	$5.\overline{3}$	$6.\overline{3}$	$8.\overline{3}$	$6.\overline{3}$	$4.\overline{3}$	4

One can see that the optimal alignment (colored in red) is:

(b)

### i. Compute the probability that S[1...4] contains P = GT without substitutions.

There are three different outcomes for this result.

$$p_1:S[1]=G \text{ and } S[2]=T$$
 
$$p_2:S[2]=G \text{ and } S[3]=T$$
 
$$p_3:S[3]=G \text{ and } S[4]=T$$

Since all S[i]s are independent of each other, we simple multiply all probabilities for each S[i].

$$P(p_1) = \frac{1}{4} * \frac{1}{4} * 1 * 1 = \frac{1}{16}$$

$$P(p_2) = 1 * \frac{1}{4} * \frac{1}{4} * 1 = \frac{1}{16}$$

$$P(p_3) = 1 * 1 * \frac{1}{4} * \frac{1}{4} = \frac{1}{16}$$

Where  $\frac{1}{4}$  is the probability to choose G or C, respectively, and 1 is the probability to choose any character from the alphabet. All three outcomes would fulfill the task, so we simply sum up all probabilities:

$$P(S[1...4] \text{ contains GT}) = P(p_1) + P(p_2) + P(p_3) = \frac{3}{16}$$

### ii. Compute the probability that S[1...6] contains P = AAA with at most one substitution.

There are three different possible outcomes for P to appear at position 1:

$$S[1...3] = YAA$$

$$S[1...3] = AYA$$

$$S[1...3] = AAY$$

Each of this possibilities has the probability  $\frac{1}{16}$ . Since we have four different start positions (1,2,3 and 4) and for each position three different outcomes, the final probability is:

$$P(S[1...6] \text{ contains AAA}) = 4 * 3 * \frac{1}{16} = \frac{3}{4}$$

Theoretical Assignment - *Practice writing an introduction / background for a paper*