Exercise Solutions  Jan. 23rd, 10 points for each question and totally 50 points.

1. In a DNA sequence set, there are totally $2 \times 10^4$ nucleotides, among which $6 \times 10^3$ are adenine, $4 \times 10^3$ are guanine, and $8 \times 10^3$ are thymine. Use this data set to estimate the probabilities for the four types of nucleotides.

Solution: The probabilities are

$$p_A = \frac{6 \times 10^3}{2 \times 10^4} = 0.3$$
$$p_G = 0.2$$
$$p_T = 0.4$$
$$p_C = 1 - p_A - p_G - p_T = 0.1$$

2. A protein sequence is 500 amino acids length. Assume amino acids occur independently in the sequence. If we are given that the probability for leucine is \(q_L = 0.04\), how many leucine residues we are expected to see in this sequence?

Solution: We expect to see $500 \times q_L = 20$ leucines in this sequence.

3. Prove Bayes’ theorem that

$$P(E | F) = \frac{P(F | E)P(E)}{P(F)}$$

where all the conditional probabilities are assumed well defined.

Proof: By conditional probability definition, we have

$$P(E | F) = \frac{P(EF)}{P(F)} \quad (i)$$

And \(P(EF)\) can be represented as

$$P(EF) = P(F | E)P(E) \quad (ii)$$

Substituting equation (ii) to equation (i), we get

$$P(E | F) = \frac{P(EF)}{P(F)} = \frac{P(F | E)P(E)}{P(F)} \quad (iii)$$

which completes the first equality of the question. For the second equality, note that \(F = FE \cup FE'\), hence

$$P(F) = P(FE) + P(FE') = P(F | E)P(E) + P(F | E')P(E')$$

Substituting this to equation (iii), we obtain the desired equation.
4. Suppose that the chance of rain tomorrow depends on previous weather conditions only through whether or not it is raining today. Suppose also that if it rains today, then it will rain tomorrow with probability 0.6; and if it does not rain today, then it will rain tomorrow with probability 0.3. Show that the process is a two-state Markov chain. Calculate all the transition probabilities.

Solution: The process defines two states of the weather, rain (R) or not rain (N). The process is a Markov chain, because the state of the next day depends only on the state of the current day, disregarding all the previous weather states. The transition probabilities can be represented in a table below:

<table>
<thead>
<tr>
<th>Transition probability</th>
<th>R</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>0.6</td>
<td>0.4</td>
</tr>
<tr>
<td>N</td>
<td>0.3</td>
<td>0.7</td>
</tr>
</tbody>
</table>

5. Consider a nucleotide sequence ‘AGCGCG’.
   (a) Use both the ‘+’ Markov model and the ‘-’ Markov model defined in the text to obtain two probabilities of this sequence. When applying Equation (3), we assume the start residue is equally likely to be each of the four nucleotides. The transition probabilities are from the table in the text.
   (b) Compare the probabilities calculated in (a). What region we can predict for this sequence, CpG island or normal region?

Solution: (a) Under ‘+’ and ‘-’ models, the probabilities of the sequence are respectively

\[
P^+(AGCGCG) = 0.25 \times 0.426 \times 0.339 \times 0.274 \times 0.339 \times 0.274 = 0.000919
\]

\[
P^-(AGCGCG) = 0.25 \times 0.285 \times 0.246 \times 0.078 \times 0.246 \times 0.078 = 0.000026
\]

(b) Since the probability under CpG island model, the ‘+’ model, is about 35 times of the probability under normal model, the ‘-’ model, we predict the sequence is from CpG island region.