Group HW5
CS 685-001/PPA 784-003/STA 695-001 Fall 2009

Due: Due Nov 5th, 2009

Problem 1 *Suppose we have the hidden state space S' and observed state space S with the transition matrix and emission matrix:*

$$\theta = \begin{array}{c} \\ \text{fair} \\ \text{loaded} \end{array} \begin{array}{c} \begin{array}{cc} \text{fair} & \text{loaded} \end{array} \\ \begin{pmatrix} 4/5 & 1/5 \\ 5/6 & 1/6 \end{pmatrix} \end{array}$$

$$\theta' = \begin{array}{c} \\ \text{fair} \\ \text{loaded} \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/12 & 1/6 & 1/12 & 1/4 & 1/24 & 1/24 \end{pmatrix}. \end{array}$$

respectively. The initial distribution is $2/3$ *for fair die and* $1/3$ *for loaded die.*

Compute the probability to observe the sequence 46.

Problem 2 *Go back to the bar hopping example. Suppose we have the transition matrix*

	A	C	G	T
Your die A	1/4	1/4	1/4	1/4
Arne's die C	1/5	1/5	2/5	1/5
Connie's die G	1/3	1/3	1/6	1/6
Dave's die T	1/6	1/3	1/3	1/6

and suppose you choose the initial bar at random (i.e., probability 1/4 for going to each bar). If you and your friends go to the bar and record which bar you go, you observe the sequence ACGTTTCGA. What is the probability to observe the sequence.

Problem 3 *Compute $P(X_{25} = A)$? Please note that you do not want to compute by hand....*

Problem 4 *Also compute a stationary distribution π. Is it unique and the same as the limiting distribution π?*

Problem 5 (CpG island) *DiAN has three tetrahedral dies. The first die corresponds to DNA that is* `G` + `C` *rich, the second die corresponds to DNA that is* `G` + `C` *poor, and the third is a fair die.*

$$\begin{array}{r} \\ \text{first die} \\ \text{second die} \\ \text{third die} \end{array} \begin{array}{c} \begin{array}{cccc} \texttt{A} & \texttt{C} & \texttt{G} & \texttt{T} \end{array} \\ \begin{pmatrix} 0.15 & 0.33 & 0.36 & 0.16 \\ 0.27 & 0.24 & 0.23 & 0.26 \\ 0.25 & 0.25 & 0.25 & 0.25 \end{pmatrix}. \end{array} \tag{1}$$

The transition matrix (warning: I completey made up!!) is

$$\begin{array}{r} \\ \text{first die} \\ \text{second die} \\ \text{third die} \end{array} \begin{array}{c} \begin{array}{ccc} \text{first} & \text{second} & \text{third} \end{array} \\ \begin{pmatrix} 31/100 & 33/100 & 9/25 \\ 29/100 & 4/25 & 11/20 \\ 1/12 & 1/12 & 5/6 \end{pmatrix}. \end{array}$$

The initial distribution for the hidden states is $1/2$, $1/4$, $1/4$ *for the first, second and third dies, respectively. What is the probability to observe the observed seqeunce data* `ACGTTTCGA` *with the sequence with hidden states, third third third first first second third third third?*

Problem 6 *Go back to the example,*

$D =$

	1	2	3	4	5	6
1	0	6	8	9	12	11
2	6	0	6	7	10	9
3	8	6	0	3	6	5
4	9	7	3	0	5	4
5	12	10	6	5	0	5
6	11	9	5	4	5	0

Reconstruct the tree vie the NJ method by NOT hand.

Problem 7 *Download* `PHYLIP` *software from* `http://www.phylip.com/` *and download the data sets 5000 data sets comprising the true 40-taxon trees and the corresponding 500-bp homologous sequences from* `http://www.atgc-montpellier.fr/phyml/datasets.php`

Problem 8 *A software* `treedist` *from* `PHYLIP` *computes Robinson-Foulds symmetric distance between two trees* `http://portal.litbio.org/Registered/Help/phylip/doc/treedist.html` *Take the first 1,000 data from the data sets you downloaded and construct the NJ trees from these data (under Kimura 2 parameter model for computing pairwise distances). Then compare the tree topologies between each reconstructed NJ tree and the true tree via Robinson-Foulds symmetric distance (so you have to do 1,000 comparisons so you do not want to do by hand).*

Problem 9 *From these results, get the five number summary statistics and get the histgram of the Robinson-Foulds symmetric distances you computed in Problem 2.*