Andrew Coonce
CS576 – Introduction to Bioinformatics
Homework 3
October 29, 2010

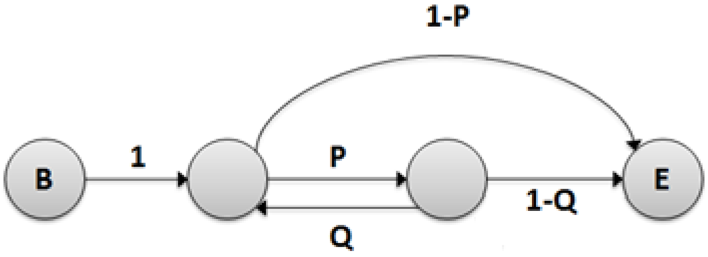
1a) The distribution of lengths, $P(\ell)$, for the state transition diagram given as Figure 1 is as follows:

$$P(\ell) = \left\{ \begin{array}{cc} 0 & if\ \ell < 4 \\ 1 - p & if\ \ell = 4 \\ p\left(q^{\ell-4}\right)(1-q) & if\ \ell > 4 \end{array} \right\}$$

1b) Given the following length distribution:

$$P(\ell) = \left\{ \begin{array}{cc} 0 & if\ \ell = 0 \\ (1-p)(pq)^{(\ell-1)/2} & if\ \ell > 0, odd \\ (1-q)p(pq)^{(\ell-2)/2} & if\ \ell > 0, even \end{array} \right\}$$

The state-transition diagram for the Markov chain is:



B and E are the silent begin and end states.

2) Given a hidden Markov model, M, with k states, it's possible to compute $P(x|M)$ with $O(k)$ space. In short, since the conditional probability distribution of the hidden state variable x_t is dependent only on the value of the hidden variable x_{t-1} according to the Markov property. Knowing this, we can calculate the distribution for x_t from x_{t-1}, discard the values used to calculate x_{t-1}, then use the distribution of x_t to calculate the distribution at x_{t+1}.

3) In modeling the DNA sequence optimization problem (Problem 2 of Homework 2) as a hidden Markov model…

3a) … the observed data is the DNA sequence.

3b) … the hidden states are the four possible dominant bases, representing the predominant base for the partition in which the base is placed.

3c) … the emission parameters represent the reward received for placing a given base in one of the hidden states. Specifically, if the base x is in the hidden state s_x where x is the dominant base of s_x, it would receive a reward of b. If the base x is some hidden state s_y where $y \neq x$, it receives a reward of c (in this case, as c is strictly less than b in most applications, we would view this reward as a penalty).

3d) … the transition parameters represent the act of switching from one partition to another (if the transition is from one state to another) or the act of extending the current partition (if it is a loop from the state back to itself). In this representation, a would be the reward (again, most frequently a penalty) associated with introducing a new partition or, equivalently, transitioning from one hidden state to another.

3e) … the Viterbi algorithm should be used to partition the DNA sequence as it is capable of finding the most likely sequence of hidden states; in this case, that would mean finding the optimal positions to place the partition boundaries.

4) Given the HMM specified in Figure 2 and the observed sequence $x = CGTCAG$…

4a) … $P(x|M)$ for the forward algorithm is:

	C	G	T	C	A	G	-
S1	0.0500000	0.0162000	0.0013018	0.0001043	0.0000334	0.0000107	0.0000202
S2	0.2000000	0.0180000	0.0058183	0.0018634	0.0001491	0.0000119	0.0000204

4b) … the most likely path of hidden states using the Viterbi algorithm is:

$S_2 \rightarrow S_2 \rightarrow \{S_1, S_2\} \rightarrow \{S_1, S_2\} \rightarrow S_1 \rightarrow S_1$

	C	G	T	C	A	G
S1	0.1000000	0.0640000	0.0058000	0.0007840	0.0006976	0.0002602
S2	0.4000000	0.0340000	0.0160000	0.0055840	0.0004624	0.0000509

4c) … $P(x|M)$ for the backward algorithm is:

	C	G	T	C	A	G
S1	0.0000055	0.0000399	0.0001088	0.0006080	0.0056000	0.0170000
S2	0.0000101	0.0000289	0.0002528	0.0007520	0.0020000	0.0080000

4d) … the posterior probabilities $P(\pi_i = 1|x, M)$ for $i = 1 \dots 6$, using the results for the forward and backward algorithms, are:

	C	G	T	C	A	G
S1	0.119826	0.554081	0.087836	0.043295406	0.385458742	0.656441718
S2	0.880174	0.445919	0.912164	0.956704594	0.614541258	0.343558282