# Hands-On Four
# Scoring using a Position Weight Matrix

## Problem

In this hands-on exercise, we are going to use the position weight matrices (PWMs), we built in Hands-On 3, to score sequences and to determine the cutoff values (thresholds) for the PWMs for both, the 5' splice site (also known as donor splice site) PWM and for the 3' splice site (also known as the acceptor splice site) PWM.
Recall that the log-odds PWM were built by considering the splice sites of the MOG gene with accession number Z48051 (at NCBI).

### A) 5' Splice Site:

Recall that the MOG gene has eight exons and seven introns:

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

In Hands-On 3, we first found the seven 9-mers for the 5' splice sites by taking the last 3 bases of the exon followed by the first six bases of each intron. Here is what we found:

$X_1$ = CAGGTAAGA
$X_2$ = AAGGTGAGT
$X_3$ = GAGGTACAG
$X_4$ = TAGGTGAGT
$X_5$ = TTGGTAAGT
$X_6$ = CAGGTGCAG
$X_7$ = TACGTAAGT

We also used the fact that the genome-wide average G and C content is 44% to fill Table 1 which represents the log-odd scores of the 9-mers of the 5' splice sites of MOG. We used log base 2.

Table 1: Log-odds of the PWM of the 9-mers for the 5' splice sites where base = 2

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| A | -0.623 | 1.184 | -1.623 | -1.623 | -1.623 | 0.699 | 0.962 | -0.038 | -0.623 |
| C | 0.310 | -1.275 | -0.275 | -1.275 | -1.275 | -1.275 | 0.310 | -1.275 | -1.275 |
| G | -0.275 | -1.275 | 1.532 | 1.725 | -1.275 | 0.725 | -1.275 | 1.310 | 0.310 |
| T | 0.377 | -0.623 | -1.623 | -1.623 | 1.377 | -1.623 | -1.623 | -1.623 | 0.699 |

1) We will use Table 1 to score the seven 9-mers for the 5' splice sites. Complete Table 2 where the score for $X_1$ has already been computed and entered.
Recall: Since $X_1$ = CAGGTAAGA then
the score of $X_1$ = 0.310 + 1.184 + 1.532 + 1.725 + 1.377 + 0.699 + 0.962 + 1.310 + (-0.623)
= 8.476

Table 2: Scores of the seven 5' splice sites of the MOG gene with PWM of Table 1

| Sequence | Score |
|---|---|
| $X_1$ = CAGGTAAGA | 8.476 |
| $X_2$ = AAGGTGAGT | |
| $X_3$ = GAGGTACAG | |
| $X_4$ = TAGGTGAGT | |
| $X_5$ = TTGGTAAGT | |
| $X_6$ = CAGGTGCAG | |
| $X_7$ = TACGTAAGT | |

2) To double-check the values you got in Table 2, run the program, **score_oligo.pl**, that:

- Takes as input
  - Donor_MOG_matrix.txt to read the values of the PWM
  - Donor_MOG.txt to read the seven 9-mers
- Computes the score of each sequence from Donor_MOG.txt
- Writes each input sequence followed by its score into a file, Donnor_MOG_scores.txt.

3) You are going to randomly choose seven 9-mers from the MOG sequence (Z48051 at NCBI) that are not 5' splice sites (but have the invariant GT in positions 4 and 5) and score them.

Table 3: Scores of 7 randomly chosen 9-mers with GT (positions 4 and 5) with PWM of Table 1

| Random Sequence | Start | End | Region | Score |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

4) Check out the 14 scores you obtained in Tables 2 and 3 and decide on a good cutoff value that can be used as threshold. How many false positives and false negatives do you have?

## B) 3' Splice Site:

Repeat all the steps of Part A for the 3' splice sites.
In Hands-On 3, we found seven 9-mers for the 3' splice sites by taking the last thirteen bases of the intron followed by the first base of the exon. Here is what we found:

$Y_1$ = GTGTCTTGGACAGG
$Y_2$ = TCCTGCCTTTCAGA
$Y_3$ = TTTTCTATTTTAGG
$Y_4$ = GTTTCTCTTTCAGA
$Y_5$ = GTGAACAATTCAGA
$Y_6$ = TTTTTGTTTTCAGG
$Y_7$ = CTCCTTCTTCTAGG

We also used 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 3' splice sites of MOG. We used log base 2.

Table 4: Log-odds of the PWM of the 14-mers for the 3' splice sites where base = 2

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | -1.623 | -1.623 | -1.623 | -0.623 | -0.623 | -1.623 | -0.038 | -0.623 | -1.623 | -0.623 | -1.623 | 1.377 | -1.623 | 0.377 |
| c | -0.275 | -0.275 | 0.310 | -0.275 | 0.725 | 0.310 | 0.725 | -1.275 | -1.275 | -0.275 | 1.310 | -1.275 | -1.275 | -1.275 |
| g | 0.725 | -1.275 | 0.310 | -1.275 | -0.275 | -0.275 | -1.275 | -0.275 | -0.275 | -1.275 | -1.275 | -1.275 | 1.725 | 1.047 |
| t | 0.377 | 1.184 | 0.377 | 0.962 | -0.038 | 0.699 | -0.038 | 0.962 | 1.184 | 0.962 | -0.038 | -1.623 | -1.623 | -1.623 |

5) We will use Table 4 to score the seven 14-mers for the 3' splice sites. Complete Table 5 where the score for $Y_1$ has already been entered.

Table 5: Scores of the seven 5' splice sites of the MOG gene with PWM of Table 4

| Sequence | Score |
|---|---|
| $Y_1$ = GTGTCTTGGACAGG | 8.853 |
| $Y_2$ = TCCTGCCTTTCAGA | |
| $Y_3$ = TTTTCTATTTTAGG | |
| $Y_4$ = GTTTCTCTTTCAGA | |
| $Y_5$ = GTGAACAATTCAGA | |
| $Y_6$ = TTTTTGTTTTCAGG | |
| $Y_7$ = CTCCTTCTTCTAGG | |

6) You are going to randomly choose seven 14-mers from the MOG sequence (Z48051 at NCBI) that are <u>not</u> 3' splice sites (but have the invariant AG in positions 12 and 13) and score them.

Table 6: Scores of 7 randomly chosen 14-mers with AG (pos. 12 and 13) with PWM of Table 4

| Random Sequence | Start | End | Region | Score |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

7) Check out the 14 scores you obtained in Tables 5 and 6 and decide on a good cutoff value that can be used as threshold. How many false positives and false negatives do you have?