Assignment 4

- In motif finding, a weight matrix (also referred to as Position Weight Matrix or Position Specific Weight Matrix) is defined as the log-odds matrix whose elements are defined as

  $$W(b, j) = \log [F'(b, j) / F(b, o)]$$

  Where b corresponds to the base and j is the index accounting for the number of bases in the motif. F' (b, j) corresponds to the frequency with which each base occurs at a specified position and can easily be calculated from the counts matrix after adjusting for zero values (see below). F(b, o) corresponds to the background frequency with which a particular base is known to occur and can be assumed to be 0.25 for all bases at all positions in the motif.

  A transcription factor argR is known to bind to a motif which can be represented with the following counts matrix built from a total of 27 binding sites documented in the literature (the counts matrix is attached as a text file which should be used by your program).

a	|	8	12	21	9	4	2	21	21	3	10
	8	5	7	25	4	2	2	25			
c	|	7	4	1	6	2	3	3	3	1	0
	2	0	7	0	3	3	24	0			
g	|	3	2	1	8	2	21	2	2	0	1
	0	1	0	1	0	15	0	2			
t	|	9	9	4	4	19	1	1	1	23	16
	17	21	13	1	20	7	1	0			

  Now write a script/program to compute the frequency matrix F(b, j) using the above counts matrix. Since log odds matrix is based on frequency matrix, to avoid taking logarithm of 0 in computing it, a revised F' (b, j) can be computed by augmenting all the base counts in counts matrix by 1

thereby artificially increasing the number of sites to 31 (put another way, a pseudocount of +1 is added to each of the real counts for each base at each position, which increases the total counts at each position in the matrix to 31). Based on this notion, compute the F’ (b, j) as well in the same script/program.

- Now use the weight matrix to scan and identify the binding sites in the attached set of upstream regulatory regions of genes by filtering to those with highest similarity to the PSM i.e, your program should predict and show only the top 30 scoring gene ids corresponding to these sequences. Upstream regulatory regions of genes defined as 400 bases upstream and 50 bases after the translational start site are provided in the fasta nucleotide format along with information about the gene id to which it corresponds to.