6.047 / 6.878 Computational Biology: Genomes, Networks, Evolution
Fall 2008

6.047/6.878 Fall 2008 - Problem Set 4

Due: Wednesday October 22, 2008 at 8pm

1. **Representing phylogenetic trees.**

   It is convenient to represent rooted phylogenetic trees as binary trees with uniquely labelled leaves. Binary trees, however, impose an inherent left-to-right ordering of nodes that has no biological or phylogenetic relevance. For example, the trees ((a,b),(c,(d,e))) and ((c,(e,d)),(a,b)) represent the same evolutionary history and we can call them *isomorphic* binary trees.

   (a) For a given phylogenetic tree with n leaves, how many binary trees are isomorphic to it (i.e. same topology, differing only in left-to-right order)? Justify your answer.

   (b) Lets say we wanted to avoid this redundancy in the binary tree representation by defining one of the many isomorphic binary trees to be the "canonical" tree (define this distinction however you wish). Give an efficient algorithm that produces a canonical binary tree T given *any* binary tree isomorphic to T. State and justify the runtime of your algorithm.

   Notes: Your algorithm for finding a canonical tree should behave such that, given any two isomorphic trees T_1 and T_2, canonical(T_1) should be the same binary tree as canonical(T_2). Assume you have any $O(1)$ binary tree operations you need, such as Left(x) for the left-hand child, Right(x), Parent(x), IsLeaf(x), Label(x), SwapChildren(x), etc.) Also assume there exists an ordering on the leaves (e.g. each leaf has a unique small ID number or short text string)
   Among other applications, this algorithm can be used to hash trees by converting your canonical form to a string.

   (c) How would you represent and canonicalize *unrooted* phylogenetic trees?

2. **Phylogenetic tree building.**

   (a) Consider the following DNA sequences:

Position	1	2	3	4	5	6
Sequence 1	A	A	C	C	G	G
Sequence 2	A	C	T	C	A	G
Sequence 3	G	T	C	C	T	T
Sequence 4	G	G	T	T	C	G

   Consider the three possible unrooted trees on four elements. For the sequences above, give the cost of each position for each tree. Also give the total cost of each tree and indicate the lowest cost tree. Assume the cost of a transition ($A \leftrightarrow G$, $C \leftrightarrow T$) is 1, the cost of a transversion (any other mismatch) is 2 and there is no cost for a match.

   (b) Perform UPGMA clustering on sequences with distances:

	a	b	c	d	e
a	0	3	11	10	12
b	3	0	12	11	13
c	11	12	0	9	11
d	10	11	9	0	8
e	12	13	11	8	0

Include all intermediate trees and distance matrices resulting from the application of UPGMA as shown in class.

Notice that the tree does not perfectly represent the original pair-wise distances shown above. Keeping the topology found by UPGMA above, what would the branch lengths need to be to perfectly match the distance metric? What algorithm seen in class does this, and how?

3. **Positive selection in the human genome.** In this problem, we will analyze a region of human chromosome 5 to identify single nucleotide polymorphisms (SNPs) that exhibit characteristic signatures of recent positive selection in human populations.

   (a) **Long haplotype** tests are one approach to detect regions of the genome that may be under selection.

      i. Briefly explain why long haplotypes are evidence for recent selection.

      ii. Why is it difficult for haplotype-based methods to identify individual polymorphism(s) under selection, especially if the selection is very strong?

      Cross-population extended haplotype homozygosity (XP-EHH) is a metric used to identify long haplotypes in one subpopulation versus another. The file `XPEHH.txt` contains XP-EHH scores for a series of SNPs in populations from Europe (`CEU`), Africa (`YRI`), and Asia (`JPT+CHB`).

      iii. Plot the scores across chromosome 5 for for the three pairwise population comparisons. You can plot XP-EHH against the position of the SNPs either in terms of their DNA sequence positions (bp) or in terms of their recombinant frequencies (cM). Which is preferable for this purpose, and why?

      iv. In which subpopulation do you see the strongest evidence for natural selection? Explain your answer.

      v. How many SNPs have XP-EHH scores above 2.0 in at least one pairwise comparison?

   (b) **Derived allele frequency.** For a SNP, we distinguish between the *ancestral allele*, the allele present in the common ancestor, and the *derived allele*, arising from a recent mutation and possibly under selection. One way to study the strength of selection on a derived allele is to determine how much it has spread through the population.

      To approach this for a given SNP, we need to know which of its alleles is the ancestral allele, and the derived allele frequency in the modern population.

      i. One way to infer the ancestral allele is to assume that it is the base observed in a closely related species – chimpanzee is often used for humans. However, this may sometimes lead to an erroneous conclusion as to which of the SNP's alleles is the ancestral allele. Explain why, and give a back-of-the-envelope estimate for how likely this is to occur for a given SNP, considering that the mean human-chimp sequence divergence is 1.23%.

      ii. The file `Derived.txt` specifies how many copies of the derived allele were found among 120 European, 120 African, and 180 Asian chromosomes for each of the SNPs we are studying. Calculate the derived allele frequencies in the population that you concluded is under selection in part (a), and plot them across chromosome 5.

      iii. How many SNPs have both the long haplotype signal above and derived allele frequency above 0.6?

(c) **Population differentiation.** A third line of evidence for recent selection is given by highly differential allele frequencies between subpopulations. One way of measuring the degree of difference is to compare the *heterozygosity* within each subpopulation to the heterozygosity of the population as a whole. If p is the frequency of an allele in a population, then the expected heterozygosity is the frequency of heterozygotes in the population at Hardy-Weinberg equilibrium, $2p(1-p)$.

The statistic F_{ST} is defined as $\frac{H_T - H_S}{H_T}$, where H_T is the heterozygosity for the total population and H_S is the average heterozygosity of the subpopulations. Roughly speaking, F_{ST} tells us how much genetic differences between subpopulations, rather than genetic diversity within subpopulations, contribute to overall genetic diversity.

i. Assume we have a population composed of two equally sized subpopulations. The overall allele frequency in the population is p, and the allele frequencies in each subpopulation are $p+d$ and $p-d$. Derive a simple expression for F_{ST} in terms of p and d.

ii. Based on the derived allele frequencies for the human subpopulations in part (b), calculate F_{ST} for the subpopulation under selection against each of the other subpopulations. How do you estimate p and d? For each SNP, average these two pairwise F_{ST} values and plot them across chromosome 5.

iii. How many SNPs pass the above thresholds and also have an average $F_{ST} > 0.6$?

(d) **Function.** Finally, to facilitate follow-up studies, we would like to restrict our investigation to SNPs that fall within known or likely functional elements in the genome.

i. The file `phastConsElements.txt` gives the coordinates (in bp) of regions that are evolutionarily conserved in vertebrate species, and the file `genes.gff` gives the exons and introns of known protein-coding genes in the region of chromosome 5 we are studying. How many SNPs are within conserved elements? exons? introns?

ii. Based on these annotations, how many SNPs pass the above thresholds and also lie within known or likely functional elements?

(e) Based on all the evidence we've now collected, which SNP is the best candidate target of selection? If it lies within a gene, search the internet to find the function of the gene.

4. **(6.878) Divergence, coalescence and human-chimp speciation**.

Please read the following paper (a copy is provided on the course web site).

Nick Patterson, Daniel J. Richter, Sante Gnerre, Eric S. Lander and David Reich (2006). Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103-1108.

The authors present evidence that human and chimp ancestors interbred at some point long after their initial divergence. Using concepts presented in class and/or discussed in the paper, explain the basis for this claim in ≤ 1 page. How do the observed data differ from what we would expect from a simple allopatric speciation? Why are the observations from chromosome X particularly telling? Why is the authors' model more consistent with the data? Can you think of any alternative explanations?