Bioinformatics I
WS 15/16
Tutor: Alexander Seitz

1	2	$\sum$ (6)

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	b_2	b_3	b_4	b_5	b_6	b_7	b_8	b_9	b_{10}
	C	A	T	T	C	C	G	T	T	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:

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

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*

Pattern recognition is a important topic for all life sciences areas and also a though challenge for the computational area. The main goal of finding repeating patterns in bio-sequences is the lead to propose functions which could be contained in new unknown sequences. Specially motifs which are involved in regulation can give a deep insight to regulatory networks and the involvement of proteins. The field of transcription factors is also highly interested in interaction motifs, which can lead to a regulation of gene expression.
For example the standard example the nucleus located protein p53, which is in very often mutated in cancer cells. A very conserved region of p53 binding side which binds to following motif on the DNA ([**?**]):

$$5' - R - R - R - C - (A/T) - (T/A) - G - Y - Y - Y - 3'$$

Directly by p53 influenced proteins are for example p21, which inhibits the cell cyklus and Bcl2, an trigger protein for apoptosis. [**?**]

sperimental approach for showing the binding of a protein is the chromatin immunoprecipitation (ChIP). With the fixation of the nucleic-acid-protein-complex via formaldehyde. Afterwards the cells are lysated via sonification. The resulting DNA and DNA-Protein-complex fragments are seperated via an immunoprecipitiation with a protein specific antibody. After several washing steps, the fixation via the formaldehyde is destroyed via heating up the sample. After another seperation step, the result is the DNA from the DNA-protein-complex. The gained can be sequenced or otherwise analysed. A critical parameter for this methode is the specifity of the antibody. Because if the specifity is to low, no pure sample can be purified [**?**].
But of course there are also computational methods, which allow the discovery of new motifs like with the Gibbs sampling mehtod implemented in or the Projection Method. These methods all challenge the problem, what is background and therefore not important for a interaction? But also there are Database like Jaspar or TRANSFAC which already contain motifs, which can be compared to a querry.
With this new algorithm we designed, we want to push the limits of space, time complexity and accuracy. The "best-motives"-Algorithm solves could therefore make motif finding much more eas-

ier and faster.

References

[1] C. Wagener and O. Müller, Molekulare Onkologie: Entstehung, Progression, klinische Aspekte ; 95 Tabellen. Thieme, 2010.

[2] S. Sohr and K. Engeland, "The tumor suppressor p53 induces expression of the pregnancy-supporting human chorionic gonadotropin (hcg) cgb7 gene," Cell Cycle, vol. 10, no. 21, pp. 3758–3767, 2011. PMID: 22032922.

[3] E. Lottspeich, Bioanalyitk. Springer, 2012.