1. **[20 points] Probability review**

   (a) **[6 points]** After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease, and that the test is 99% accurate (i.e. the probability of testing positive given that you have the disease is 0.99, as is the probability of testing negative given that you don’t have the disease). The good news is that this is a rare disease, striking only 1 in 10,000 people. Why is it good news that the disease is rare? What are the chances that you actually have the disease?

   (b) **[14 points]** It is quite often useful to consider the effect of some specific propositions in the context of some general background evidence that remains fixed, rather than in the complete absence of information. The following questions ask you to prove more general versions of the product rule and Bayes’ rule, with respect to some background evidence $E$.

   • Prove the conditionalized version of the general product rule:
     \[
     P(A, B | E) = P(A | B, E) P(B | E)
     \]

   • Prove the conditionalized version of Bayes’ rule:
     \[
     P(A | B, E) = \frac{P(B | A, E) P(A | E)}{P(B | E)}
     \]

2. **[30 points] MLE for Bayesian networks**

   Consider $N$ genes that are known to be involved in lipid biosynthesis. Say that we know roughly how they regulate each other’s expression. However, in order to understand the molecular level mechanism of lipid synthesis or of related phenotypes (such as obesity), we need a more detailed picture of their regulatory interactions. Thus, you decided to represent these interactions by using the Bayesian network and learn its parameters from microarray expression data.

   Denote by $x_1, \ldots, x_N$ the variables representing the expression levels of those $N$ genes. For simplicity, we assume that each $x_i$’s value is discretized to two levels, i.e., $x_i \in \{\text{up, down}\}$. Let’s assume that the conditional independence assumptions are already known based on the prior knowledge; so the structure of the Bayesian network is fixed. Given the data $D = \{x[1], \ldots, x[M]\}$, where $x[m]$ consists of an instantiation of all the variables $x_1, \ldots, x_N$, your goal is to learn the parameters $\theta_{x_1 | \text{pa}_1}, \ldots, \theta_{x_N | \text{pa}_N}$ by using MLE.

   (a) **[5 points]** You decided to use the table CPDs to represent the statistical dependencies between $x_i$ and its parents $\text{pa}_i$. Then, for the following sub-network including $x_1$ and its parents $x_2$ and $x_3$, describe how $\text{pa}_i$ and the parameters $\theta_{x_i | \text{pa}_i}$, determine the distribution over $x_1$. (Hint: conditional distribution)
(b) **[10 points]** Write down the likelihood function $L(\theta : D)$. Show how the decomposition of the global problem to independent sub-problems allows us to devise efficient solutions to the MLE problem.

(c) **[15 points]** Prove that the MLE solution of the parameters are:

$$\hat{\theta}_{x_i|p_a} = \frac{M[x_i, p_a]}{M[p_a]} \text{ for } i = 1, \ldots, N$$

(Hint: Use the fact that the conditional probability is legal, i.e., $\sum x_i \theta_{x_i|p_a} = 1$.)

3. **[50 points]** HMM Write a program that implements a 2-state HMM for detecting G+C rich regions in the Dictyoglomus thermophilum H-6-12 sequence. Conceptually, state 1 will correspond to the more frequent 'A+T rich' state, whereas state 2 will correspond to the less frequent G+C-rich state. Specifically: The starting parameter values (taken from Klein et al (2002), PNAS 99: 7542-47) should be as follows:

Transition probabilities $a_{ij}$ are $a_{11} = .999, a_{12} = .001, a_{21} = .01, a_{22} = .99$.

Initiation probabilities for each state (i.e. the transition probabilities from the 'begin' state into state 1 or 2) should be .996 for state 1, and .004 for state 2; these should be held fixed throughout the Viterbi training.

Emission probabilities (which should also be held fixed) are $e_A = e_T = .291, e_G = e_C = .209$ for state 1; $e_A = e_T = .169, e_G = e_C = .331$ for state 2.

Use Viterbi training as described in class to find improved parameter estimates for the transition probabilities, holding the emission and initiation probabilities fixed at the above values. Run the training for 10 iterations, where for each iteration you:

- Use dynamic programming to find the highest probability underlying state sequence.
- Using this state sequence, compute
  - The number of states of each of the two types (1 and 2), and the number of segments of each type (where a segment consists of a contiguous series of states of the same type, that is preceded and followed by states of the opposite type or the beginning or end of the sequence).
  - New transition probabilities to be used in the next iteration.

Your output should provide:

- the name and first line of the .fna file
- the information described above (in 2. i.e. numbers of states and segments, and new probability values), for each of the 10 iterations. Give probabilities to 4 decimal places only.
- the list of G+C-rich segments (corresponding to the segments having state 2 as the underlying state) after the final (10th) round of Viterbi training, sorted by genomic position.
- your description of the first 5 of the segments found above, as found by looking up the Genbank annotations.


(Acknowledgement: This problem was generated by Benjamin Vernot for Prof. Phil Green’s GS540.)