# Homework 4

**Instructions.** As usual, you may work together on homework, but you must write up the solutions on your own. If you turn in an electronic copy, please provide it in pdf format. This homework is due by the beginning of class on Thursday, March 2nd.

1. In this first question you will assess whether a couple of motifs are overrepresented in a set of sequences using Markov chain *theory* (not Monte Carlo simulation). I pose three questions, but you may not be able to answer them in order, so gather your evidence and provide one answer. Consider answers that allow overlaps throughout this homework.

   (a) What order Markov chain should be used to model the sequences? Does it matter? When you settle on an order, use it for the rest of this homework. (Hint: I have posted an old perl script `torder` that I have shared with previous versions of this class that does nested tests of Markov chain order. It runs as `./torder < data_file` and should give you what need without further modification.)

   (b) Is there an overrepresentation of motif GGTCAAAGGT in the sequences?

   (c) Is there an overrepresentation of motif GGCGCC in the sequences?

2. One difficulty with motifs is that they are not exact: nature can often tolerate a great deal of mismatch. How can you modify the theory to handle motifs GGTNAAAGGT and GG[CG][CG]CC, where N is any nucleotide and [CG] matches C or G? For this question, write down modified versions of the equations $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$ derived in the Markov chain notes as they apply to these ambiguous motifs, but *do not* implement the tests.

3. Another difficulty with motifs is that they often function in the DNA duplex and may function on either strand. How can you modify the theory to handle a match on either the forward or reverse strand for the original motifs GGTCAAAGGT and GGCGCC? Again, just write down modified versions of the equations $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$, but *do not* implement the tests.

4. In order to address the next question, we will need the variance $\mathrm{Var}(W)$ of $W$, the distance between occurrences. Use first step analysis to derive a general system of equations that can be solved for the variance.

5. Your client is interested in the spacing of the two motifs. There are two different random variables, $W_1$ the wait time until the first occurrence of GG[CG][CG]CC after GGTNAAAGGT, and $W_2$, the wait time until the first occurrence of GGTNAAAGGT after GG[CG][CG]CC. One of them may be much more common than the other, in which case you need only do the analysis for that one. Is the spacing of these two motifs when they occur together unusual? (For this problem you need to consider the ambiguity in the motifs–otherwise there is no data to analyze. Carefully set up the transition probability matrix before attempting to write the equations you will need to solve.)

6. Do you have any concerns with the applicability of the theory we have used in this homework? Why?

7. The ambiguity in the motifs is far more than I allowed above. I kept it “simple” to prevent the state space from blowing up too big on you. Possible motifs are often “scored” by how well they match a position scoring matrix (PSM) obtained from biologically confirmed instances of the motif. If we assume the sites in the motif are independent but not identically distributed, then the PSM provides the nucleotide probabilities $p_{iN}$ for the probability of base $N$ in the motif at position $i$. Then, the likelihood of potential motif sequence $M_1M_2\cdots M_k$ is $p_{1M_1}p_{2M_2}\cdots p_{kM_k}$. The higher the likelihood, the better the potential motif sequence matches the binding site motif. Map out a process to use this score in the context of a Markov chain of the order chosen in question 1 to identify the occurrences of these two motifs and assess whether the spacing of these two motifs is unusual in these data. (**Bonus:** You are not asked to implement this method. If you do, it will earn you bonus points.)