DosR bonding sites in MTB

Alejandro Amador

2020/18/09

It is somewhat a weird challenge since this was already partially worked in the final part of the chapter. We can, however, add some extra functionality to the code so that we may replicate and extend slightly the discussion on the book. We proceed by adding just some extra functions to find the consensus k-mer, to do this we merely obtain the count matrix from the best motifs found (by either randomized or Gibbs sampling) and produce the consensus string.

In the chapter they do not mention how many iterations they used for randomized/Gibbs algorithms but here I used 5,000 runs for both of them, obtaining in each one the same consensus strings. Also, I do not know how this information can be formally translated into knowing what genes are regulated in this biological process, but I do think that the genes that are related to the k-mer should be the ones in which this k-mers appears the most. Because of this, I also counted how many times the consensus motif appeared in each of the 10 genes with, at most, four mismatches. The results can be seen at Table 1.

k	**Consensus**	**Gene 1**	**Gene 2**	**Gene 3**	**Gene 4**	**Gene 5**	**Gene 6**	**Gene 7**	**Gene 8**	**Gene 9**	**Gene 10**
8	**CATCGGCC**	38	44	42	47	43	48	42	42	34	41
9	**GTGGCCGCC**	26	17	28	17	16	37	19	19	24	24
10	**CTATCGGCCC**	9	12	12	8	8	11	8	10	10	9
11	**GGACTTCCGGC**	3	5	5	4	2	8	5	4	3	4
12	**GGACTTCCGGCC**	2	3	4	3	1	4	2	1	2	3
13	**GGACTTCCGGCCC**	2	2	2	3	1	3	2	1	2	1
14	**GGGACCTACGTCCC**	2	2	1	0	2	1	1	1	2	1
15	**GGACTTACGGCCCTA**	1	1	1	0	1	1	1	1	2	0
16	**GGGACCTACGTCCCTA**	1	1	1	0	1	1	1	1	1	1
17	**GGGACCTACGTCCCTAG**	0	1	1	0	1	0	1	0	1	1
18	**GGGACCTACGTCCCTAGC**	0	0	1	0	1	0	1	0	1	0
19	**GGGACCTACGTCCCTAGCC**	0	0	1	0	1	0	1	0	0	0
20	**GGGGACCTACGTCCCTAGCC**	0	0	0	0	1	0	1	0	0	0
21	**CGGGACCTACGTCCCTAGCCG**	0	0	0	0	1	0	0	0	0	0
22	**GGGACCTACGGCCCTAGCCCCG**	0	0	0	0	0	0	0	0	0	0

Table 1: Consensus k-mers found by both randomized/Gibbs and their corresponding appearances in each gene with, at most, four mismatches.

We can appreciate embedding of sub-strings into the longer k-mers. Moreover, as k goes to higher values, the consensus appears in less genes, which could indicate that, in such genes, the pattern does indicate a related biological function or a so-called *hidden message* and not just coincidental appearances of the k-mers. With this reasoning, it could be inferred that genes 3,5,7 and 9 are the regulated ones.