Due February 9, 2022 at 11:59PM.

Problem 1

Sequencing errors

(15 points)

Let us define the following random variables:

- S_n - the *true unobserved* sequence at position n.
- O_n - the *observed* sequence at position n.
- E_n - an indicator variable. $E_n = 1$ if an error is observed at position n. $E_n = 0$ otherwise.

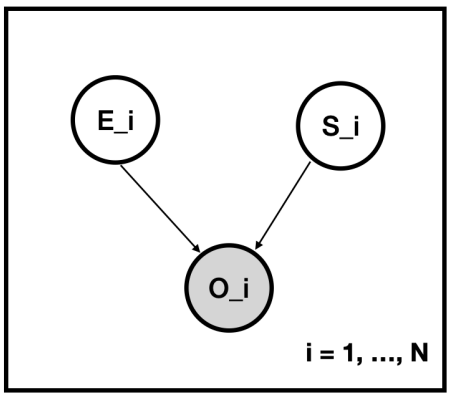
If no errors are realized, then the observation is the true sequence (e.g. $P(O_n = X \mid S_n = X, E_n = 0) = p_X$).

If an error is realized, the observed distribution looks as follows:

$$P(O_n = A \mid S_n = T, E_n = 1) = p_{AT}$$
$$P(O_n = T \mid S_n = A, E_n = 1) = p_{TA}$$
$$P(O_n = C \mid S_n = G, E_n = 1) = p_{CG}$$
$$P(O_n = G \mid S_n = C, E_n = 1) = p_{GC}.$$

The remainder of the possible errors have probability mass zero. If there isn't a probability explicitly provided above, simply treat it as a general random variable (e.g. $P(S_1 = A)$ or $P(E_1 = 1)$).

The model in plate notation looks as follows:



(a) If the observed sequence is AAA, what is the probability that the true sequence has exactly one T? Formally, one probability you need to keep track of is: $P(S_1 = A, S_2 = A, S_3 = T, E_1 = 0, E_2 = 0, E_3 = 1 \mid O_1 = A, O_2 = A, O_3 = A)$. Using Bayes' rule, you should be able to write this in terms of things you know.

(b) If the observed sequence is AAA, what is the probability that the true sequence is TTT?

(c) If the observed sequence is AGC, what is the probability the true sequence is ACC?

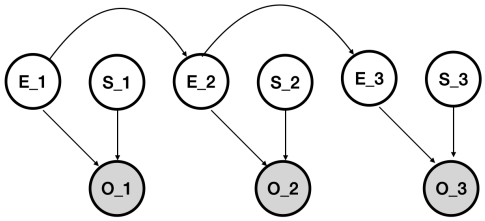
(d) If the observed sequence is AAA and you are told there is exactly one error, what are the possible sequences and what are their probabilities?

(e) If the observed sequence is AAA and you are told there is *at least* one error, what are the possible sequences and what are their probabilities

Now, let's make this model a bit more realistic. The probability of an error is no longer independent, but rather, it depends on the previous state. Here are the updated error probabilities (for $n > 1$):

$$\begin{aligned} P(E_1 = 1) &= p_E \\ P(E_n = 1 \mid E_{n-1} = 0) &= p_E \\ P(E_n = 1 \mid E_{n-1} = 1) &= 2p_E \\ P(E_n = 0 \mid E_{n-1} = 1) &= 1 - 2p_E \\ P(E_n = 0 \mid E_{n-1} = 0) &= 1 - p_E. \end{aligned}$$

All other distributions remain the same (with the appropriate conditioning). The model in plate notation for a sequence of length three looks as follows:



(f) If the observed sequence is AAA, what is the probability that it is correct?

(g) If the observed sequence is AAA and you know there are exactly two errors, what are the possible sequences and their probabilities?

(h) In general, for $n > 1$, what is the probability of an error at position n? To make things simpler, let us only worry about the error distribution and not the actual observation and sequence. If you previously did it like that, impressive and totally fine.

(i) In general, for $n > 1$, what is the probability that at least one occurs? Again, let us only worry about the error distribution and not the actual observation and sequence.

Problem 2

RNA-seq modeling

(15 points)

Please refer to the slides for definitions on effective length and its friends. Assume there is positional sequencing bias bias in your generative model. That is, the probability of starting a fragment any particular site is not uniform.

Let i be the index of the read we are about to observe. We define,

- O_i - the orientation of the fragment (i.e. strand). In real-life this can start at the 'left' or the 'right' and continue backwards. This is because DNA is a double helix and the gene has a reverse complement.
- S_i - an integer indicating the position of where the fragment begins.
- l_i - the length of the transcript generating the fragment.

We also define:

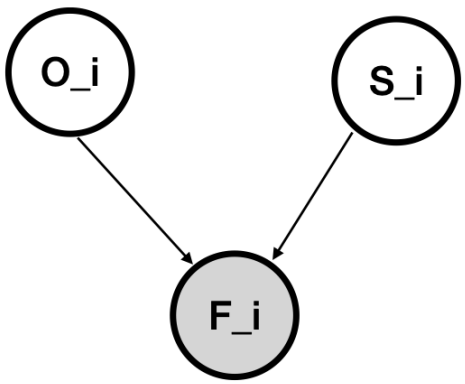
- $P(O_i = 0) = P(O_i = 1) = 0.5$
- $P(S_i \in (1, \frac{1}{2}l_i)) \propto \frac{1}{l_i}$
- $P(S_i \in (\frac{1}{2}l_i + 1, l_i)) \propto \frac{3}{l_i}$

The generative process is as follows:

1. draw an orientation, O_i.
2. draw a start position, S_i.
3. draw a fragment length, F_i.

For simplicity, you can assume all transcripts are of even length.

The generative model in plate notation looks as follows:



(a) For any transcript, what is the probability of a fragment starting in the first half of the transcript?

(b) For any transcript, what is the probability of a fragment starting in the second half of the transcript?

(c) What is the expected effective length for any distribution F_i? Be conscious of the indices.

(d) Now, assume F_i follows the following un-normalized probability mass function:

$$P(F_i = f) \propto \exp\left(-\frac{(f-\mu)^2}{2\sigma^2}\right).$$

While in (c) we had to condition on the start position, orientation, etc, as you saw, it can get pretty hairy. Instead, we are going to drop these assumptions and simply approximate the marginal (above) with that mass function. Thus, you can simply compute $E[F_i]$ in a straightforward manner as in the lecture. Again, if you did this the completely rigorous way, fantastic and you also get credit.

Compute the effective length for:

(a) $l_i = 1000, \mu = 200, \sigma = 20$,

(b) $l_i = 1000, \mu = 200, \sigma = 100$.

Please include the code.

(e) How did the values change and is this reasonable behavior?