Task 5

Alex, Alex and Helena

11/24/2021

The file *MA00081.sites.fasta* contains **25 sequences** with a **regulating factor of the HOMEO family** in *Arabidopsis thaliana*. The regulation factor in each sequence is indicated in capital letters, thus in the first sequence it is CAATTATT. The different tasks to be carried out are presented below:

1. Extract the subsequence with capital letters from each sequence and align them all.
2. Perform the representation of the multiple alignment as a logo sequence. Can be done with R with the seqLogo package (bioconductor). Discuss the results.
3. Create the absolute frequency matrix. Present the consensus sequence. Is there the consensus sequence between the original sequences?
4. Obtain the matrix of relative frequencies and calculate KL divergence.
5. Calculate thew log likelihood for each of the 25 subsequences displayed in uppercase.
6. Choose a complete sequence of the 25, that is, with 15 nucleotides, and represent graphically the log of the PSSM plausibility when moving through the entire sequence (see fig 10.1 "Introduction to Computational Genomics. A case studies approach" of N.Cristianini & M.W. Hann)
7. Write three sequences where the PSSM matrix has a low score, that is, the sequences are negative from the point of view that the PSSM matrix represents a discriminator.

1. Extract the subsequence with capital letters from each sequence and align them all.

Our file provides a list of 25 sequences. Let us see how the sequences look when we read the file:

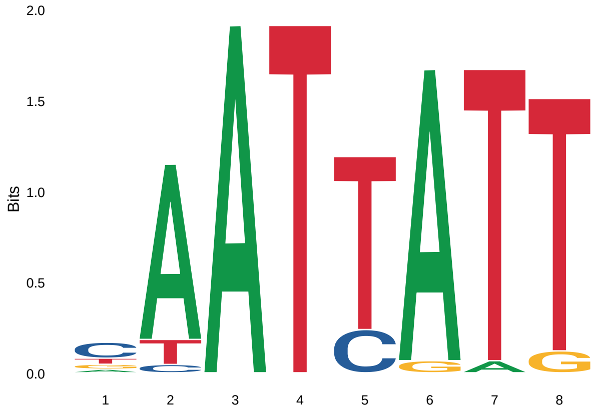
```
 [1] ">MA0008\tAthb-1\t1"  "gagaagaCAATTATT"     ">MA0008\tAthb-1\t2" 
 [4] "gttggttCAATTATT"     ">MA0008\tAthb-1\t3"  "CAATTATTgcagaga"    
 [7] ">MA0008\tAthb-1\t4"  "tgacattCAATTATT"     ">MA0008\tAthb-1\t5" 
[10] "caaatcaCAATTATT"     ">MA0008\tAthb-1\t6"  "CAATTATTacgggcg"    
[13] ">MA0008\tAthb-1\t7"  "CAATTATTgcaaagc"     ">MA0008\tAthb-1\t8" 
[16] "cataCAATTATTgga"     ">MA0008\tAthb-1\t9"  "cgccgtgCAATCATT"    
[19] ">MA0008\tAthb-1\t10" "tgCAATCATTagctg"     ">MA0008\tAthb-1\t11"
[22] "ttagaccCAATCATT"     ">MA0008\tAthb-1\t12" "acatgCAATCATTca"    
[25] ">MA0008\tAthb-1\t13" "ctcgtctCAATCATT"     ">MA0008\tAthb-1\t14"
[28] "aactcGAATTATTgg"     ">MA0008\tAthb-1\t15" "cttGAATTATTggat"    
[31] ">MA0008\tAthb-1\t16" "accagggTAATTATT"     ">MA0008\tAthb-1\t17"
[34] "agTAATTATTgcttg"     ">MA0008\tAthb-1\t18" "TAATTATTgcacaac"    
[37] ">MA0008\tAthb-1\t19" "tcccAAATTATTgca"     ">MA0008\tAthb-1\t20"
[40] "attaAAATTATTgca"     ">MA0008\tAthb-1\t21" "ggcttgAAATTATTg"    
[43] ">MA0008\tAthb-1\t22" "cagcGTATTATTgca"     ">MA0008\tAthb-1\t23"
[46] "tggcacgGCATTATT"     ">MA0008\tAthb-1\t24" "aaTTATTATGcggtc"    
[49] ">MA0008\tAthb-1\t25" "aTTATTGAGcgcgaa"    
```

The regulatory regions (or promoters), which are in capital letters in the original fasta file are the following (shown in table):

	Regulatory sequence
1	CAATTATT
2	CAATTATT
3	CAATTATT
4	CAATTATT
5	CAATTATT
6	CAATTATT
7	CAATTATT
8	CAATTATT
9	CAATCATT
10	CAATCATT
11	CAATCATT
12	CAATCATT
13	CAATCATT
14	GAATTATT
15	GAATTATT
16	TAATTATT
17	TAATTATT
18	TAATTATT
19	AAATTATT
20	AAATTATT
21	AAATTATT
22	GTATTATT
23	GCATTATT
24	TTATTATG
25	TTATTGAG

2. Perform the representation of the multiple alignment as a logo sequence. Can be done with R with the seqLogo package (bioconductor). Discuss the results.

Here, we show the multiple alignment as a logo sequence of the 25 sequences.



It can be observed that this regulatory region is pretty similar and follows a pattern in the 25 sequences. There is a notable predominance of a specific nucleotide in mostly all positions, with exception of the first one. We could read that the motif would be XAATTATT. In the first position, there is no nucleotide substantially appearing with more frequency than the rest.

3. Create the absolute frequency matrix. Present the consensus sequence. Is there the consensus sequence between the original sequences?

The absolute frequency matrix is basically obtained by counting the frequencies of nucleotide in each position. In this case, the frequency matrix would be like this:

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
A    3   21   25    0    0   24    1    0
C   13    1    0    0    5    0    0    0
G    4    0    0    0    0    1    0    2
T    5    3    0   25   20    0   24   23
```

The consensus sequence is a new sequence formed by the most frequent letter used at each position. Therefore, observing the frequency matrix, we would suggest that the consensus sequence is: CAATTATT. This sequence is found in the first 8 sequences of our sample.

4. Obtain the matrix of relative frequencies and calculate KL divergence.

Since we are estimating a probability based on a small sample, and it is quite possible that we never observe symbols that have small probability, we will add what is called a **pseudocount** to each entry of the matrix. This way, we are making the probabilities unlikely but not impossible.

Hence, the relative frequency matrix is obtained by adding 1 and dividing by the total of characters in each position:

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
A 0.14 0.76 0.90 0.03 0.03 0.86 0.07 0.03
C 0.48 0.07 0.03 0.03 0.21 0.03 0.03 0.03
G 0.17 0.03 0.03 0.03 0.03 0.07 0.03 0.10
T 0.21 0.14 0.03 0.90 0.72 0.03 0.86 0.83
```

The KL divergence measures how different the motif is from the background distribution, which we assume is uniform in the PSSM. This number is calculated using the equation given in page 165 in Cristianini & Hann's "Introduction to Computational Genomics. A case studies approach". The value of the KL of this motif is

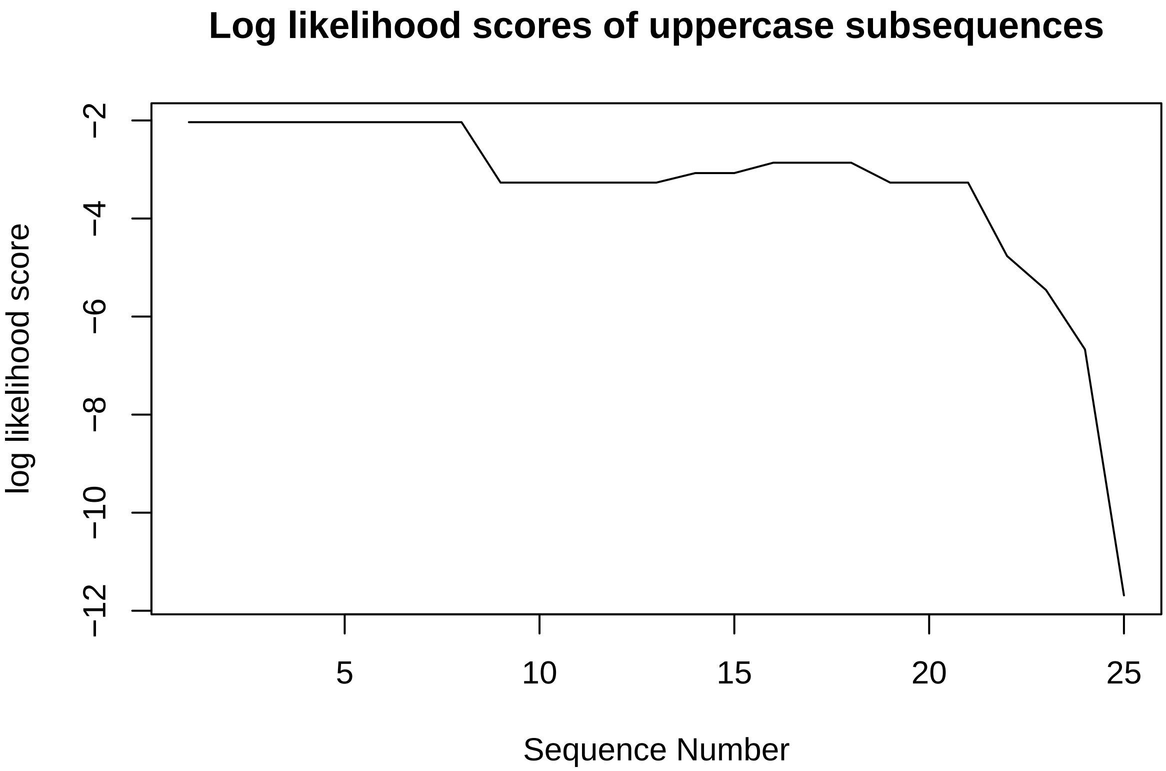
```
[1] 5.732189
```

5. Calculate the log likelihood for each of the 25 subsequences displayed in uppercase.

The way we interpret this problem is that we are asked calculate the log-likelihood for each of the 25 uppercase subsequences that we extracted in problem 1.

log likelihood	Promoters
-2.03560647673031	CAATTATT
-2.03560647673031	CAATTATT
-2.03560647673031	CAATTATT
-2.03560647673031	CAATTATT
-2.03560647673031	CAATTATT
-2.03560647673031	CAATTATT
-2.03560647673031	CAATTATT
-2.03560647673031	CAATTATT
-3.26775015802294	CAATCATT
-3.26775015802294	CAATCATT
-3.26775015802294	CAATCATT
-3.26775015802294	CAATCATT
-3.26775015802294	CAATCATT
-3.07359414358198	GAATTATT
-3.07359414358198	GAATTATT

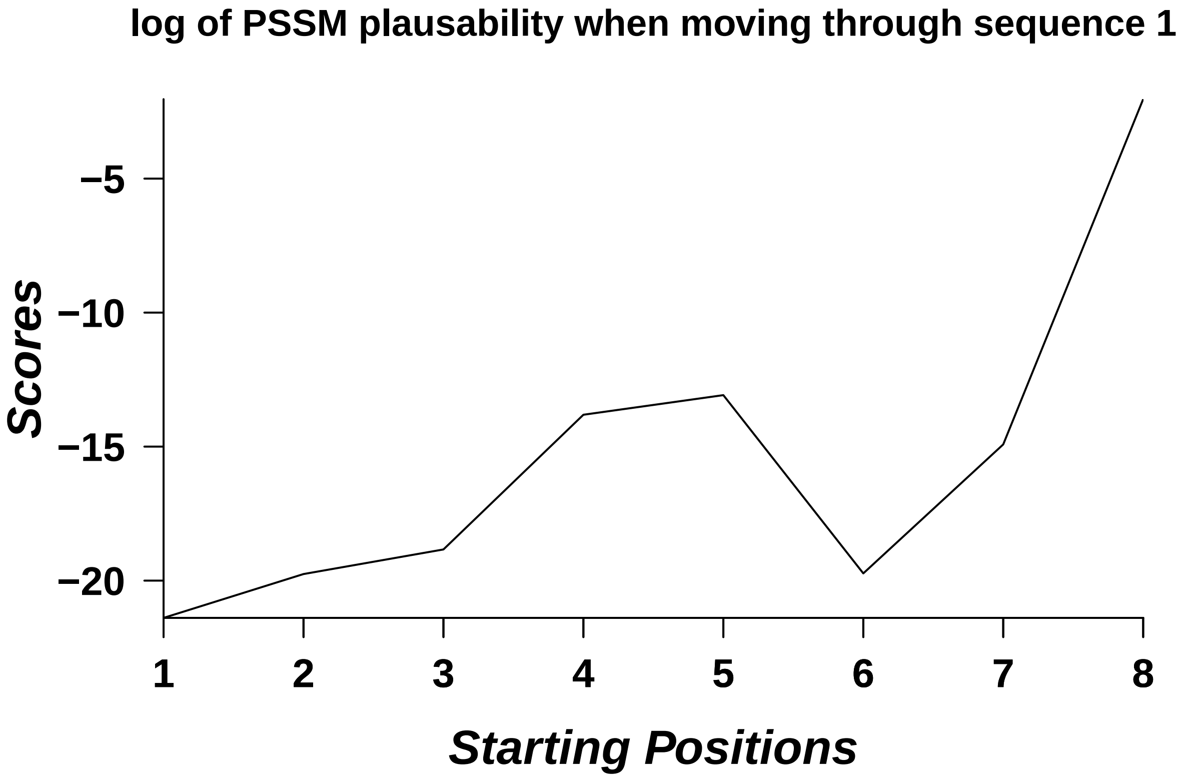
log likelihood	Promoters
-2.86228504991478	TAATTATT
-2.86228504991478	TAATTATT
-2.86228504991478	TAATTATT
-3.26775015802294	AAATTATT
-3.26775015802294	AAATTATT
-3.26775015802294	AAATTATT
-4.76527015425306	GTATTATT
-5.458417334813	GCATTATT
-6.6702165753884	TTATTATG
-11.6870908697848	TTATTGAG



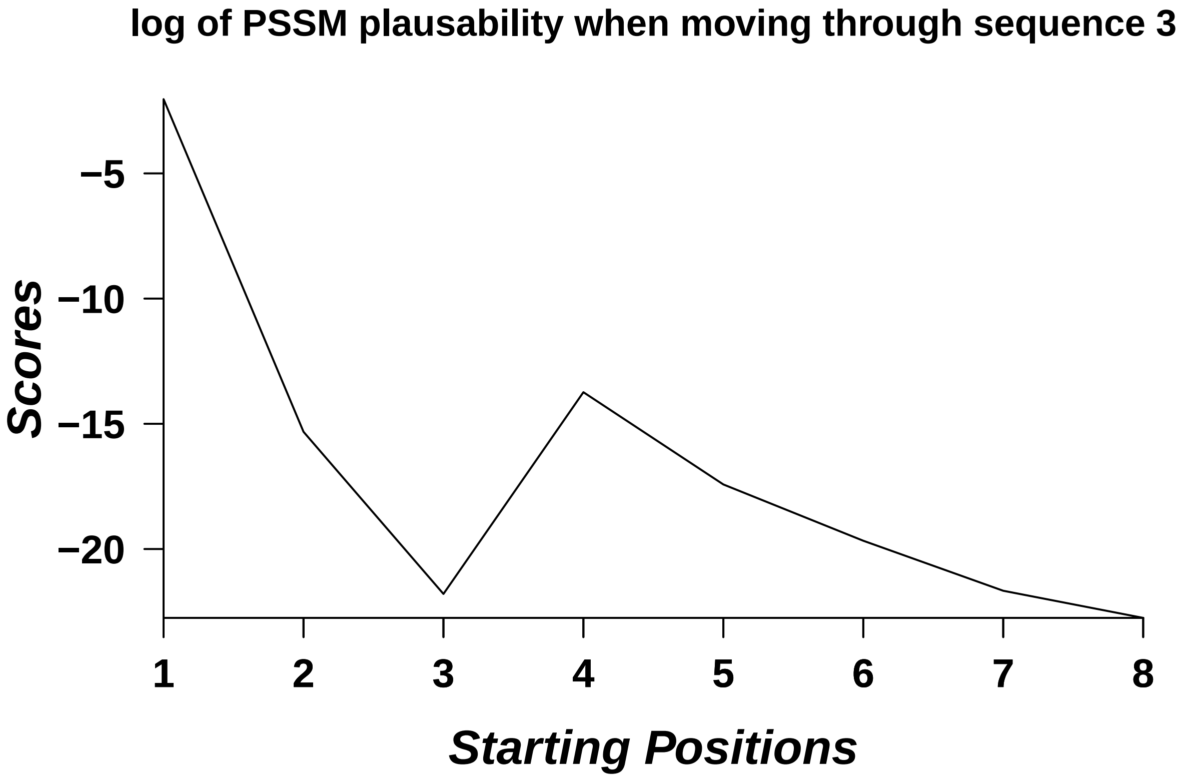
6. Choose a complete sequence of the 25, that is, with 15 nucleotides, and represent graphically the log of the PSSM plausibility when moving through the entire sequence (see fig 10.1 "Introduction to Computational Genomics. A case studies approach" of N.Cristianini & M.W. Hann)

We select two different sequences for illustration purposes: the first and the third.

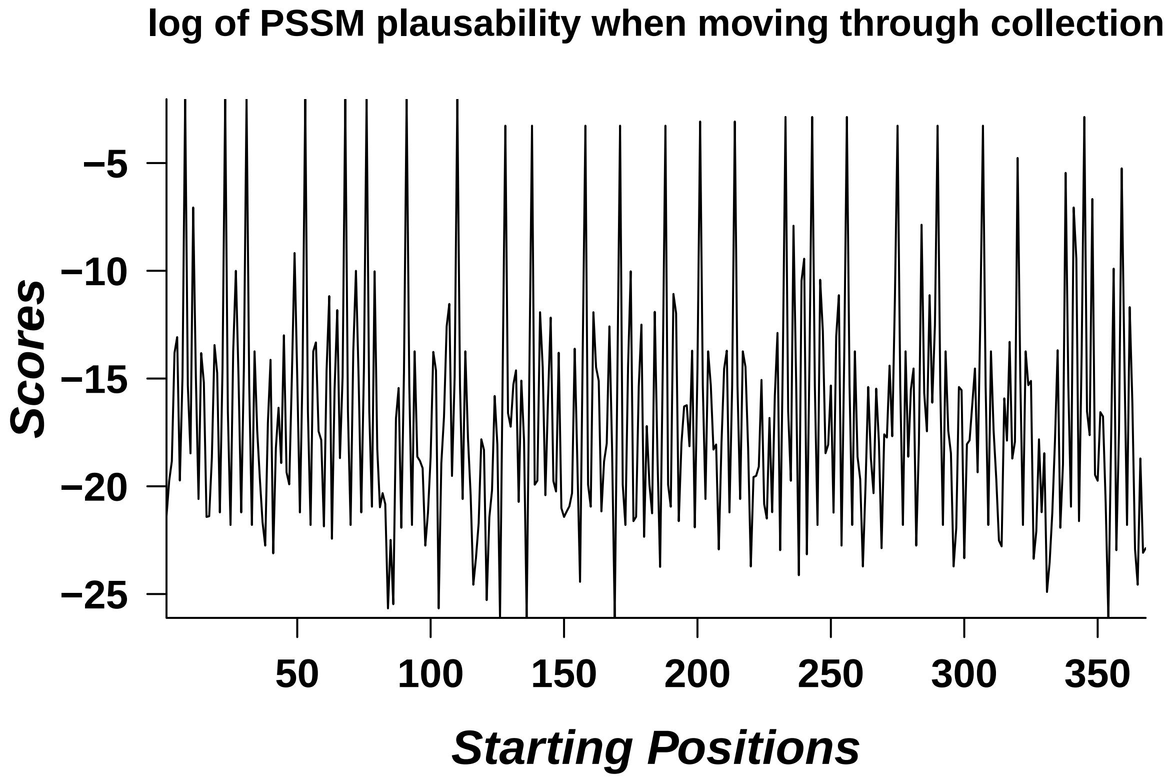
First, we analyze the first sequence GAGAAGACAATTATT.

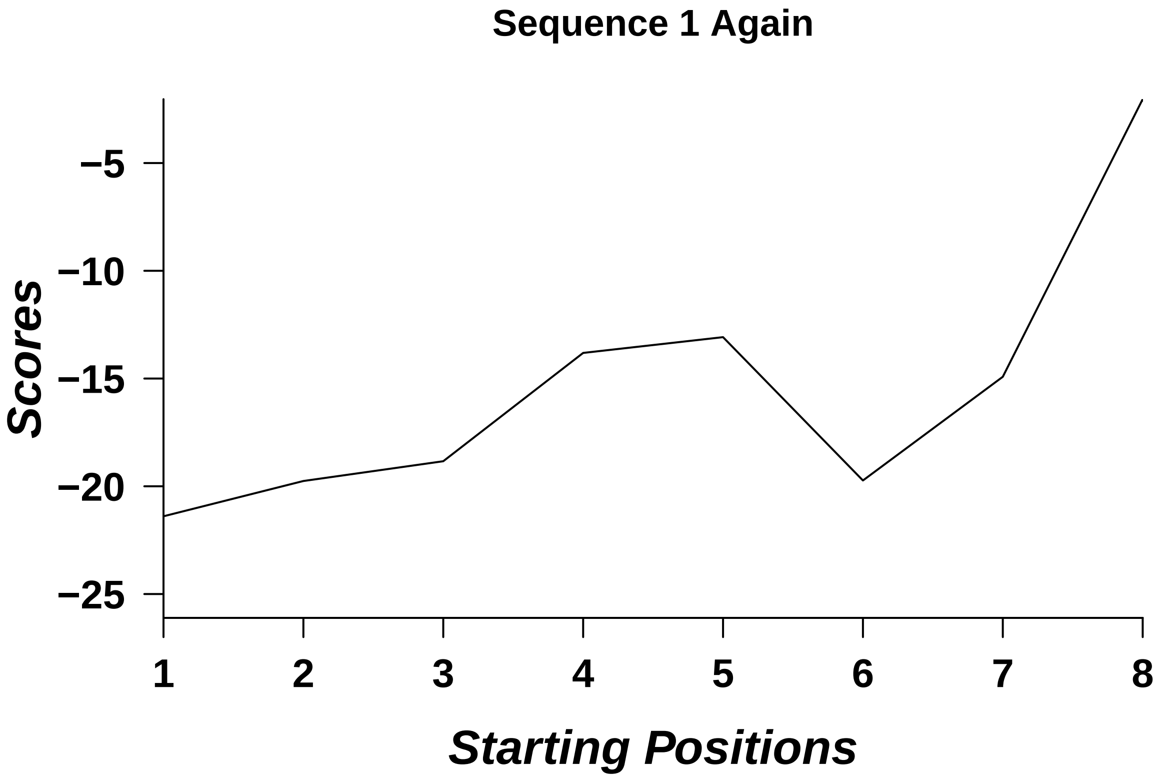


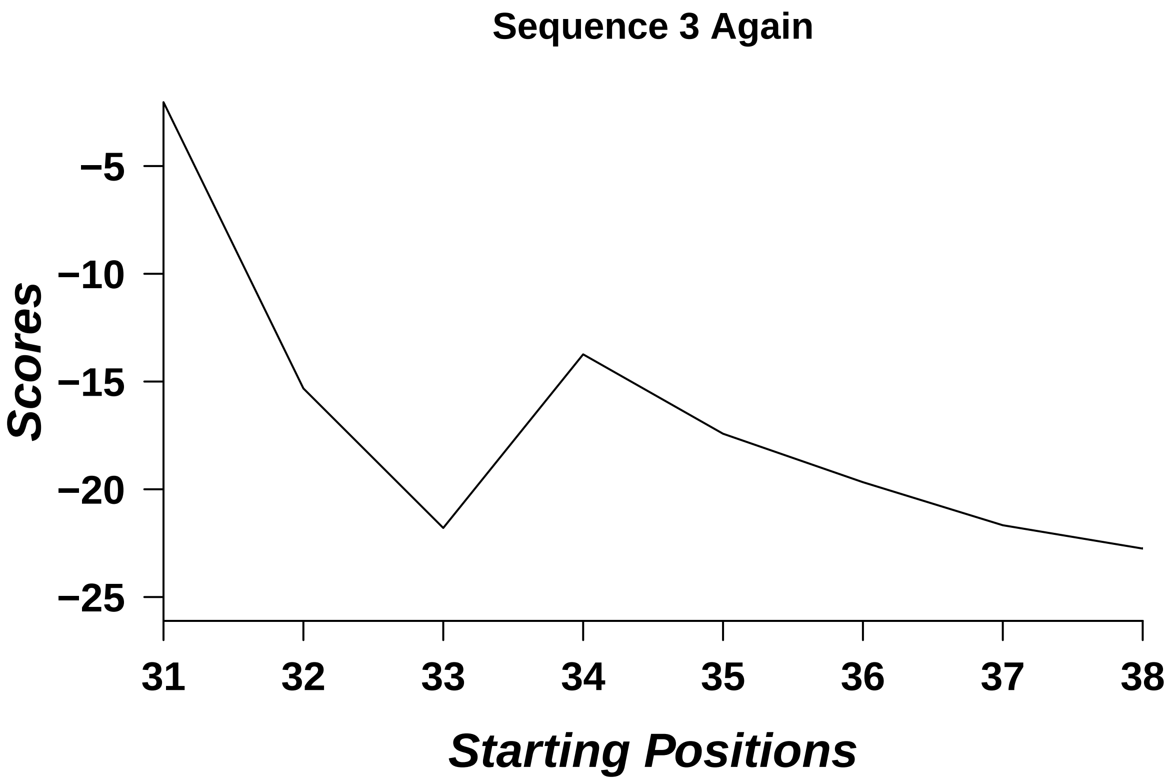
Next, we analyze the third sequence CAATTATTGCAGAGA.



Just to check, let us concatenate all the strings and do the same exercise as in the book.


log of PSSM plausability when moving through collection
Scores
−5
−10
−15
−20
−25
50
100
150
200
250
300
350
Starting Positions


Sequence 1 Again
Scores
−5
−10
−15
−20
−25
1
2
3
4
5
6
7
8
Starting Positions



7. Write three sequences where the PSSM matrix has a low score, that is, the sequences are negative from the point of view that the PSSM matrix represents a discriminator.

In order to create sequences with a low score, we should take those nucleotides that are a low probability to appear. For example, the sequences could be: AGCAACCC, AGGGGCGC and AGGCATCC.

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
A 0.14 0.76 0.90 0.03 0.03 0.86 0.07 0.03
C 0.48 0.07 0.03 0.03 0.21 0.03 0.03 0.03
G 0.17 0.03 0.03 0.03 0.03 0.07 0.03 0.10
T 0.21 0.14 0.03 0.90 0.72 0.03 0.86 0.83
```