**Hypoxia Motif Search in *Mycobacterium tuberculosis***

We will first work with [MEME](http://meme-suite.org/tools/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. After submitting the process, click on "MEME html output". Notice that the motif logo has been generated under "Discovered Motifs". Click the down arrow under "more" to see the starting positions of motifs. To download the motif logo, click the right arrow above the logo, navigate to the "Download-logo" tab, and click "Download".

If the queue on the MEME server is too long, you can use [alternate instance](http://alternate.meme-suite.org/tools/meme).

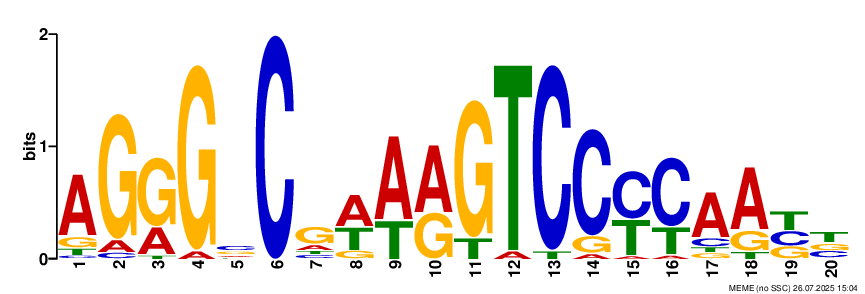
**Indicate the starting positions of the substrings of length 20 identified by MEME.**

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

In order to visualize the information contained in these sequences, we will copy them into WebLogo (http://weblogo.threeplusone.com)o generate a motif logo.

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

Motif Logo for Hypoxia-Related Binding Motifs in Mycobacterium tuberculosis



A motif diagram of one motif in Mycobacterium tuberculosis. The consensus sequence, indicated by the largest/top letter in each position, is AGGGCCGAAAGTCCCCAATT.

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](http://meme-suite.org/tools/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?**

The motif is 20 bases long. The sequence is AAGGGCCGAAAGTCCCCAAT, which only differs from the first sequence by an additional A at the beginning and a missing T at the end.

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

It is better to start with a smaller motif and work your way up from there if possible, because there will be more matches for a short sequence. As you increase motif length, matches become less and less likely, which can narrow down your search or eliminate biologically relevant options.

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).

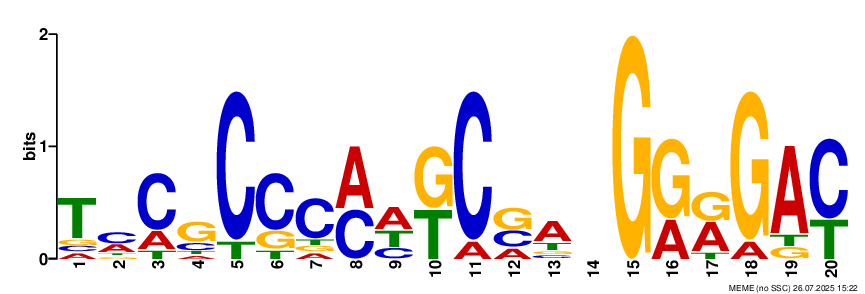
**How would you generate these strings? Justify your answer.**

To determine statistical significance, we would need to generate many random motif sequences with the same length. After generating scores for each sequence, we can see how our motif compares to the rest. If the motif of interest scores sufficiently higher than the rest of batch, we conclude that the score is not as high as it is just due to random chance and the motif is not a random sequence.

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](http://bioinformaticsalgorithms.com/software_challenges/motifs/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.)**

Motif Logo Upstream 25

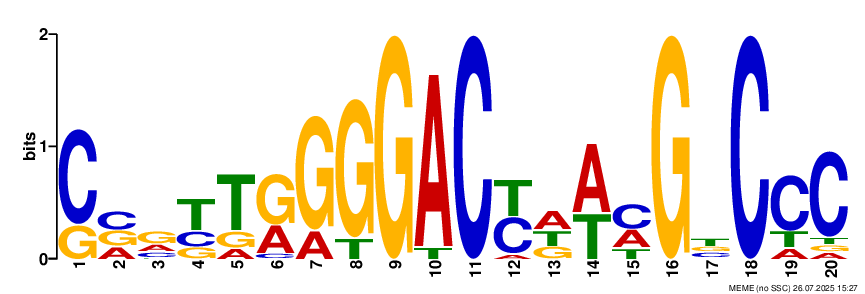


The sequence is TCCGCCCAAGCGACGGGGAC. (The base in position 14 does not have a majority.)

Next, we will consider upstream regions of length 100 bp ([upstream100.txt](http://bioinformaticsalgorithms.com/software_challenges/motifs/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.)**

Motif Logo Upstream 100

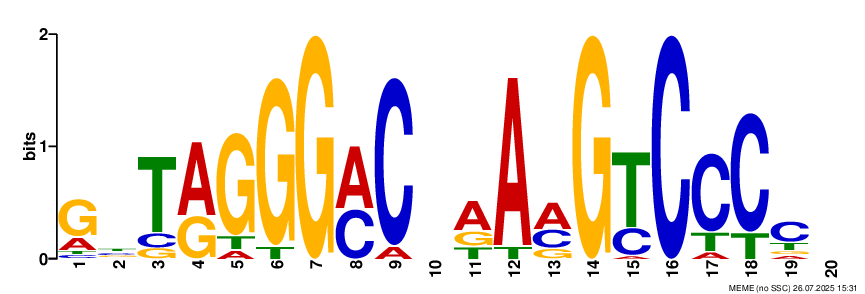


The sequence is CCGTTGGGGACTAACGTCCC.

Finally, we will consider upstream regions of length 1000 bp ([upstream1000.txt)](http://bioinformaticsalgorithms.com/software_challenges/motifs/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.)**

Motif Logo Upstream 1000



GTTAGGGACTAAAGTCCCCT.

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?**

Upstream 100, 500, and 1000 similarities with the motif generated from upstream250. The 4th-20th bases of the three sequences are very close to the whole 1-20 sequence from upstream 250, with a few insertions/deletions. Upstream 25 does not display much similarity, perhaps because the binding motif is more than 25 bases upstream.