CS 309: Computational Biology

Lab 2: DosR binding site motif search in Mycobacterium tuberculosis

By: Amna Khalid, Ebo Dennis

Number 1:

N = 100

Best Motifs =
['AGATCACCGGGTTTCCCCGG', 'CGACCACCGCGCCGGAGCCC',
'ATCGGAGGGCTCCGGCGCGG', 'AATGCCGCCGTTGGGCCGCG',
'ATTGGCCGCGGGCGGCCCGC', 'AGCCGCTCGCGGTGGCGGGC',
'GCAGCATCGGTTCGGCGCGG', 'ATCACCTCAATCCGATGACC',
'GCACATGGCCGCCGACCCGC', 'CCACGAGGCCGCCGCCGGCC',
'GGACGTCCGCGACGACGCGT', 'CCCCTACCCTGGCTGCCCCG',
'GATCATCGGCCAGGGCGCCG', 'GCACCGGCGCTTTGGCGAGG',
'GGTGGTGCCCACCCGCGCGG', 'ACCGGATCGACACGGCGGGG',
'GATGGTGGGCGGCGGTCCGG', 'CATGATTGGATTCGACGCCC',
'CGTGAATCCCGCTGGCGGTC', 'AGTGGATCACCGACGGGCGC',
'AGGGACCGCGGGTGGCGCTG', 'CGTGCCGCGCGTCGGCGAGT',
'TCTGCTCGCCCATGGCCCTC', 'TGAGGTGGCCTACGGCGAGG',
'CGATGTGCCCGGTCGCGCCG', 'GGTCGCGGCCGATGGCGTCC',
'GCACATACCCTTCCGCGTCG', 'GGTCGCCACGGCTGGCGATG',
'CGTCGTCGGCCTCGGCGTCG', 'CGACGGTCACCATGTCGCGG',
'GATCACCGGGGAAGGCGCTG', 'GGTCAGCGCCTTCCCCGGTG',
'ATACCAGCCCGGATGCGCCG', 'GGTGGGGACCAACGCCCCTG',
'CGTGATAACGCGCGGCGCCG', 'CGACGTGCCGGTGCCAGCCG']

Number 2:

N = 200

Best Motifs =
['GTCCCCAGCCCCAAGGCCGA', 'AGTCCCGGCCACTGGGGTCA',
'AACCCTGGCTTCGATGGCGC', 'GGCCCTGTCCGCGTCCGTGT',
'CGCCGTAGCCGATTGGCCGC', 'CGCCGTTGCGCCGGGTGCGT',
'AATGCGGCCGACGAGCGGGC', 'GGCCCGGCCGCCATCGGCCC',
'CACCCCGGGGAGCACGTTCC', 'GGCCCCACCCACGAGGCCGC',

'CGCCGTGGCGGTGACAACGA', 'CGCGGCGGCGGCGATACCCC',
'GGCCAGGGCGCCGGGCTTCC', 'GACCTGAGCCAGTTCACCCC',
'CACCCGCGCGGAAAGCCCCG', 'CGGCGGGGCCGCGTGAGCGC',
'GGCCTTGTCGACCACGTTGT', 'TACCCCGACGACAACTCCGT',
'CCCCGCGTCGACGTGCCAGT', 'AGCCCTGGCCACGATGGGCT',
'AGCGCGGGCCATTTGTCCGC', 'AGCCGTCGCCATTGTCGCGC',
'CACCGAAGCCGACATCGCCC', 'CCTGCGGCCGCCCATAGCGC',
'CGCCGCGTCCACCTCGGTCA', 'CAGCCCGGCCAGCACGCCGT',
'AACCCGTGGGCACATACCCT', 'TGCCGTGCCGAAGGCGGTGT',
'GGCGTCGGCCAACGCCGCGA', 'CACGACGGTCACCATGTCGC',
'GGCCCGGACGGCCAGGGTGA', 'CACCGCGGCTGTCGACGCGC',
'CACCTGAGCCGTCGGCTCGC', 'CGCCCTGACGGGTTGCGTCT',
'AGCGGCGGCGCAAACCGCGT', 'CACGCTGCCGAATATGACCC']

N = 1000

Best Motifs =
['TGACCGACGTCCCCAGCCCC', 'CGACCACCGCGCCGGAGCCC',
'TATCGGAGGGCTCCGGCGCG', 'CATTGGCCACAATCGGGCCA',
'GATTGGCCGCGGGCGGCCCG', 'CGCTCGCGGTGGCGGGCCGT',
'CGACGGTGATTCCGGGTCCG', 'GCCCGGCCGCCATCGGCCCG',
'CCATCGCCGCGGTCAAGCCG', 'TGACCACCGCCGACGGGCAC',
'GCACCATGACCGCCTGGCCA', 'GCCCGGTCGCCACGCGGCGG',
'CGATCATCGGCCAGGGCGCC', 'GGACCATCGCCTCCTGACGC',
'GCAACGTCGGGCCGGGCCAG', 'GGATCGACACGGCGGGGCCG',
'CGATGGTGGGCGGCGGTCCG', 'TCATCGCCGCATCGGTGGCA',
'CGATCCTCGGCATCGGGCCG', 'TCACCGACGGGCGCGGACAA',
'AGCCCATCGTGGCCAGGGCT', 'CCATTGTCGCGCACAAACCG',
'CATCGCCCGACACCTGCCCG', 'GGTCCATCGACCCGCGGCCC',
'CATCGGCCGCGACCAAGCCC', 'CATCGACGGTCCCCGGGGCT',
'AACCCGTGGGCACATACCCT', 'TGACGAAGGCGACATGGCCC',
'CGATCACCGTAACAGGACCG', 'TCACGACGGTCACCATGTCG',
'TGTTCATGGCTGCCGGGCCT', 'CCACCGCGGCTGTCGACGCG',
'CCATCGCCGGCGGCAGTGCG', 'GAACCGACGGGATGTATCCG',
'CTATCGCCGGGGCTGGGCCG', 'GCTCCGACGTGCCGGTGCCA']

N = 2000

Best Motifs =

['CGTGACCGACGTCCCCAGCC', 'CGGCGCCATCGAAGCCAGGG',
'GGGCTCCGGCGCGGTGGTCG', 'TGCCGCCGTTGGGCCGCGGA',
'CGTCACCGGCATCCGCATCG', 'CGGCGCTGGTGGCCTACTGC',
'TGCGGCCGACGAGCGGGCGC', 'GGCCGCCATCGGCCCGTCGA',
'CGTCGATGACGGCGCAGTAC', 'CGCCGGCAGCGTGCGTGACG',
'GGGGACCGAAGTCCCCGGGC', 'CTCCGCTGGCAGCCCGAGCG',
'CATCGGCCAGGGCGCCGGGC', 'CGTCGGTAGTACACCCATGC',
'GGGCAACGTCGGGCCGGGCC', 'CGGCGGCCTTGGCCGCCCGG',
'GGTGGGCGGCGGTCCGGTCA', 'AGGCGGCGGCATCCGATTCG',
'CCTCGGCATCGGGCCGGTAG', 'GATCACCGACGGGCGCGGAC',
'GGGGACCATTGACCCTGTTG', 'CGTCGCCATTGTCGCGCACA',
'CGCCAGCATCGGAGGTACCC', 'CGGGTCCATCGACCCGCGGC',
'GGACGCCATCGGCCGCGACC', 'GGCCGATGGCGTCCTCATCG',
'CTTCCGCGTCGTACTGGTCA', 'GGTGGGTGCCGTGCCGAAGG',
'CGTCGGCCTCGGCGTCGGCC', 'GGGCAGCGTTGCACTCGGTC',
'CGCCAGCGGAGGACCTTTGG', 'GGTCAGCGCCTTCCCCGGTG',
'CGCCGGCGGCAGTGCGTGCC', 'CGCCAGTAACGTACCGCTGA',
'CGGCGGCGCAAACCGCGTTC', 'GGGCTCCGACGTGCCGGTGC']

Number 3

```
best_motifs, bestScore = getBestMotifSet(dna, 20, 36, 300, 2000)
```

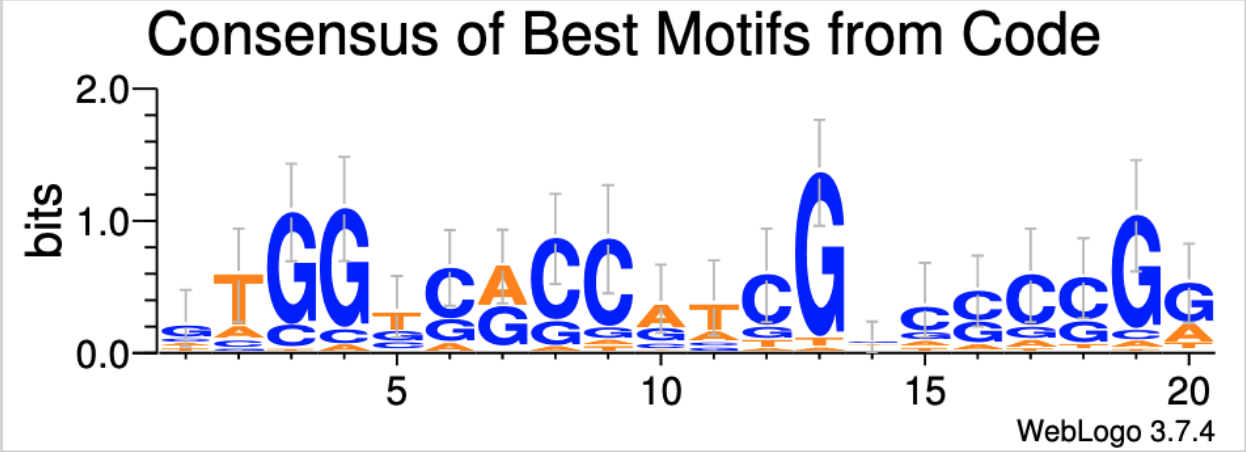
['TTCGTGACCGACGTCCCCAG', 'GCCGGCGCCATCGAAGCCAG',
'CTGGTCGCCACTGGAAAGGG', 'TTGGCCACAATCGGGCCACA',
'GTGGTCGCGATCGAACCCGA', 'TTGGCCACCGGCGCTGGTGG',
'GTGGCGACGGTGATTCCGGG', 'CCGGCCGCCATCGGCCCGTC',
'AAGCAAACCATCGAACCCGG', 'CTGGTGACCACCGCCGACGG',
'CTGGGGACCGAAGTCCCCGG', 'CTGGCAGCCCGAGCGGGGGG',
'ATCATCGGCCAGGGCGCCGG', 'GTGGTCGACAAGGTCGCCGA',
'TTGGGCAACGTCGGGCCGGG', 'GCGGCGGCCTTGGCCGCCCG',
'ATGGTGGGCGGCGGTCCGGT', 'GAGCGAACCATCTACCCCGA',
'ATCCTCGGCATCGGGCCGGT', 'GTGGTCACCATGGTGTCCGG',
'TTGGGGACCATTGACCCTGT', 'GCCGTCGCCATTGTCGCGCA',
'TTGGTCGGAATCGTCACCGA', 'GAGGTGGCCTACGGCGAGGA',
'GAGGACGCCATCGGCCGCGA', 'GTGGCCACTGTCGAGACCGG',
'CTGGTCAGTCTCGACAGCGA', 'AGGGTCGCCACGGCTGGCGA',
'GTCGTCGGCCTCGGCGTCGG', 'CTGGGCAGCGTTGCACTCGG',
'GGGGGCCCGGACGGCCAGGG', 'ATGGTCAGCGCCTTCCCCGG',

'GATCGGGCCATCGCCGGCGG', 'GTGGGGACCAACGCCCCTGG',
'GTGATAACGCGCGGCGCCGG', 'GAGGGCTCCGACGTGCCGGT']

Score = 269

Number 4:
Consensus = GTGGTCACCATCGGCCCCGG

Number 5:

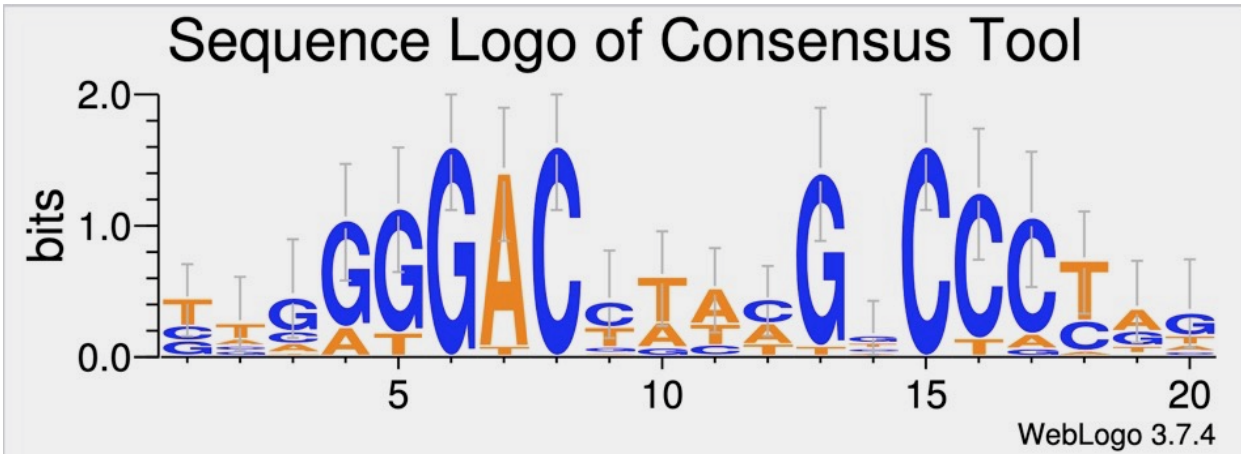


Number 6:

Best Motifs =
[‘TTCGTGACCGACGTCCCCAG’, ’TTGGGGACTTCCGGCCCTAA’,
’GCCGGGACTTCAGGCCCTAT’, ’CATGGGACTTTCGGCCCTGT’,
’GAGGGGACTTTTGGCCACCG’, ’CCAGGGACCTAATTCCATAT’,
’TTGAGGACCTTCGGCCCCAC’, ’CTGGGGACCGAAGTCCCCGG’,
’TTAGGGACCATCGCCTCCTG’, ’TGGATGACTTACGGCCCTGA’,
’TTGGGGACTAAAGCCTCATG’, ’TCGGGGACTTCTGTCCCTAG’,
’TTGGGGACCATTGACCCTGT’, ’TTGAGGACCTAAGCCCGTTG’,
’CACGGGTCAAACGACCCTAG’, ’GGCGGGACGTAAGTCCCTAA’,
’GAAGTGACGAAAGACCCCAG’, ’CGGAGGACCTTTGGCCCTGC’,
’GTGGGGACCAACGCCCCTGG’]

Score = 118

Consensus = ‘TTGGGACCTACGCCCCTAG’



Number 7:

There are more fully conserved positions in the sequence logo from the consensus tool compared to the sequence tool of our Gibbs Sampler code where no position is fully conserved. In addition, the difference between the scores is very large (118 for consensus tool vs 269 using our Gibbs sampler code). Furthermore, our Gibbs sampler code returns the best motifs for all 36 given sequences, while the consensus tool returns only 19.

There is a pattern of 3 conserved G's and 3 conserved C's in the consensus tool graph. It takes an interesting pattern of increasing G's (position 4-6) and then decreasing C's (position 15-17). In the code's consensus graph, there is also a pattern of 4 conserved C's (position 15-18). Additionally, there are more conserved A's and T's seen in the consensus tool graph in comparison to the code's consensus graph.

In regards to the similarities, the more conserved positions in both sequence logos are mostly G's and C's. The positions with the same bases are:
2 - T, 3 - G, 4 - G, 7 - A, 8 - C, 9 - C, 12 - C, 13 - G, 15 - C, 16 - C, 17 - C and 20 - G.

On exploring why the Consensus only gave 19 motifs instead of 36 (which are the number of motifs the code generates), the assumption is that the consensus algorithm online reaches a convergence point on those motifs by only keeping motifs that help gain a better overall profile. The algorithm online also further removes some motifs that do not seem to fit into the constraints of best motifs i.e. it removes the outliers from the data to get an overall improved score.