

Lecture 3 Homework

<http://thegrantlab.org/>

Dr. Barry Grant

Your Name: Steven Gan

UCSD Email: digan@ucsd.edu

PID: A59020397

To complete this homework you must be working on your own copy in **Google Docs**. Once you have filled in your answers to **Q1-Q4** in the space provided click **File > Download > PDF document** and upload to gradescope (link can be found on the class website).

Q1. [4pts] Consider the following multiple alignment of Transcription Factor Binding site DNA sequences

	1	2	3	4	5
Sequence 1	-	G	A	G	C
Sequence 2	C	T	A	G	A
Sequence 3	C	G	A	-	A
Sequence 4	A	G	C	G	A

Give the average profile (frequency matrix) of the above alignment by filling out the table below. The first position of the first column (i.e. position in the alignment) has been done for you, now complete the rest. You will use this table for answering questions 2 and 3 below.

	1	2	3	4	5
A	0.25	0	0.75	0	0.75
C	0.5	0	0.25	0	0.25
T	0	0.25	0	0	0

G	0	0.75	0	0.75	0
-	0.25	0	0	0.25	0

Q2. [2pts] What is the highest scoring sequence match to your profile above (question 1) and what is its score?

Sequence: CGAGA

Score: 4.4

Q3. [2pts] Using your completed profile table above (from question 1) score the following two sequences (**S1** and **S2**):

S1. CTGGC -0.1

S2. AGCTA 0.3

Q4. [2pts] Following the heuristic threshold for a positive match proposed in Harbison et al. [Nature (2004) 431:99-104.] namely using the threshold for a positive match = 60% x Max Score
Are either of the two sequences in question 3 potential transcription factor binding sites? If so, why?

None them are potential TF-binding sites using the 60% threshold, since both scores are lower than the threshold.

	1	2	3	4	5
Sequence 1	-	G	A	G	C
Sequence 2	C	T	A	G	A
Sequence 3	C	G	A	-	A
Sequence 4	A	G	C	G	A

Frequency	1	2	3	4	5
A	0.25	0	0.75	0	0.75
C	0.5	0	0.25	0	0.25
T	0	0.25	0	0	0
G	0	0.75	0	0.75	0
-	0.25	0	0	0.25	0

Counts	1	2	3	4	5
A	1		3		3
C	2		1		1
T		1			
G		3		3	
-	1			1	

PSSM	1	2	3	4	5
A	0	-1.6094379	0.95551145	-1.6094379	0.95551145
C	0.58778666	-1.6094379	0	-1.6094379	0
T	-1.6094379	0	-1.6094379	-1.6094379	-1.6094379
G	-1.6094379	0.95551145	-1.6094379	0.95551145	-1.6094379
-	0	-1.6094379	-1.6094379	0	-1.6094379

Query Seq	Score
CTGGC	-0.0661398
AGCTA	0.30158498

Max Score	Threshold
4.40983245	2.64589947
CGAGA	