Exercise: Behavior and Limitations of Motif Finding

Jeremy Buhler

August 25, 2017

Today, we'll look at some behaviors of the popular MEME motif-finding software. MEME is a powerful tool for discovering putative regulatory motifs in DNA sequences. You can find it at `http://meme-suite.org/tools/meme`.

To begin, we'll focus on a motif-finding problem in *D. melanogaster* promoter sequences. The sequences we will work with are in the FASTA-formatted file `dmel-promoters.fna`. These sequences were extracted from just upstream of the transcription start sites of twelve genes, all of which are found on the dot chromosome *and* exhibit expression specific to fly heads. We hypothesize that these genes, which share both common expression and a common chromosomal environment, are regulated by one or more shared transcription factors.

> **Question 1**: *Can you think of an alternate mechanism that could explain why these genes are co-expressed but do not require them to share a binding site for a common transcription factor?*

To investigate this hypothesis, we'll use MEME to look for putative motifs shared among these sequences. In the submission form at the above URL, click "Browse" and tell MEME that you want to search `dmel-promoters.fna`. Finally, hit "Start Search" to launch the motif finder.

You might have to wait a few minutes for the search to finish, depending on the server load. The search page will refresh periodically. When the search is done, new links to appear on "Results" section of the page. Use the "MEME HTML output" link to view the search results.

1 Investigating a MEME Motif

MEME returns several putative motifs, each with a very significant E-value. These motifs vary between 37 and 50 bases in length. You can click on the down arrow under the "More" heading next to each motif to see a detailed alignment of its instances.

> **Question 2**: *How does the range of motif lengths reported by MEME compare to the lengths of binding sites typically reported in the literature? Why might such a large region exhibit similarity across multiple promoters?*

Unfortunately, we can't tell much about a putative motif's function just by looking at the raw MEME output. However, just as BLAST lets us compare new sequences to a database of well-annotated functional elements, MEME has a companion tool, Tomtom, that can compare motifs to known transcription factor binding sites.

Let's submit the first motif in the output to Tomtom. Click the right-facing arrow under "Submit/Download" next to the first motif. The "Tomtom" option should already be selected in the resulting dialog. Hit "Submit" to pull up the Tomtom submission page for your motif.

Just like BLAST, Tomtom offers a choice of organism-specific databases. Under "Select target motifs," specify "FLY (Drosophila melanogaster) DNA" in the first drop-down menu. The second menu now offers several motif databases to search. Select the "Fly Factor Survey" database, and hit "Start Search" to continue.

Question 3: *How were the motifs in the Fly Factor Survey database obtained? Are they supported by experiments? If so, what kind? (Hint: a Google search for "fly factor survey database" will be informative.)*

For each motif from Fly Factor Survey that matches part of our putative motif, Tomtom gives an estimate (in particular, an E-value) of how significant the match is, and a link to more information about the gene whose protein product is associated with the motif. Among other things, these links provide information about the functional roles associated with the protein.

Question 4: *How much of your putative motif is matched by known motifs from the Fly Factor Survey? Based on the E-value statistics, how strong are the matches?*

Question 5: *For the* first *motif match, is there functional evidence from the linked gene description that suggests that this gene could plausibly bind the specified site?*

2 Playing with MEME's Parameters

The quality and quantity of MEME's output is quite sensitive to its search parameters. Let's try a couple of variants of our original search to see how these parameters affect the results we obtain.

Repeat the MEME search, but this time, open the "Advanced options" tab and limit the length of the motifs returned to at most 20, rather than the default of 50.

Question 6: *Do the motifs returned by this search appear at the same locations as those in the previous search? How do their E-values compare to those of the motifs found previously?*

Question 7: *Why, given a choice of lengths, did MEME return such long motifs before? Why might it be advantageous to let MEME find very long motifs? Why might it be disadvantageous?*

Another important parameter for MEME is how many times a motif is permitted to occur in a sequence. By default, MEME assumes that a motif occurs either once or not at all in each input sequence. However, we can instead tell MEME that a motif could occur *any* number of times in each sequence.

Question 8: *Is there a biological reason to think that a given motif might occur multiple times in one promoter?*

To see if a motif might occur multiple times, look at the "Motif Locations" tab in the MEME results. By default, it shows the (at most one) instance in each sequence that contributed to MEME's reported motif model. If we instead select the "Motif Sites+Scanned Sites" option, it will also show (in lighter colors) other parts of the sequences that might match one of the motifs but were not used to build the motif model. Try this now.

Question 9: *What does the result suggest about how often these motifs occur in the input sequences?*

If a motif occurs in multiple instances per sequence, it might be a good idea to consider *all* these instances when determining the motif's specificity in the first place. To do this, rerun the MEME search, still limiting the maximum motif length to 20, but this time, under "Select the site distribution," choose "Any number of repetitions."

Question 10: *How do the E-values of the resulting motifs compare to those observed previously? How specific are the resulting motifs compared to those obtained previously?*

Question 11: *To what extent do these motifs overlap the ones you obtained previously? Do they hit the same known motifs in the Fly Factor Survey Database?*

Question 12: *What are some advantages and disadvantages of having MEME build its motif models from more than one instance per sequence?*

3 Breaking MEME

For our last exercise, we'll apply MEME to what should be quite an easy problem. The file `planted-100.fna` contains 20 sequences, each of length 100 bases (*much* shorter than the previous examples). These sequences were generated uniformly at random with equal base frequencies. Into each sequence, we planted an instance of a single motif of length 15. The motif has consensus sequence `TTTTGTCGCATTCCG`, but each instance differs from this consensus by four substitutions at randomly chosen positions. All instances were planted on the forward strand.

This is a very strong motif, in the sense that a set of unrelated 15-mer sequences with this degree of similarity would be highly unlikely to occur by chance in the background sequence; MEME would therefore give it a significant E-value. You can see the actual starting coordinate of each instance in its sequence in the sequences' FASTA headers.

Ask MEME to find the planted motif in this file. To make MEME's job easier, change the search parameters to tell it that there is *exactly* one motif instance per sequence, that the minimum and maximum lengths are both 15, and that it should search only the forward strand.

> **Question 13**: *How well does the motif found by MEME match the known consensus? How many of the 20 instances were found (almost) correctly?*

> **Question 14**: *How does this motif differ qualitatively from the ones found by MEME previously? In particular, look at the heights of letters in the sequence logo. Why might this motif be harder to find than the ones we found previously?*

The file `planted-600.fna` contains another planted motif of length 15, this time with consensus `TCACTATATCAGTCT`, again with four substitutions in each instance. This time, however, the 20 random background sequences are each 600 bases long. Despite the somewhat longer background length, the planted motif remains highly significant. Try to find this motif with MEME, using the same parameters as before.

> **Question 15**: *How well does the motif found by MEME match the known consensus? How many of the 20 instances were found (almost) correctly?*

> **Question 16**: *Does the fact that a motif is strong guarantee that it will be found by MEME? Do these planted motifs seem biologically plausible?*