De volta à semana 5 | Lições | Anterior | Início do curso

Avaliar colegas: Bioinformatics Application Challenge

Avaliar até Dezembro 14, 11:59 PM PST

Avaliações 5 a serem concluídas

Week 5 - Attempt 1

por Annie Lennek
Enviado em 4 de Dezembro de 2016

♡ curtida ⚐ Sinalizar este envio

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
174
120
161
198
214
157
159

The respective starting positions are:

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

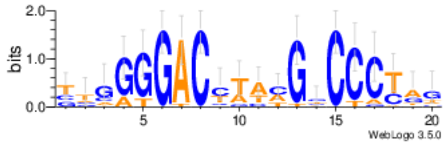
○ 0 pts
The learner provides more than three missing starting positions.

○ 1 pt
The learner provides at most three missing starting positions.

● 2 pts
The learner provides correct answer.

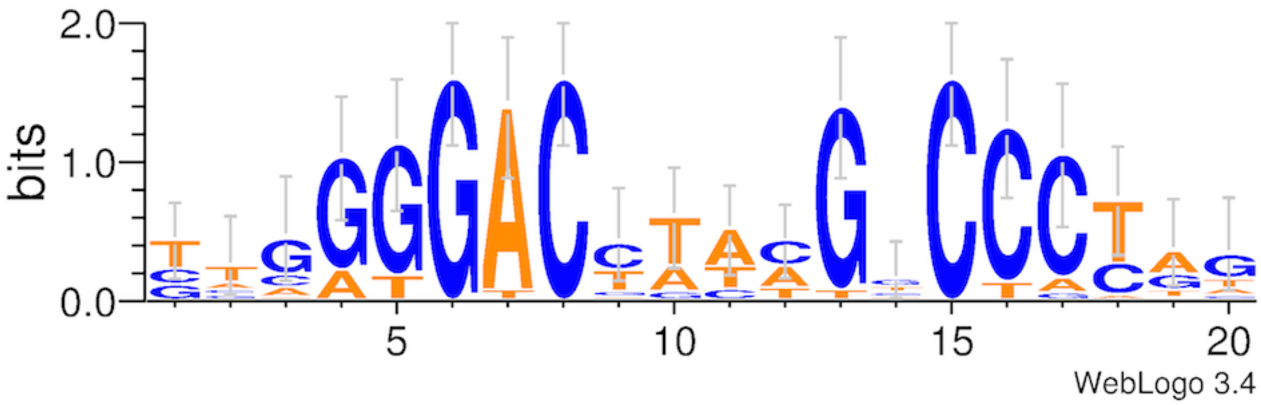
In order to visualize the information contained in these sequences, we will copy them into WebLogo (http://weblogo.threeplusone.com) to generate a motif logo.

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



20-mer motif logo created from upstream250.txt by Consensus: ttggggacctacgnccctag

The learner should provide motif logo shown below. (Note: the learner may have a differently colored motif logo.)



● 1 pt
Yes

○ 0 pts
No

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.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.

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

57
139
172
114
107
136
143
155
159
137
200
118
186
178
173
160
62
201
216
165
45
154
204
187

The starting positions are:

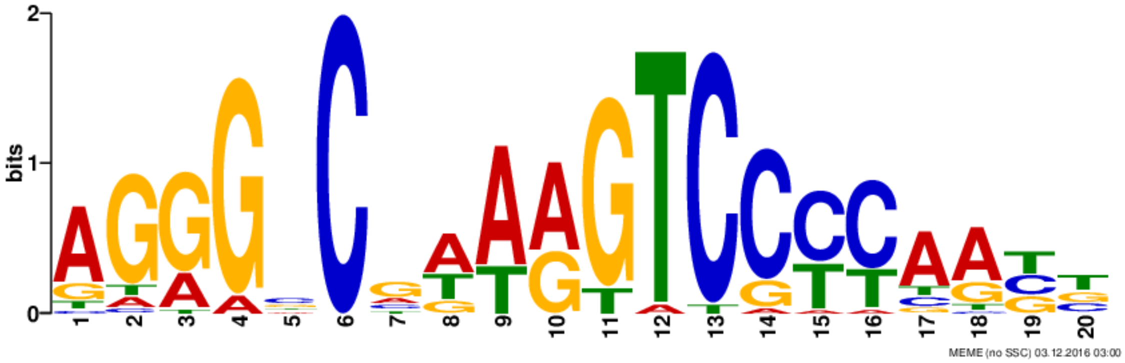
57 139 172 114 107 136 143 155 159 137 200 118 186 178 173 160 62 201 216 165 45 154 204 187

○ 0 pts
The learner provides more than three missing starting positions.

○ 1 pt
The learner provides at most three missing starting positions.

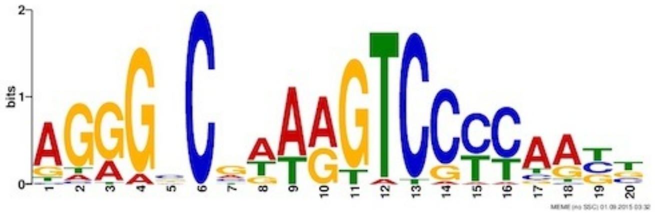
● 2 pts
The learner provides the correct answer.

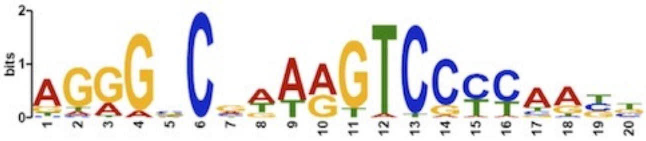
Upload the image file obtained after downloading this motif logo.



20-mer motif logo created from upstream 250.txt by MEME: agggccgaaagtccccaatt

The learner should provide one of the motif logos shown below.





● 1 pt
Yes

○ 0 pts
No

We will now run a third motif finder, W-AlignACE (Chen and Jiang, 2006). Choose upstream250.txt as your file to upload, enter 20 as the number of columns to align and the number of sites to expect, and change the fractional background GC-content to 0.66 (the GC-content of *Mycobacterium tuberculosis*). Then click "submit".

We will not ask you to output the results of your program. Why do you think that this is? (hint: W-AlignACE is based on Gibbs sampling). Make sure to answer in your own words.

Since the initial motif is chosen at random, its influence will likely be felt in the final motif output, especially if only a few number of iterations are performed.

W-AlignACE is based on Gibbs sampling, which we saw in the main text. Gibbs sampling is a randomized algorithm (i.e., based on tossing coins and rolling dice to find motifs). As a result, as stated on the W-AlignACE homepage, it is not possible to guarantee that the algorithm will return the same result every time.

○ 0 pts
The learner gives no answer or provides copied text from the W-AlignACE homepage.

○ 1 pt
The learner attempts an explanation but does not mention randomized algorithms or returning the same result.

○ 2 pts
The learner explains that W-AlignACE will not return the same answer each time but fails to connect it to randomized algorithms.

● 3 pts
The learner explains that Gibbs sampler is based on random chance, meaning that the same result may not be returned every time.

Did all three programs generate similar motifs? Provide a brief (1-2 sentence) explanation.

The consensus and the W-AlignACE motif are fairly similar, while the MEME motif seems rather different. One explanation is that the motif we are searching for is longer than 20 nucleotides. Alternatively, there could be multiple motifs of length 20 in upstream250.txt.

(A good answer for "No") No, the three programs generated mostly dissimilar motifs, with only vague similarities between the motifs generated by Consensus and MEME.

(A good answer for "Yes) Yes, the motifs do seem to be similar if we don't overlap them exactly but instead slide them over by 1-2 nucleotides in order to align similarities.

○ 0 pts
The learner provides a one-word answer.

● 2 pts
The learner attempts a well-reasoned explanation but does not provide a complete idea.

○ 4 pts
The learner answers "No" and provides a well-reasoned explanation.

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 40 nucleotides long. (b) The 20-mer motif logo from MEME is quite similar to the first half of the 40-mer motif logo.

(a) 2 points: The motif has length 40.

(b) 2 points (1 point for identifying similarity, 1 point for a reasoned explanation). This motif is much longer then previous one found by MEME. But the first half is similar to the previous one, and it picks the same conserved positions (GGG(C/A)C and G(T/G)CCC), which forms palindromic region.

○ 1 pt
(See above grading scale.)

○ 2 pts
(See above grading scale.)

○ 3 pts
(See above grading scale.)

● 4 pts
(See above grading scale.)

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

I think it would be better to start with shorter motifs. As you lengthen the motifs, you should see the shorter motifs settle in place within the longer and improve the score.

It is better to start with shorter motifs. As we have seen, it takes less time to find shorter motifs, and some algorithms may even be unable of identifying longer motifs. Since longer motifs will contain shorter motifs as substrings, we may be able to find longer motifs by first finding shorter motifs and then attempting to expand them into longer motifs.

○ 0 pts
The learner offers no answer.

○ 1 pt
The learner says that we should start with longer motifs and offers minimal explanation.

○ 2 pts
The learner says that we should start with shorter motifs, but offers minimal explanation.

● 4 pts
The learner says that we should start with shorter motifs and offers a reasonable explanation.

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.

I would set up a weighted dice to generate each nucleotide in the string with the probability of rolling both G and C set to one half the GC content of the organism. Similarly the probability of rolling A and T would be 1/2*(1-GC content). For this example, G and T would be rolled at 0.33, and the probability of rolling A and T would be 0.17.

There are several possible answers. Here are three:

- 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.

○ 0 pts
The learner does not provide an answer.

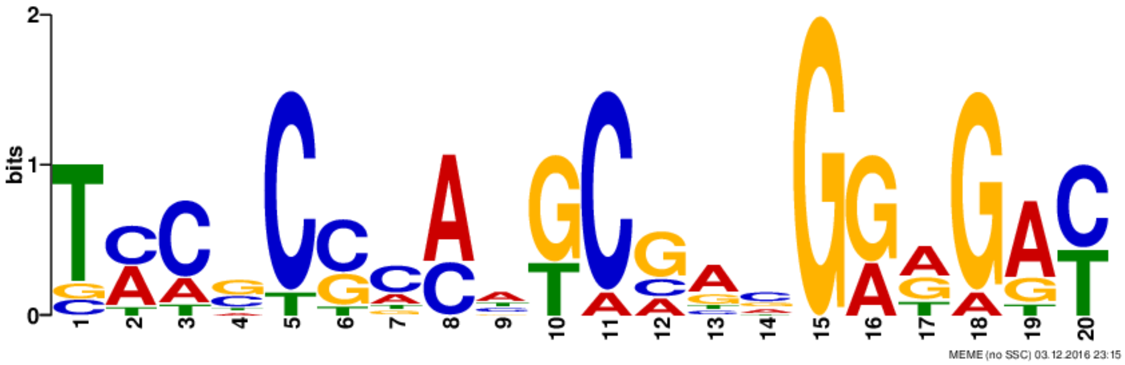
○ 1 pt
The learner gives a reasonable answer but does not justify it.

○ 2 pts
The learner gives a reasonable answer and provides only limited justification.

● 3 pts
The learner gives a reasonable answer (it does not have to be one of the above two) and justifies how it would help determine significance.

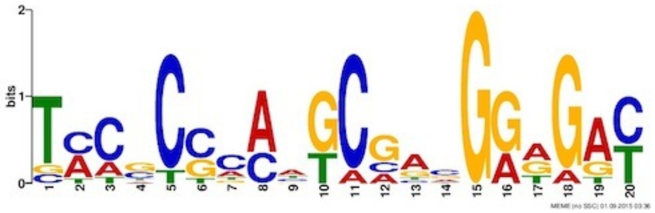
We have begun to confirm our colleague's suspicion that we should consider motifs of length about 20. However, thus far, we have only analyzed the 250 bp regions upstream of each gene. This makes us wonder whether we will identify the same motif for upstream regions of different lengths. First, we will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

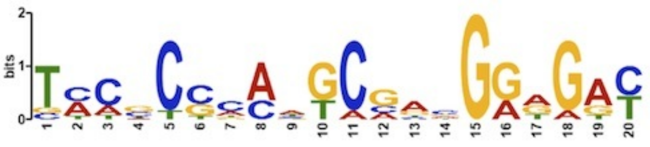
Upload the motif logo obtained by running MEME on upstream25.txt. (Remember to specify a motif of length 20 in the advanced options.)



20-mer motif logo created from upstream 25.txt by MEME: tccgcccaagcgacggagac

The motif logo is one of the ones shown below:





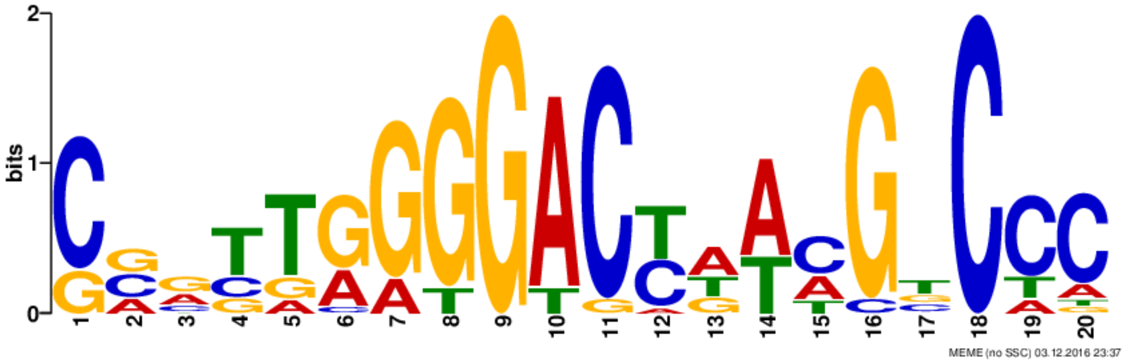
○ 0 pts
The learner does not report the correct motif logo.

○ 1 pt
The learner does not specify the correct logo but does report a similar logo.

● 3 pts
The learner reports the correct motif logo.

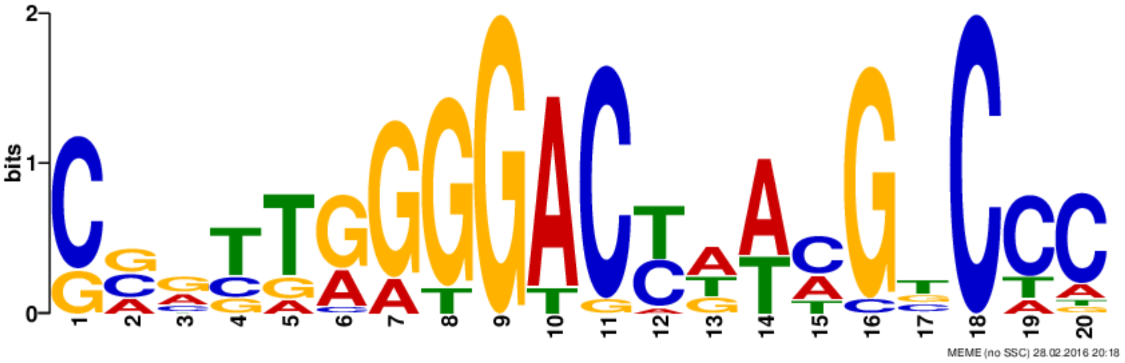
Next, we will consider upstream regions of length 100 bp (upstream100.txt).

Upload the motif logo obtained by running MEME on upstream100.txt. (Remember to specify a motif of length 20 in the advanced options.)



20-mer motif logo created from upstream100.txt by MEME: cggttggggactaacgtccc

The motif logo is shown below:



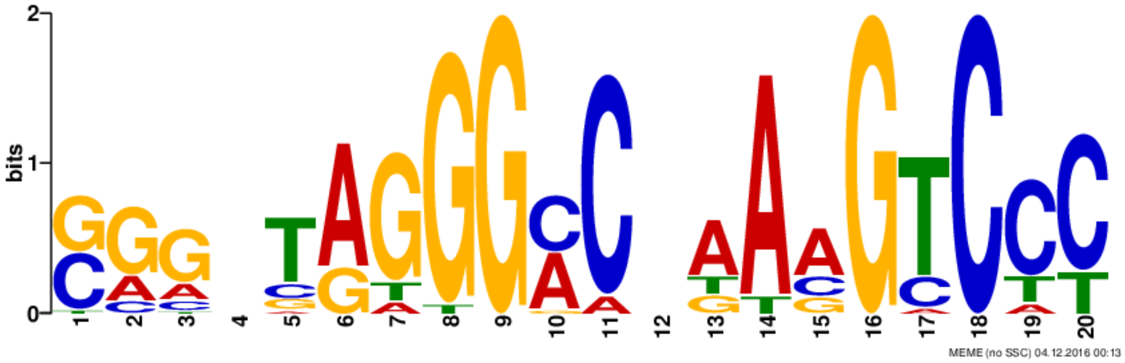
○ 0 pts
The learner does not report the correct motif logo.

○ 1 pt
The learner does not specify the correct logo but does report a similar logo.

● 3 pts
The learner reports the correct motif logo.

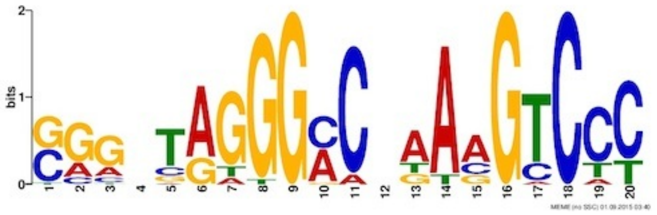
Next, we will consider upstream regions of length 500 bp (upstream500.txt).

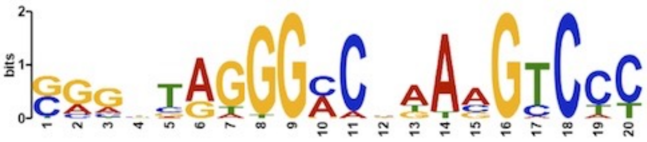
Upload the motif logo obtained by running MEME on upstream500.txt. (Remember to specify a motif of length 20 in the advanced options.)



20-mer motif logo created from upstream500.txt by MEME: gggntagggccnaaagtccc

The motif logo is one of the ones shown below:





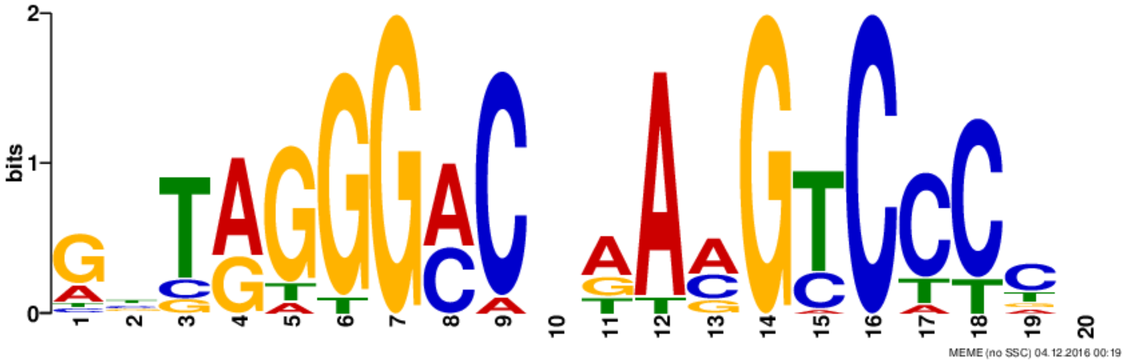
○ 0 pts
The learner does not report the correct motif logo.

○ 1 pt
The learner does not specify the correct logo but does report a similar logo.

● 3 pts
The learner reports the correct motif logo.

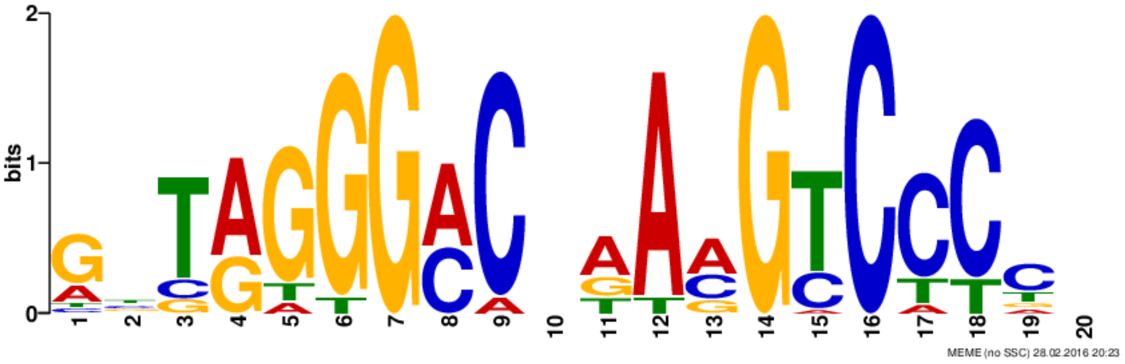
Finally, we will consider upstream regions of length 1000 bp (upstream1000.txt).

Upload the motif logo obtained by running MEME on upstream1000.txt. (Remember to specify a motif of length 20 in the advanced options.)



20-mer motif logo created from upstream1000.txt by MEME: gntagggacnaaagtccccn

The motif logo is shown below:



○ 0 pts
The learner does not report the correct motif logo.

○ 1 pt
The learner does not specify the correct logo but does report a similar logo.

● 3 pts
The learner reports the correct motif logo.

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 created from upstream100, upstream500 and upstream1000 are similar with perhaps some left or right shifting so they align. The motif created from upstream25 looks rather different.

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).

○ 0 pts
The learner did not correctly identify any of the datasets.

○ 1 pt
The learner correctly identified 1 of the datasets.

○ 2 pts
The learner correctly identified 2 of the datasets.

○ 3 pts
The learner correctly identified 3 of the datasets.

● 4 pts
The learner correctly identified 4 of the datasets.

(optional) Please provide any additional general feedback that you would like to give here.

Well done

Submit Review

You must fill in all the fields above.

Comentários

Visível para os colegas de turma

compartilhe suas ideias...