1	2	$\sum$ (6)

Bioinformatics I
WS 15/16
Tutor: Alexander Seitz

Jonas Ditz
& 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_1	p_2	p_3	p_4	p_5
A	$0.\overline{3}$	0	$0.\overline{3}$	0	0
C	0	0	0	1	0
G	0	0	0	0	1
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 C	b_2 A	b_3 T	b_4 T	b_5 C	b_6 C	b_7 G	b_8 T	b_9 T	b_{10} C
p_1	$-1.\overline{6}$	$-0.\overline{3}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$
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	C	A	T	T	C	C	G	T	T	C
0	0	0	0	0	0	0	0	0	0	0	0
p_1	-2	$-1.\overline{6}$	$-0.\overline{3}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$	$-1.\overline{6}$
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:

C	A	T	T	C	C	G	T	T	C
	p_1	p_2	p_3	p_4	-	p_5			

(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:

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

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:

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

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