Following decades-long work of many, we will study sequence motif analysis.

1. **Modeling motifs**. Read Stormo [2]'s brief history of computational motif analysis.
   (a) On canvas, you will find a fasta file containing thousands of known 9bp motifs of a transcription factor. Estimate the parameters of an independent sites model for the motif.
   (b) Read the Markov chain notes through the section on Likelihood. Does a *nonhomogeneous* Markov chain fit the motif better than the independence model of Part a? Compute a p-value using:
      i. theory and,
      ii. a Monte Carlo method.
   (c) The Markov chain allows for dependence between random variables, but it is of a very specific form and there are many other kinds of dependence. Can you find evidence of other kinds of dependence beyond what you might have already discovered in Part b? Provide numeric measures of confidence for your conclusions.

2. **Testing for presence a known motif.** In this question, you will examine 2,135 promoter regions (1000 bp) upstream of genes, known to be transcribed under a particular biological condition of interest, for evidence of the motif you studied in Question 1. For this question, I want you to have the background knowledge of Sections 2.5, 2.7–2.10 from the Markov chain notes. In particular, from the perspective of modeling, pay attention to Sections 2.5 and 2.10. The other sections are theoretical background supporting complete understanding of the models.
   (a) The Neyman-Pearson Lemma states that the most powerful statistic for detecting the signal of a simple hypothesis H_1 against another simple hypothesis H_0 is the likelihood ratio. (A simple hypothesis is one without any unknown parameters.) Formulate a likelihood ratio that would most powerfully detect the presence of the motif you modeled in Question 1 in its natural habitat, *i.e.* the mouse genome.
   (b) Unfortunately, the Neyman-Pearson Lemma does not also provide a sampling distribution for the most powerful statistic it suggests. Propose and implement, on the data provided, a Monte Carlo method to detect which promoters likely have at least one copy of the motif. In your writeup, justify your choice of Monte Carlo method and why you did not choose other available Monte Carlo approaches.
   (c) Explore a little, using code and data, and discuss further the impact of your choice for H_0. Does it matter which H_0 you use? If so, which H_0 should you use? If H_0 involves unknown parameters, what is the best way to obtain estimates for those parameters?

3. **Testing for motif enrichment.**

   Read McLeay and Bailey's overview [1] of motif enrichment analysis. How does motif enrichment analysis differ from what we did in Question 2? Describe how you could supplement the data of Question 2 to implement the methods advocated there.

4. **Detecting novel motifs.**

References

[1] Robert C. McLeay and Timothy L. Bailey. "Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data". In: *BMC Bioinformatics* 11.1 (2010), p. 165. ISSN: 1471-2105. URL: https://doi.org/10.1186/1471-2105-11-165.

[2] Gary D. Stormo. "DNA binding sites: representation and discovery". In: *Bioinformatics* 16.1 (2000), pp. 16–23. ISSN: 1367-4803. URL: https://dx.doi.org/10.1093/bioinformatics/16.1.16.