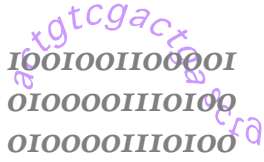


Hands-On Seventeen
Hidden Markov Model

Problem One

Consider the "Distressed Student Model" and the "Successful Student Model" discussed in class. Assume that each model has a starting state and that the transition probabilities of going from the start state to L, C, or B are 1/3.

Suppose that we have the following states visited by 3 students:

Student 1 : LLLLCLLBCLLC
Student 2 : LBBLBBBBBBBL
Student 3 : CCCLCCLBCLCL

A) Compute the probabilities of the paths for each student,

a) in the "Distressed Student Model", and
b) in the "Successful Student Model".

B) Compute the log likelihood ratios of the paths for each student,

a) in the "Distressed Student Model", and
b) in the "Successful Student Model".

Problem Two

Let us suppose that every DNA nucleotide in a sequence belongs to either a "normal" region (denoted by N) or to a GC "rich" region (denoted by R). Assume the starting state is N.

We can represent the problem by using a hidden Markov model, λ, with the following parameters:

Two states: N (for normal) and R (for GC rich) Alphabet = {A,C,T,G}.

The transition probability matrix is A, and π gives the initial probabilities:

$$A = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix} \quad \pi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

In other words: p(N|N) = 0.9, p(R|N) = 0.1
p(N|R) = 0.2, p(R|R) = 0.8

The probabilities of emission are:

$b_N(A) = 0.3$ $b_N(T) = 0.3$ $b_N(G) = 0.2$ $b_N(C) = 0.2$

$b_R(A) = 0.1$ $b_R(T) = 0.1$ $b_R(G) = 0.4$ $b_R(C) = 0.4$

a) Why is it a Hidden Markov Model? Clearly explain.

b) Compute p(TTC). Show all your work.

c) Name two additional methods you could have used for the computation of p(TTC).

d) Find the sequence Q that most likely generated TTCCC. Show all your work.

e) Find the sequence Q that most likely generated TTCCCC. Show all your work.

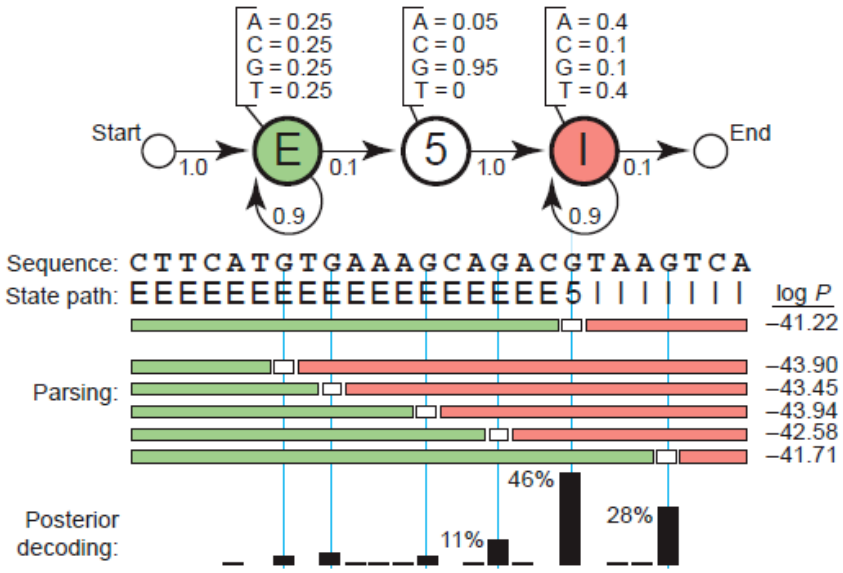
f) [Optional] Use the two methods mentioned in c) and compute p(TTC).

Problem Three

"What is a hidden Markov model" by Sean Eddy appeared in "Nature Biotechnology" in 2004.

Carefully read and understand the whole paper. As mentioned in the article, Figure 1 is a toy HMM for 5' splice site recognition. The 3 non-silent states are denoted by: E for exon, 5 for 5' splice site, and I for intron. The emission probabilities of the 4 bases for the 3 states are given (they are written above the states in the diagram). The transition probabilities of going from one state to the other are also given. Note that all state paths will always have the same pattern: a number of E's (at least one) followed by exactly one "5" followed by a number of I's (at least one).
Suppose we are given the following observed sequence:
CTTCATGTGAAAGCAGACGTAAGTCA (same as in Figure 1 of article)

a) How many different state paths can generate (or equivalently, accept) the above observed sequence?



b) Figure one gives six possible state paths that can generate (or accept) the observed sequence: CTTCATGTGAAAGCAGACGTAAGTCA.
Compute the probability: P, of each state path and then take the log base 10: log P, to obtain the 6 numbers reported in Figure 1.

c) [Optional] Explain "Posterior decoding" that appears at the bottom of the figure. In other words, understand and reproduce the 3 percentages shown in the figure: 11%, 46%, and 28%.