**EXERCISE 3**:

The motif that I found this week has a consensus site of GCCACTTGA versus SCACTTGAV of last week. I found two other motifs. The second motif was MCGGGC. The third motif was GCGAGYCCGG. I do not think either of these two motifs is relevant. They are artifacts. The e-value for second motif is 10-fold greater than the first motif. The third motif has an e-value in the same order of magnitude as the first sequence but less information content and a greater length. This is just based off of general speculation from skimming the given information. In order to further investigate this, I would compare all of these motifs to the same reference databases, and determine which is homologous to the database.

JASPAR\_CORE\_2014.meme

I did find it useful. You can observe the query motif against the JASPAR CORE 2014 database. The homology of the query motif can be observed amongst different multicellular eukaryotes.

When you change the database to flyreg.v2.meme – an obvious change occurs. The number of motifs matched decreases from 32/593 (matched motifs/total motifs in database) to 15/75. This is most likely because the flyreg.v2.meme database is more specific towards *Drosophila* than JASPAR CORE 2014 which has several multicellular eukaryotes. The e-values for ‘tin’ are noticeably lower (by 100 fold order of magnitude) in the flyreg.v2.meme database as well, which is a better indicator of homology.

**EXERCISE 4**:

Exercise 4: Now look at the sequence-specific information for each motif. Did you find all your Tin sites? How does the output compare with the searches you did in Exercise 1? The FASTA headers for the sequences you used contain PMIDs for papers with verified Tin sites (some are the same as the ones you used earlier in Ex 1 and Ex 2). Look up the verified sites. How well did you do in finding known sites?

No I did not find all of the Tin sites (if the Tin sites are considered the “confirmed binding sites” from exercise 3 of last week). The consensus sequence for the best motif discovered is GCCACTTGA. The output is similar to the searches I did in exercise 1.

Exercise 1 output: for sequence C, D, and E

Abbass-MacBook-Pro:week\_9 aarizvi$ python RIZVI.seqsearch.py sequence\_C SCACTTGAV

position is: 653

motif is: CCACTTGAG

Abbass-MacBook-Pro:week\_9 aarizvi$ python RIZVI.seqsearch.py sequence\_D SCACTTGAV

total number of matches is: 0

Abbass-MacBook-Pro:week\_9 aarizvi$ python RIZVI.seqsearch.py sequence\_E SCACTTGAV

position is: 942

motif is: GCACTTGAG

position is: 953

motif is: CCACTTGAC

position is: 1710

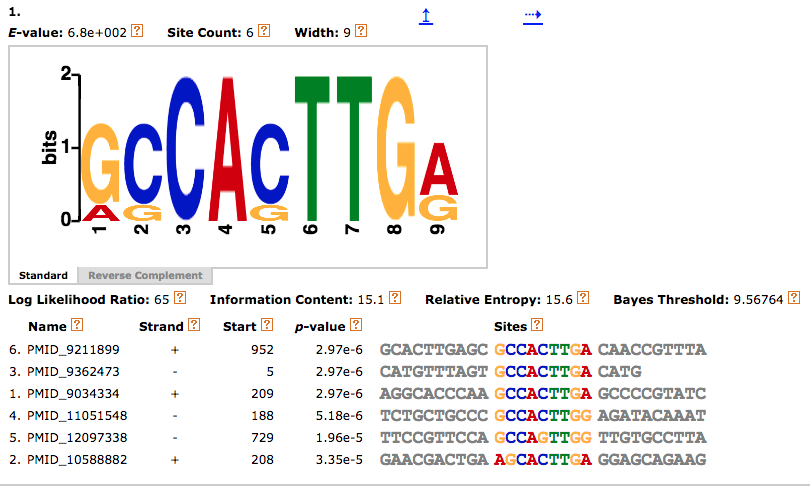
motif is: TTCAAGTGC (reverse strand)

position is: 1759

motif is: CTCAAGTGG (reverse strand)

total number of matches is: 4

Output from MEME:



The output is quite similar. Sequence C motif when searched for SCACTTGAV is quite similar to the motif CCACTTGAG (position: 653) is quite similar to motif 1 (PMID\_9034334) site. Sequence D has no matches. Sequence E had 4 matches – the GCACTTGAG (position: 942) sequence is similar to motif 2 (PMID\_10588882) and CCACTTGAC (position: 953) is similar to motif 6 and 3 (PMID\_9211899 and PMID\_9362473, respectively). Interestingly, all 3 forward strands had similar matches, while the reverse strands did not – I believe that this is most likely due to the reverse nucleotide dictionary that was used in the script.

I looked up the verified PMIDs (they are listed above). I think that my program did a fairly good job at looking up the sites. CACTTGA is found in most of the papers. This is the same as one of the sites that was found in my 7-mer word count program. TCAAGTG is my other most common 7-mer (the reverse complement of the above other 7-mer). I found this in one paper (PMID\_9034334). So because of this, I would say that my program did a fairly good job at finding known sites.