CS 530: Geometric and Probabilistic Methods in Computer Science
Homework 3 (Fall '06)

1. Download the complete amino acid sequence for the chromosome of the bacterium, *Buchnera*, from the class webpage. Amino acids are represented by single characters from the set:

   $$\{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y, \$\}$$

   Using MATLAB, compute the distribution of the amino acids. Plot the distribution and compute the entropy of the *Buchnera* chromosome.

2. Using the Huffman coding algorithm (executed by hand if you wish), devise a variable-length prefix code based on the code alphabet, $\{A, G, C, T\}$. Compute the efficiency of the code you devise and compare it to the efficiency of the fixed-length "genetic code" used by all living cells. Is your code more or less efficient?

3. Assuming that the amino acid sequence of the *Buchnera* chromosome can be modeled as a first order Markov process, compute the conditional entropy per amino acid. [Hint: Use the amino acid sequence to build a 21×21 transition matrix, $P_{t\,|\,t-1}$ and compute the conditional entropy, $H_{t\,|\,t-1}$, as described in the class notes.]

4. Using MATLAB, compute the eigenvectors and eigenvalues of the transition matrix you constructed in the last problem. Is the Markov process irreducible and aperiodic? If so, plot the limiting distribution and compare it to the distribution of amino acids you plotted in the first problem.

5. You are going to successively flip a coin until the pattern HHT appears; that is, until you observe two successive heads followed by a tail. In order to calculate some properties of this game, you set up a Markov chain with the following states: S, H, HH, and HHT, where S represents the starting point, H represents a single observed head, HH represents two successive heads, and HHT is the sequence you are looking for. Observe that if you have just tossed a tails, followed by a heads, a next toss of a tails effectively starts you over again in your quest for the HHT sequence. Set up the transition probability matrix.

6. Let $x_i^{(n)}$ denote the quality of the n-th item produced by a production system with $x_0^{(n)}$ being "good" and $x_1^{(n)}$ being "defective." Suppose that $\mathbf{x}$ evolves as a Markov chain whose transition probability matrix is:

$$\begin{bmatrix} .99 & .12 \\ .01 & .88 \end{bmatrix}$$

What is the probability that the fourth item is defective, given that the first item is defective?

7. Consider the Markov process whose transition probability matrix, $\mathbf{P}$, is given by:

$$\begin{bmatrix} .4 & .0 & .0 & .1 \\ .0 & .7 & .3 & .2 \\ .3 & .2 & .3 & .4 \\ .3 & .1 & .4 & .3 \end{bmatrix}$$

Suppose that the initial distribution is $\mathbf{x}^{(0)} = [.1\ .1\ .5\ .3]^{\mathrm{T}}$.

(a) Compute $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \mathbf{x}^{(3)}$ and $\mathbf{x}^{(4)}$

(b) Compute $\mathbf{P}^2, \mathbf{P}^3$ and $\mathbf{P}^4$.

(c) Compute $\mathbf{x}$ where $\mathbf{x} = \mathbf{P}\mathbf{x}$.

8. Hazel and Naomi each have a brown paper grocery bag containing three exotic fruits. Initially, Hazel's bag contains three persimmons and Naomi's bag contains three kumquats. Once every day, Hazel and Naomi swap one fruit chosen at random from their bags.

(a) Derive an expression for the probability that Hazel will have $k+1$ kumquats today given that she had k kumquats yesterday.

(b) Derive an expression for the probability that Hazel will have k kumquats today given that she had k kumquats yesterday.

(c) Derive an expression for the probability that Hazel will have $k-1$ kumquats today given that she had k kumquats yesterday.

(d) Give the transition matrix for the Markov process.

(e) Is the Markov process irreducible? aperiodic? Prove your answers.

Hint: Observe that the number of kumquats which Hazel has in her bag always equals the number of persimmons which Naomi has in her bag and *vice versa.*

9. In this exercise you will simulate the Ising model, a standard model of the emergence of spatial organization in ferromagnetic materials. The input to your program, written in MATLAB, is a matrix representing a 64×64 toroidal lattice of randomly oriented spins, *i.e.*, the spin at each location equals $+1$ with probability $\frac{1}{2}$ and -1 with probability $\frac{1}{2}$. It will return the matrix of spins after n iterations of the Gibb's sampling procedure, using the conditional p.m.f. computed as described in class. This matrix can be displayed as a grey scale image using *gdisplay*. Hand in your MATLAB code and hardcopy of the results of your simulation for $n = 10^3, 10^4, 10^5$, and 10^6.