Hands-On Three
Constructing Position Weight Matrix

Problem

In this hands-on exercise, we are going to build two position weight matrices (PWMs), one for the 5' splice site (also known as donor splice site) and the second for the 3' splice site (also known as the acceptor splice site). We are going to build the PWM by considering the splice sites of the MOG gene.
Go to NCBI and retrieve the record (data entry) with accession number Z48051.

A) 5' Splice Site:

1) MOG has eight exons and seven introns:

```
CDS             join(1166..1253,3274..3621,10106..10219,11597..11617,
                11860..11880,14238..14354,14658..14678,15129..15142)
```

We are going to consider the last 3 bases of the exon followed by the first six bases of each intron for the seven introns. Short sequences of length 9 are known as 9-mers.

1) We are going to determine the seven 9-mers representing the seven 5' splice sites.
For the first intron, we have the following 9-mer: caggtaaga, where cag are the last three bases of exon one and gtaaga are the first 6 bases of intron one. Let us denote it by X_1 = caggtaaga.
Determine the remaining six 9-mers and list them here:

X_1 = caggtaaga
X_2 =
X_3 =
X_4 =
X_5=
X_6 =
X_7 =

2) Copy the seven 9-mers and paste them in the window at http://weblogo.berkeley.edu/logo.cgi to create a logo. Note that you will have to remove all the "X_j" before hitting the "Create Logo" button.
Compare the logo obtained with the one we studied in the course reproduced here:

5′ splice site

3) Fill in Table 1 that lists the seven 9-mers representing the 5'splice sites of MOG. The first row, corresponding to X_1, is already filled in.

Table 1: The 9-mers representing the seven 5' splice sites of MOG

	1	2	3	4	5	6	7	8	9
X_1	C	A	G	G	T	A	A	G	A
X_2									
X_3									
X_4									
X_5									
X_6									
X_7									

4) Use Table 1 to fill Table 2 which represents the probability distribution of each base in each of the 9 positions. Note that this is the Position Weight Matrix representing the 9-mers.

Table 2: PWM of the 9-mers of the 5'splice sites of MOG

	1	2	3	4	5	6	7	8	9
A									
C									
G									
T									

5) Use Table 2 and Laplace rule for pseudocounts to build Table 3.

Table 3: PWM with pseudocounts using Laplace's rule

	1	2	3	4	5	6	7	8	9
A									
C									
G									
T									

6) Use Table 3 and the fact that the genome-wide average G and C content is 44% to fill Table 4 which represents the log-odd scores of the 9-mers of the 5' splice sites of MOG. Use log base 2.

Table 4: Log-odds of the PWM of the 9-mers from Table 3 where base = 2

	1	2	3	4	5	6	7	8	9
A									
C									
G									
T									

7) To double-check the values of Table 4, run the Python program, **create_pwm.pl**, that:

- Takes as input the seven 9-mers of part 1) of the problem, saved in a file: Donor_MOG.txt
- Computes the probability of occurrence of each base at each position.
- Computes the 36 log-odds scores (log of observed/expected) where expected is 0.28 for A's and T's, and 0.22 for C's and G's, and with pseudocount of 1.
- Writes the 36 values of the entries of the PWM into a file, Donor_MOG_matrix.txt; Donor_MOG_matrix.txt has 4 rows with 9 values on each line, as in Table 4.

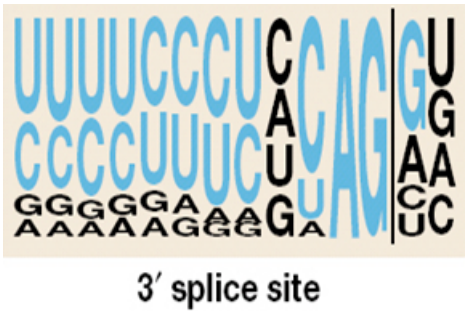
B) 3' Splice Site:

Repeat all the steps of Part A for the 3' splice sites where we consider 14-mers, the last thirteen bases of the intron followed by the first base of the exon.

8) Determine the seven 14-mers and list them here:

Y_1 = gtgtcttggacagg
Y_2 =
Y_3 =
Y_4 =
Y_5=
Y_6 =
Y_7 =

9) Copy the seven 14-mers and paste them in the window at http://weblogo.berkeley.edu/logo.cgi to create a logo. Note that you will have to remove all the "Y_j" before hitting the "Create Logo" button.
Compare the logo obtained with the one we studied in the course reproduced here:



10) Fill in Table 5 that lists the seven 14-mers representing the seven 3' splice sites of MOG.

Table 5: The 14-mers representing the seven 3' splice sites of MOG

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Y_1														
Y_2														
Y_3														
Y_4														
Y_5														
Y_6														
Y_7														

11) Use Table 5 to fill Table 6 which represents the probability distribution of each base in each of the 9 positions. Note that this is the Position Weight Matrix representing the 14-mers.

Table 6: PWM of the 14-mers of the 3'splice sites of MOG

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A														
C														
G														
T														

12) Use Table 6 and Laplace rule for pseudocounts to build Table 7

Table 7: PWM with pseudocounts using Laplace's rule

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A														
C														
G														
T														

13) Use Table 7 and the fact that the genome-wide average G and C content is 44% to fill Table 8 which represents the log-odd scores of the 14-mers of the 3' splice sites of MOG.
Use log base 2.

Table 8: Log-odds of the PWM of the 14-mers from Table 7 where base = 2

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A														
C														
G														
T