BCB 568 Homework Assignment 1

Schuyler Smith

January 25, 2018

1.
- All nucleotides are independent.
- All nucleotides are randomly distributed.
- All nucleotides occur with equal probability.

2. **a.** If we can use $m : \Omega_{\text{codon}} \to \Omega_{aa}$ to represent the relation of a specific codon to its amino acid, then $\mathcal{C}_{aa}$ is the set of all codons that translate to amino acid aa (i.e. $\mathcal{C}_F = \{\text{TTT}, \text{TTC}\}$). For our model to work we need to make the same assumptions from problem-1, which allows us to assume that when an amino acid is observed then any of the codons comprising $\mathcal{C}_{aa}$ are equally likely to be what we are observing. That would mean the probability for any codon considering amino acid probabilities is

$$\begin{aligned} P(C = c) &= P(C = c \mid A = m(c))P(A = m(c)) \\ &= \frac{1}{|\mathcal{C}_{m(c)}|} \cdot P(A = m(c)) \\ &= \frac{P(A = m(c))}{|\mathcal{C}_{m(c)}|} \end{aligned}$$

If we sum the probabilities for all codons for any amino acid it takes us back to the original probability for amino acids, P(A = a):

$$\sum_{c \in \mathcal{C}_{aa}} P(C = c) = \sum_{c \in \mathcal{C}_{aa}} \frac{P(A = m(c))}{|\mathcal{C}_{m(c)}|} = \sum_{c \in \mathcal{C}_{aa}} \frac{P(A = a)}{|\mathcal{C}_{aa}|} = |\mathcal{C}_{aa}| \cdot \frac{P(A = a)}{|\mathcal{C}_{aa}|} = P(A = a),$$

b. If we again use the same assumptions from problem-1, we know that $P(A = a) = |\mathcal{C}_{aa}|/64$. So for any $c \in \Omega_{\text{codon}}$, it can be modeled as

$$P(C = c) = \frac{P(A = m(c))}{|\mathcal{C}_{m(c)}|} = \frac{|\mathcal{C}_{m(c)}|}{64} \cdot \frac{1}{|\mathcal{C}_{m(c)}|} = \frac{1}{64}$$

3. **a.** An ORF has been defined as a sequence of codons beginning with a START-codon and terminated by a STOP-codon. In this way, the shortest ORF, in theory, would be of length 1, for the methionine from the START-codon. Essentially we want to say that the length L of a randomly generated ORF is going to be dependant on the probability p of observing a STOP-codon where the only alternative is finding an amino acid codon. The p we will define so that it will be equal to the probability of seeing the nucleotides in their sequence to compose one of the three STOP-codons

$$p = P(TAA) + P(TAG) + P(TGA)$$

The probability of these codons is defined by the probability of the nucleotides $q_{nucleotide}$ such that

$$\begin{aligned} p &= q_T q_A q_A + q_T q_A q_G + q_T q_G q_A \\ &= 2 q_T q_A q_G + q_T {q_A}^2 \end{aligned}$$

If we define finding a STOP-codon as a success and finding an amino acid codon as a failure we can model p as a geomettric random variable thusly

$$P(L = l) = p(1-p)^{l-1}$$

b. *Loxodonta africana* (african elephants) have mean GC content of $q_{GC} = 0.477$. To sufficiently answer the question we need to make the assumption that $q_C = q_G = 0.477/2 = 0.2385$ and that $q_A = q_T = (1 - 0.477)/2 = 0.2615$. We can use these ratios for the equation we got for part (a) and have

$$\begin{aligned} p &= 2(0.2385)(0.2615)^2 + 0.2615^3 \\ &= 0.0505 \end{aligned}$$

So if we model L as we did before and looking for a cutoff l to where we want to call the cutoff for what is most likely a true gene we can write the cumulative density function as

$$P(L \leq l) = 1 - (1-p)^l,$$

If we set our false possitive level α and say that an ORF is not a true gene if $P(L \geq l) \leq \alpha$. At $\alpha = 0.05$, we see

$$P(L \geq l) = (1-p)^1 = 0.05. l = \frac{\ln 0.05}{\ln (1-p)} = \frac{\ln 0.05}{\ln 0.9495} = 57.81 \approx 58.$$

So using these estimates we would say that an ORF with $l \leq 58$ is not likely to be a true gene. The reasoning being that if it exists at a length longer than what is statistically likely by random chance, the reason must be for a biological purpose.

c. Though the assignment comments that GC content may range anywhere from 20-80% I chose to model GC content at 0.42-0.56 with incrememnts of 0.02. I thought this would be interesting as it demonstrates the difference even a slight deviation from equilibrium can make. It turned out that it was not as drastic as I thought, but still has some interesting points. The graph is in another file, taken from an R graph output.

4. **a.** Isochores, I, have a GC content of q^I_{GC} and q^I_A, q^I_G, q^I_C, q^I_T asthe proportions each nucleoptide in I. For an ORF o, $o \in I$ denotes that o is contained in isochore I. To resolve our model we need to make additional assumptions

- Within each isochore I, nucleotides are independently randomly distributed
- Genes and isochores are adjacent, no introns or "junk" DNA
- Each ORF is found entirely in a single isochore

A revised model for ORF length that accounts for isochores and using the geometric random variable model for STOP-codons we applied earlier here as $P(L = l|o \in I)$

$$\begin{aligned} P(L = l) = {} & P(L = l|o \in L_1)P(o \in L_1) + P(L = l|o \in L_2)P(o \in L_2) \\ & + P(L = l|o \in H_1)P(o \in H_1) + P(L = l|o \in H_2)P(o \in H_2) \\ & + P(L = l|o \in H_3)P(o \in H_3), \end{aligned}$$

b. From the table of isochores we were given, on average, isochores have abnormally low GC-content. Lower GC-content precludes STOP-codons from being found. Knowing that, without accounting for isochores, our previous estimates for L of ORFs would be too conservative and that they acceptable l for concluding a true gene would be longer than estimated.