HW 3:

Submission deadline: 17 April 2020

PS: HWs submitted after the deadline will not be accepted

Email your HW to:

nizamettinaydin@gmail.com

You must place **I2B-HW3** in the "Subject" field.

You will include two attachments in your email: 1. A pdf file with answers to (a), (b), and (c); 2. A perl program file (xxxxxxxx8-hw3.pl)

You must name your file by using as your matriculation number-hw3.pl (for example, 07011068-hw3.pl)

In Expectation-Maximization algorithm there are two steps: Expectation step and Maximization step.
In the expectation step, background residue frequencies are calculated based on those residues that are not in the initially aligned sites. Column specific residues are calculated for each position in the initial motif alignment. Using this information, the probability of finding the site at any position in the sequences can then be calculated.
Residues not in a motif are background frequencies used to determine probability of finding site at any position in a sequence to fit motif model.
An initial, random alignment for 6 sequences are given here. The motif we are searching for is six bases wide (motifs at each sequence are highlighted).

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

In this case;

a. Fill in the following observed count table by calculating background and observed residue count of each residue for initial alignment.

Nucleotide	Motif position (0 = Background)						
	0	1	2	3	4	5	6
A	20	1	3	1	0	3	2
C	25	1	1	1	0	3	3
G	17	4	1	1	3	0	0
T	22	0	1	3	3	0	1

b. Using the count information in (a) fill in the residue frequency/probability ($P_{ca} = n_{ca} / N_c$) table. (round the fraction to two digits)

Nucleotide	Motif position (0 = Background)						
	0	1	2	3	4	5	6
A	0.24	0.17	0.5	0.17	0	0.5	0.33
C	0.30	0.17	0.17	0.17	0	0.5	0.5
G	0.20	0.67	0.17	0.17	0.5	0	0
T	0.26	0	0.17	0.5	0.5	0	0.17

c. Fill in the following adjusted residue frequency table. In calculation of residue frequencies for each position (except 0), use pseudocounts as discussed in slide 32 ($P_{ca} = (n_{ca} + b_{ca}) / (N_c + B_c)$, assuming B_c=1. (round the fraction to two digits)

Nucleotide	Motif position (0 = Background)						
	0	1	2	3	4	5	6
A	0.24	0.18	0.46	0.18	0.04	0.46	0.32
C	0.30	0.18	0.18	0.18	0.04	0.46	0.46
G	0.20	0.61	0.18	0.18	0.46	0.04	0.04
T	0.26	0.04	0.18	0.46	0.46	0.04	0.18

d. Write a Perl program generating tables in (a), (b) and (c) when the input is initial random alignment given above.