You may work in groups, but the final writeup must be done by the individual. If you share code, have one team member turn in the code. Other team members should indicate which code they used and how they used it. If you change someone else's code, turn in the version you used, but properly credit the code.

This homework deals with codons, open reading frames, and coding sequences (CDS).

Table 1: Gene code ($*$ indicates a STOP codon, ATG is a START codon)

TTT-F	TCT-S	TAT-Y	TGT-C
TTC-F	TCC-S	TAC-Y	TGC-C
TTA-L	TCA-S	TAA-*	TGA-*
TTG-L	TCG-S	TAG-*	TGG-W
CTT-L	CCT-P	CAT-H	CGT-R
CTC-L	CCC-P	CAC-H	CGC-R
CTA-L	CCA-P	CAA-Q	CGA-R
CTG-L	CCG-P	CAG-Q	CGG-R
ATT-I	ACT-T	AAT-N	AGT-S
ATC-I	ACC-T	AAC-N	AGC-S
ATA-I	ACA-T	AAA-K	AGA-R
ATG-M	ACG-T	AAG-K	AGG-R
GTT-V	GCT-A	GAT-D	GGT-G
GTC-V	GCC-A	GAC-D	GGC-G
GTA-V	GCA-A	GAA-E	GGA-G
GTG-V	GCG-A	GAG-E	GGG-G

1 Naive codon model

The simplest possible model for the probability of observing a specific codon in a coding sequence is $P(C = c) = 1/64$ for $c \in \Omega_{\text{codon}}$, the set of 64 codons. List all the assumptions one must make in order for this model to work (there are at least 3).

2 A less naive codon model

a) Obviously, amino acids are not equally probable in the coding sequences of real proteins. Suppose you know the relative abundance of amino acids, expressed as probabilities $P(A = a)$ for all $a \in \Omega_{aa}$ in the set of 20 amino acids. Propose another model for codon probabilities that satisfies these amino acid probabilities

b) Under what conditions does the model reduce to the naive model above?

3 Build a model for the length of a random ORF

a) Describe a model for the length of a random open reading frame (ORF). An ORF is defined as a sequence of codons beginning with a START codon (ATG) and ending with a STOP codon (TAA, TAG, or TGA). The smallest ORF will have a length of 1 (e.g. ATG-TAA). The model should account for nonuniform nucleotide proportions, but it should not consider amino acid proportions.

b) Choose an organism with sequenced genome and use it to set parameters in your model. Build a test for the hypothesis that the observed length of a coding sequence (CDS) satisfies (or not) your model from question a). Note that we are only considering mono-exonic genes, which have no introns and no splicing. This test could be used as a rudimentary classifier to separate random ORFs from true coding genes (where selection preserves long ORFs). Some early genomics projects used the arbitrary cutoff of 100 codons to distinguish between true genes and random ORFs. What cutoff would you recommend for your organism?

c) Different organisms vary enormously in GC content, from 20% in the malaria parasite *Plasmodium falciparum* to 80% in some bacteria. Even between fairly close clades there can be huge differences in GC. For example monocots and dicots differ by around 10%. Prepare plots showing how GC proportion affects the expected length of a random gene.

4 ORF length in a genome with isochores

In the last question, you built a model of random ORF length parameterized by GC content. However, biology is never so simple. The GC content of vertebrates has a banded composition [1]. The bands, known as isochores, have lengths of around 500-1000 kb and are highly-variable in GC content. The overall distribution of isochores is conserved across vertebrates and can be broken into 5 classes, summarized in the table below.

Table 2: Vertebrate isochore overall averages. Adapted from Table 1 of [1]. Standard deviation is provided in parentheses.

name	%GC	relative amount
L1	36.2(0.3)	21.8(2.5)
L2	39.1(0.6)	36.4(0.8)
H1	43.3(0.7)	26.7(3.8)
H2	48.3(0.5)	12.5(1.3)
H3	54.8(0.5)	2.9(0.8)

a) Extend your random ORF length model to account for isochores. List the assumptions you make. The mechanisms that partition genomes into isochores are a matter of intense debate, but you are not expected to dig into the literature. Formulate your model based only on the data provided here.

b) How would failure to account for isochores affect your inferences about ORFs?

References

[1] Costantini, M. *et al.* (2009) The evolution of isochore patterns in vertebrate genomes. *BMC genomics* 10, 146