6.096 Algorithms for Computational Biology

Problem Set 3

Professor: Manolis Kellis

Instructions

This problem set is due on Wednesday May 4th, 2005. For problems with a programming component, the algorithms should be implemented using python. When handing in your solutions, you should turn in your python code. Collaboration is permitted, but the write-up and the code should be your own. Please acknowledge your collaborators.

Suggested readings

- Pevzner: Hidden Markov Models page 387 Clustering and Trees page 339.
- Durbin et al. Markov Chains and Hidden Markov Models page 46.

1 HMMs

1.1 Dishonest casino

A dishonest casino uses either a fair or a biased coin. The fair coin comes out heads or tails with equal probability. The biased coin comes out heads with probability $\frac{3}{4}$ and tails with probability $\frac{1}{4}$. The probability of transitioning from a biased to a fair coin and from a fair to a biased coin is $\frac{1}{10}$.Using the Hidden Markov Model(HMM) just defined representing a dishonest casino compute the most probable sequence of states that generate the following sequence of coin tosses: HHHHHTTTTT. You should fill out a 2 by 10 dynamic programming table.

1.2 Probability at a given position of being generated by a state

With the same HMM, what is the probability that the 'T' at the seventh position is generated by a biased coin?

1.3 Probabilities of longer sequences

Using the Viterbi algorithm python code provided (or writing you own from scratch if you prefer), use the previous HMM to find the most probable sequences of hidden states that generate the following sequence of coin tosses: HHHHHHHHHHHHHHHHHHHHTTTTTTTTTTTTTTT (that is 20 Heads in a row followed by 15 Tails in a row).

1.4 CpG islands

Regions near genes contain CpG di-nucleotides with higher frequency than in the rest of the genome. These regions are called pCG-islands. The transition probabilites in CpG-islands differ from transition probabilities in non-CpG islands. The following transition probability tables taken from page 50 in Durbin's book are obtained from the statistics of annotated genomic sequences.

The transition probabilities for ACGT outside CpG-islands:

0.180	0.274	0.426	0.120
0.170	0.368	0.274	0.188
0.161	0.339	0.375	0.125
0.079	0.355	0.384	0.182

The transition probabilities for ACGT inside CpG-islands:

0.300	0.205	0.285	0.210
0.322	0.298	0.078	0.302
0.248	0.246	0.298	0.208
0.177	0.239	0.292	0.292

The above 2 tables specify the transition probabilites within each group of the HMM. Now it is your task to design the transition probabilities accross the groups. Use your HMM to search the strech of genomic sequence provided. The number of CpG-islands will vary depending on the parameters for switching between CpG-island and non CpG-island. With some reasonable parameters you may get 4-5 islands.

2 Ultrametric

In class, we defined defined *Ultrametric Distance* as a set of pairwise distances that satisfy that for any three points two distances are equal and the third is smaller.

2.1 Ultrametric and UPGMA tree

Does the following distance matrix between 4 points (a, b, c, d) define an ultrametric? If it does, build the UPGM binary tree obeying the pairwise distances.

	a	b	c	d	e
a	0	1	3	3	5
b	1	0	3	3	5
c	3	3	0	2	5
d	3	3	2	0	5
e	5	5	5	5	0

2.2 Uniqueness

Give a sketch of the proof of correctness of UPGMA. First show that the algorithm returns the correct tree when the distances are ultrametrix. Then show that the tree is unique.

3 Hierarchical clustering

Given a connected graphs with weights, A *Minimum Spanning Tree* is a tree that connects all the vertices in the graph whose total weight is minimal. Kruskal's algorithm, taught in 6.046, is a greedy algorithm that calculates a minimim spanning tree. It starts out with all vertices being a forest i.e. a collection of single-vertex trees. It adds an edge of minimum weight connecting elements of the forest until we have a single tree.

This algorithm is direclty related to hierarchical clustering. Hierarchical clustering is a bottom-up approach for clustering. It starts with n single-element clusters. Then it combines the two closest clusters, which leave us with n-1 cluster, and so on. In general, the ith partition combines the two closets clusters from the (i-1)the partition and has $n - i + 1$ clusters.

Once the clusters have more than one element, there are several ways of computing the distance between the newly formed cluster and the other clusters. These different approaches, give different final trees. Some ways of defining the distace are: s-link, the smallest distance between any pair of elements

between the two clusters, c-link, the largest distance between any pair of elements between the two clusters, and a-link, the average distance between the elements of the two clusters.

3.1 Minimum Spanning Tree and Kruskal's algorithm

One of the choices for calculating the distance between clusters makes Kruskal's algorithm for Minimum spanning trees and Hierarchical clustering equivalent. Which one is it? Explain the correspondence that makes the two algorithms equivalent.

3.2 Hierarchical CLustering and Phylogenetic Trees

Explain the method that is used for calculating distance in the hierarchical clustering done by phylogenetic trees (UPGMA).

4 BONUS Hierarchical clustering and Phylogenetic Trees

This bonus question gives you the opportunity to play with code and real biological data. Modify the hierarchical clustering python code provided so that it can calculate the three different cluster metrics: s-link, c-link, and a-link. Use the modified code to run on the biological data provided. You will get a tree for each metric. Compare the trees. The biological data provided is the output of clustal-w, a popular multiple sequence alignment software, on 15 sequences corresponding to 15 different species. Note that you need to take 100 - the clustalw score as the distance between species.