

# 6.095/6.895 Fall 2005 - Problem Set 2

Due: October 10, 2005 at 8pm

If you feel that any question is ambiguous, state any assumptions you need to make and consult them with the TA during office hours (Fridays at 1pm) or via email.

1. One biological application of Hidden Markov Models is to determine the secondary structure (i.e. the general three dimensional shape) of a protein. This general shape is made up of alpha helices, beta sheets, turns and other structures. In this problem we will assume that the amino acid composition of these regions is governed by an HMM.

In order to keep this problem relatively simple, we do not use actual transition values or emission probabilities. We will use the state transition probabilities of:

	Alpha Helix	Beta Sheet	Turn	Other
Alpha Helix	0.7	0.1	0.0	0.2
Beta Sheet	0.1	0.5	0.3	0.1
Turn	0.3	0.3	0.3	0.1
Other	0.2	0.2	0.0	0.6

where, for example,  $P(\text{Alpha Helix} \rightarrow \text{Turn}) = 0$ . The start state is always Other and the emission probabilities<sup>1</sup> are:

Amino Acid	Alpha Helix	Beta Sheet	Turn	Other
M	0.30	0.15	0.00	0.05
Q	0.30	0.05	0.15	0.15
V	0.15	0.30	0.10	0.20
I	0.15	0.30	0.10	0.15
P	0.05	0.00	0.35	0.20
G	0.05	0.20	0.30	0.25

- (a) Draw this HMM. Include the state transition probabilities and the emission probabilities for each state.
  - (b) What is the probability of the second emitted character being an *M*?
  - (c) Give the most likely state transition path for the amino acid sequence *PGIMQV* using the Viterbi algorithm. Include the Viterbi algorithm table for discovering this path.
2. Page 50 of Durbin gives the following transition probabilities for the models of being inside a CpG island (M1) and outside a CpG island (M2):

M1	A	C	G	T
A	0.180	0.274	0.426	0.120
C	0.171	0.368	0.274	0.188
G	0.161	0.339	0.375	0.125
T	0.079	0.355	0.384	0.182

M2	A	C	G	T
A	0.300	0.205	0.285	0.210
C	0.322	0.298	0.078	0.302
G	0.248	0.246	0.298	0.208
T	0.177	0.239	0.292	0.292

where, for example,  $P_{M1}(A \rightarrow G) = 0.426$ .

For both models we will give the initial state a uniform probability over the nucleotides.

You must write code to simulate the generation of sequences from M1 and M2 and to find the probability of a sequence being generated by M1 or M2. Note, because this is a Markov *Chain*, the emission gives us the state (the state sequence is not hidden).

<sup>1</sup>Actual amino acid frequencies can be found in T.E. Creighton, *Proteins: Structures and Molecular Properties*, 2nd Edition, page 256.

- (a) Simulate 10000 sequences of length 6 from M1. Of the sequences simulated, how often is the probability of the sequence being generated by M2 higher than the probability of it being generated by M1?
- (b) Now, simulate 10000 sequences of length 6 from M2. Of the sequences simulated, how often is the probability of the sequence being generated by M1 higher than the probability of it being generated by M2?
- (c) What do the two parts above tell you about the potential for error when classifying sequences into regions of CpG islands using a simple likelihood estimate?
- (d) If you had a prior on how often a sequence was part of a CpG island, how would you use this information to better classify sequences?

In addition to your answers above, include a printout the code you used.

3. Perform a hierarchical clustering (i.e. give the cluster tree) of the genes  $a \dots e$  where their pairwise distances are given by:

	a	b	c	d	e
a	0	3	8	7	8
b	3	0	4	8	8
c	8	4	0	5	6
d	7	8	5	0	6
e	8	8	6	6	0

Using:

- (a) Single link clustering ( $D(X, Y) = \min_{x \in X, y \in Y} D(x, y)$ )
  - (b) Complete link clustering ( $D(X, Y) = \max_{x \in X, y \in Y} D(x, y)$ )
  - (c) Average link clustering ( $D(X, Y) = \frac{1}{|X||Y|} \sum_{x \in X, y \in Y} D(x, y)$ )
4. The clusters generated by a given algorithm are often highly dependent on the distance function used to compare two items. In the case of microarray data, we may vary the function we use on the basis of what types of clusters we are looking for. For each of the situations below on multidimensional microarray data, give some properties we may want in a distance function and suggest an example of one (such as Euclidean distance, correlation, etc.). Provide some justification for your choice. For this question you may assume our microarray data directly measures the log of the level of expression of a gene without any errors.
    - (a) We want to cluster genes that have similar levels of expression for each of the dimensions.
    - (b) We want to cluster genes whose expression patterns may differ in magnitude but follow the same overall pattern.
    - (c) We want to cluster genes whose expression patterns may be different in magnitude as above but may also differ in sign.
  5. **6.895 Problem.** The clustering algorithms we have seen so far in class are direct applications of a variety of methods from machine learning. Several issues with microarray data have led to specialized algorithms. For each of the issues below, identify a journal or conference article that attempts to deal with it. Pick one of the four and describe the methods used in more detail (limit response to one or two paragraphs).
    - (a) The data may be noisy due to technical (i.e. microarray fabrication inaccuracies) or biological (i.e. variation in DNA of the individual) reasons.
    - (b) Genes that respond to a stimulus may do so at different times in time series microarray data.
    - (c) Certain genes may not be similar to any other gene on the basis of microarray data and thus should not be assigned to any cluster.
    - (d) The size or number of clusters is difficult to provide as a parameter.