UC San Diego

Assignment No. 1 – Arsham Mikaeili Namini

---

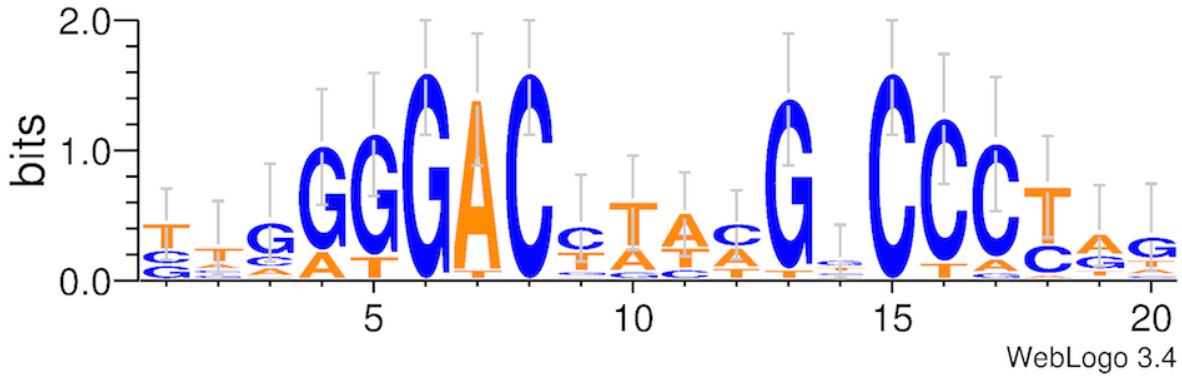
1. We will begin by running three different software tools on the dataset provided. First, upload upstream250.txt to Consensus (Hertz and Stormo, 1999). Set the desired pattern width equal to 20 (keep all other parameters the same) and click "submit".After the program has run, scroll to the bottom of the page and click "next". Under "Matrix 1", you will see 19 sequences corresponding to the substrings of the input strings having length 20 that are generated as a motif matrix. The elements in the column to the left of these sequences have the form XXX/YYY, where YYY represents the starting position of each sequence in the original string of length 250.

   Provide all of the starting positions of the strings of length 20.

   187 141 117 162 156 199 139 202 175 161 180 169 180 169 174 120 161 198 214 157 159

2. In order to visualize the information contained in these sequences, we will copy them into WebLogo to generate a motif logo.

   Upload the image file obtained after generating this motif logo (with default parameters).



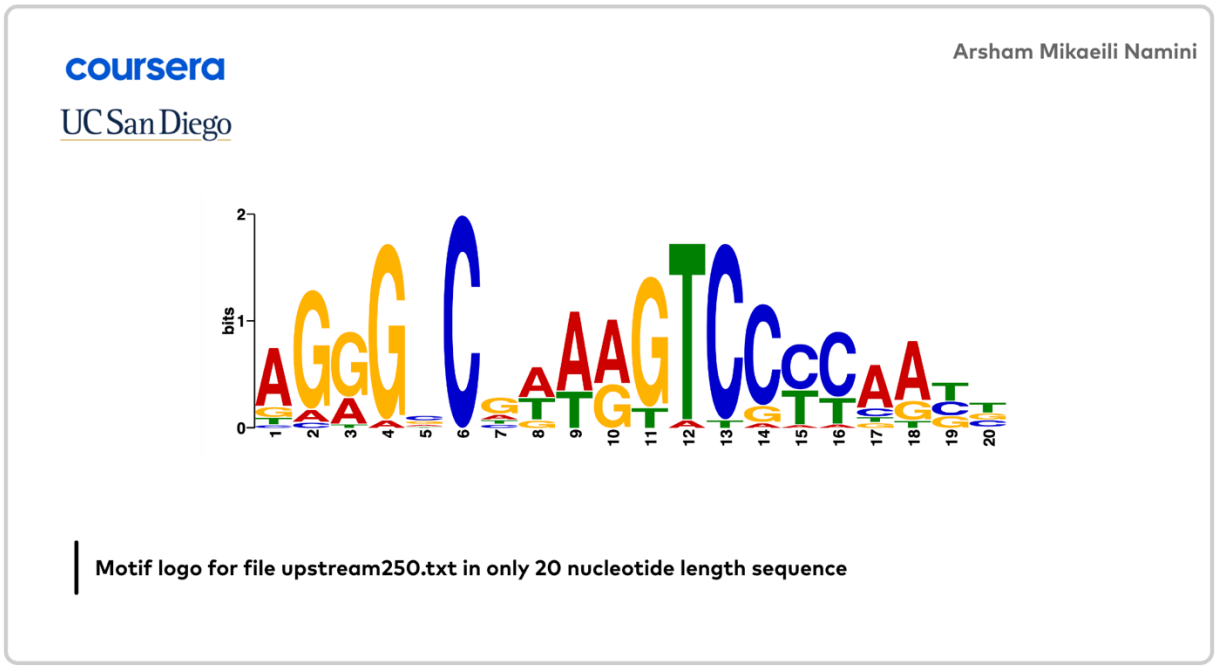
3. We will perform similar tasks with MEME (Bailey and Elkan, 1994). Upload upstream250.txt, and tell MEME to find 1 motif instead of 3. Then click on advanced options and change the minimum width to 20 and the maximum width to 20.

If the queue on the MEME server is too long, you can use alternate instance.

Indicate the starting positions of the substrings of length 20 identified by MEME.

57 139 107 172 172 114 136 159 143 155 186 178 200 118 137 173 201 160 62 216 165 45 204

4. **Upload the image file obtained after downloading this motif logo.**



5. **Did the programs generate similar motifs? Provide a brief (1-2 sentence) explanation.**

No, the programs generated mostly dissimilar motifs, with only vague similarities between the motifs generated by Consensus and MEME.

Although your biologist colleague told you that the motif is probably about 20 bp long, you are skeptical, so you decide to run a motif finding program that finds a motif over a wide range of different lengths.

Run MEME again on upstream250.txt, but this time, use the default parameters for minimum width (6) and maximum width (50). Note: this process may take a few minutes to run.

(a) How long is the motif produced by MEME?

(b) Is the motif logo produced by MEME similar to the one produced before for a motif of length 20?

a. The motif is also 20 nucleotides long. b. although it's not exactly the same they have similarities and it's noticeable that at the second run motif displace 1 nucleotide to the left.

6. **When using motif software with fixed motif lengths, is it better to start with short motifs or long motifs? Why?**

   generally, it's better to start with short motifs because long motifs are made of few short motifs and since our motif length is fixed, it's promising to search for short motifs first. on the other hand, the chance of finding good long motifs is quite low when there are no promising short motifs.

7. To evaluate the statistical significance of an identified motif, we need to ensure that a motif with the same or even larger score is unlikely to occur in a collection of "typical" DNA strings (of the same length).

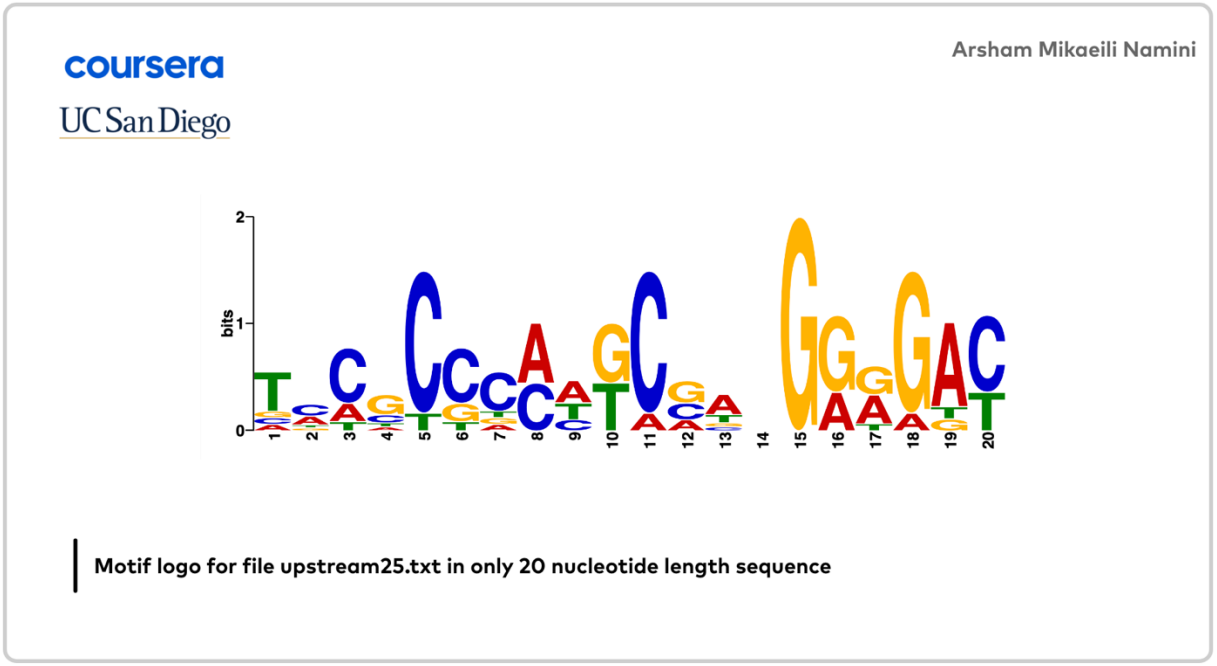
   How would you generate these strings? Justify your answer.

   - consider other known sequences of the same length having no motifs
   - randomly generate strings (ideally having the same GC-content as the species in question).

If the motif in question has a very low probability of occurring in randomly generated strings (or a low frequency in the known sequences), we can conclude that it is statistically significant.

8. We have begun to confirm our colleague's suspicion that we should consider motifs of length about 20. However, thus far, we have only analyzed the 250 bp regions upstream of each gene. This makes us wonder whether we will identify the same motif for upstream regions of different lengths. First, we will consider upstream regions of length 25 bp (upstream25.txt).

   Upload the motif logo obtained by running MEME on upstream25.txt. (You should specify that we are finding a single motif of length 20, as we did before.)



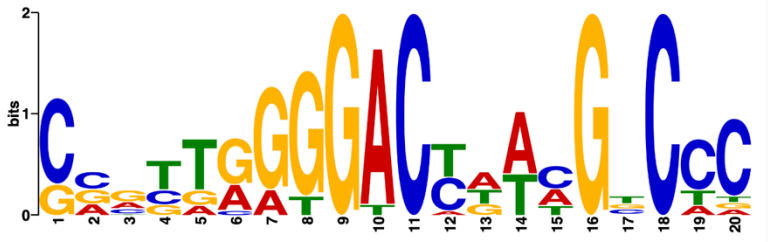
Next, we will consider upstream regions of length 100 bp (upstream100.txt).

Upload the motif logo obtained by running MEME on upstream100.txt. (Remember to specify in the options that we are looking for a single motif of length 20.)

coursera

UC San Diego

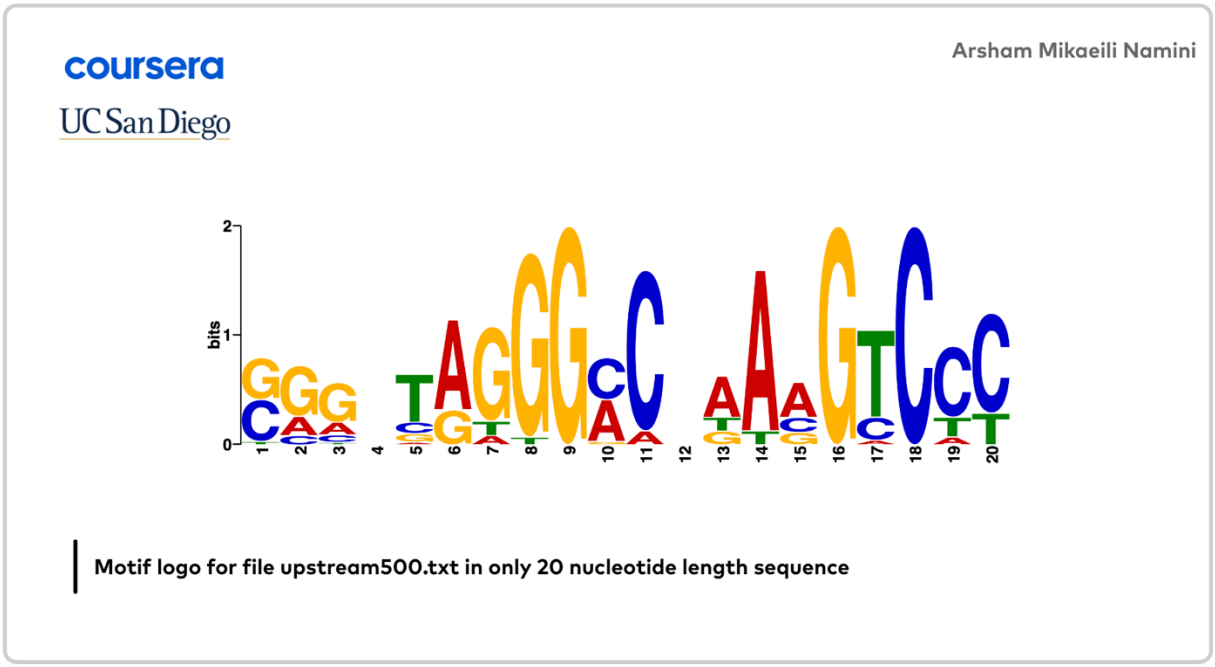
Arsham Mikaeili Namini



Motif logo for file upstream100.txt in only 20 nucleotide length sequence

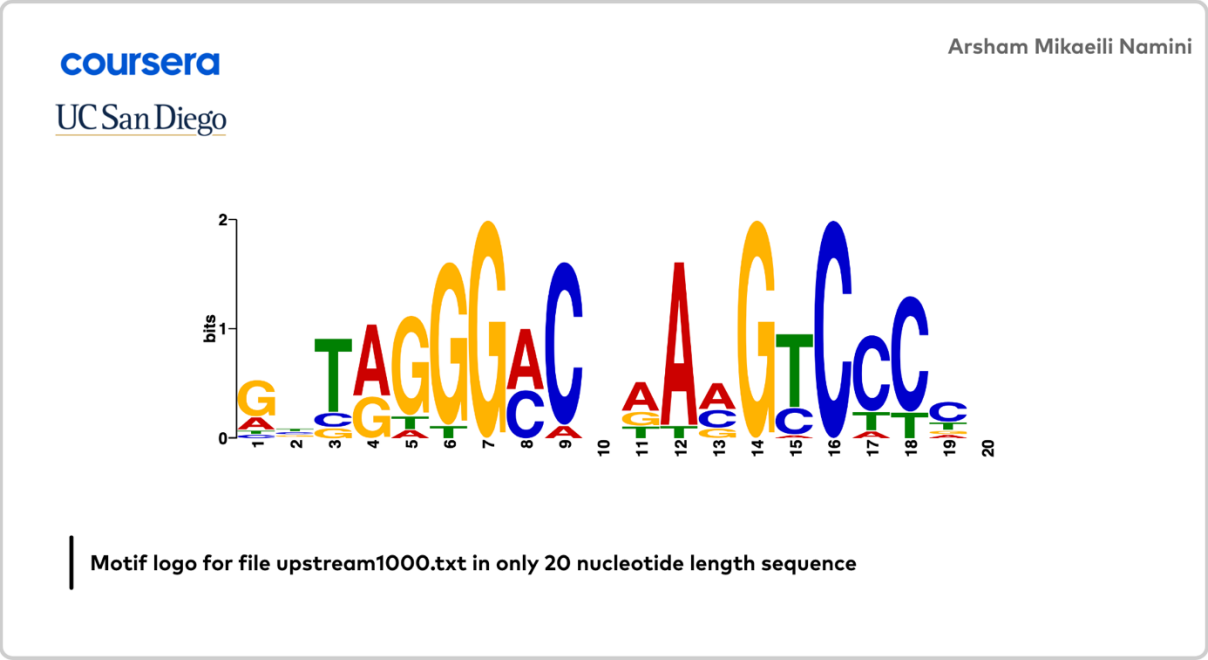
Next, we will consider upstream regions of length 500 bp (upstream500.txt).

Upload the motif logo obtained by running MEME on upstream500.txt. (Remember to specify in the options that we are looking for a single motif of length 20.)



Finally, we will consider upstream regions of length 1000 bp (upstream1000.txt).

Upload the motif logo obtained by running MEME on upstream1000.txt. (Remember to specify in the options that we are looking for a single motif of length 20.)



9. We will now compare the different motif logos generated from varying the length of upstream regions.

Which of the motif logos that you created are similar to the motif logo generated from upstream250.txt?

The motifs produced by upstream100, upstream500, and upstream1000 are all similar to the motif produced by upstream250, but the motifs produced by upstream25 does not resemble the others. (1 point for including each of upstream100, upstream500, and upstream1000; 1 point for not including upstream25).