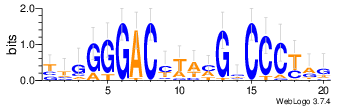
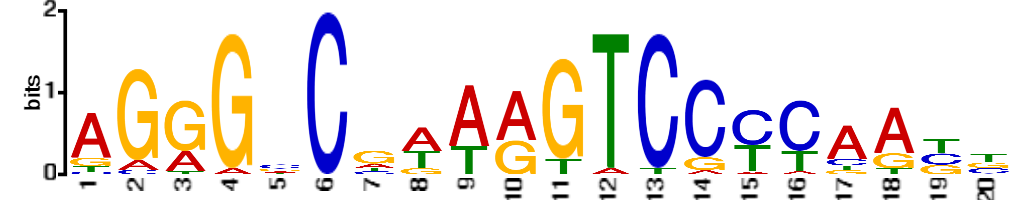
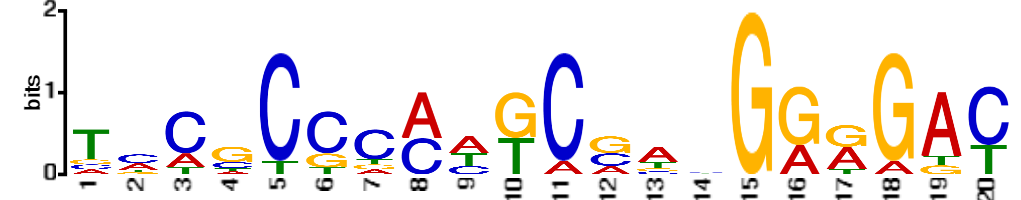
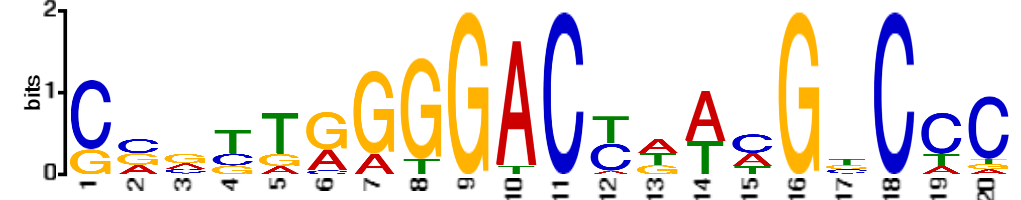
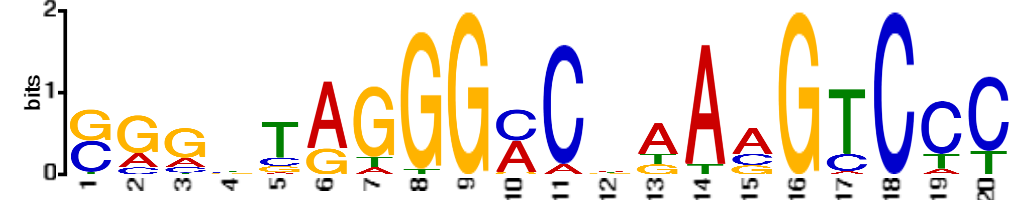
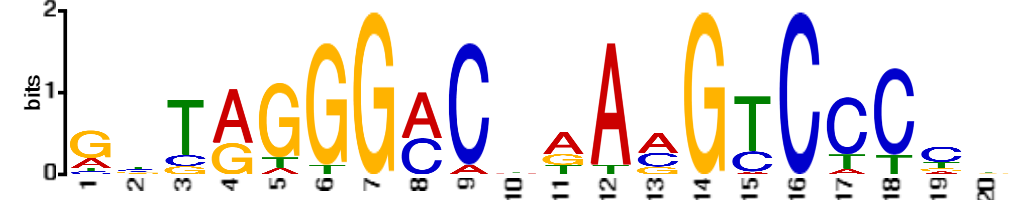
**Software Challenge**

**Sequence Motif Finding**

1. 187, 141, 117, 162, 156, 199, 139, 202, 175, 161, 180, 169, 174, 120, 161, 198, 214, 157, 159
2. 
3. 57, 139, 107, 172, 114, 136, 159, 143, 155, 186, 178, 200, 188, 137, 173, 201, 160, 62, 216, 165, 45, 204
4. ****
5. The motifs generated by the 2 programs differ but there are some areas of similarity (as though they are roughly shifted 2 places relative to each other). Namely, in both cases in the first half of the strings, there is a preponderance of G’s in 3 consecutive positions as well as a large amount of C’s 2 positions removed downstream from that. In the second half of the motifs, it seems that both set of motifs contain positions where C is highly conserved in 3 consecutive places (and in a 4th, a somewhat even C/T split) and both motifs conserve G upstream 2 positions of the beginning of the C’s.
6. It is 20 BP in length.
7. It is very similar, yes. Many of the same columns are conserved.
8. It is better to start with short motifs because short motifs are substrings (sub-motifs) of larger motifs, so the short motifs (which is faster relatively) can be used to “warm-start” the search for larger ones.
9. (I am assuming that the preface to the question is referring to a “better” score rather than a “larger” one). I would generate these “typical” DNA strings by randomly assigning nucleotides to each position according to a uniform profile (or a profile based on an estimate the actual distribution of bases in DNA, say of the species of interest). These collections of strings naively represent the “typical” DNA strand since they are roughly derived from the population-level distributions of nucleotides in DNA and, by comparing the chosen motifs to these random (representative) strings, we can determine whether the motif found is “real” or maybe just an “accident”.
10. ****
11. ****
12. ****
13. ****
14. The motifs logos most similar are those from the 100, 500, and 1000 BP data. Evidently, the 25 BP file is too short to compass the whole motif. The others are very similar to the 250 BP data set, encompassing all the important regions—just shifted slightly.