6.047/6.878/HST.507
Fall 2010 - Problem Set 4: Evolution and phylogenomics

Due: Monday, November 22 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) Let us 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. **Variant discovery in high-throughput sequencing**

One application of high-throughput sequencing is to discover differences in a particular sample (an individual, a cancer cell, or a strain or closely-related species) when compared to a reference genome. In this problem you will design a method to accurately identify single nucleotide polymorphisms (SNPs) from aligned short sequencing reads.

(a) An important consideration in identifying SNPs from sequencing data is the error rate of the underlying sequencing technology. In many situations, the sequencing error rate is higher than the true polymorphism rate. What implications does this have for identifying SNPs?

(b) Describe a method to estimate the empirical error rate, given a particular aligned sequencing dataset. What types of errors should you allow and assess? How does this change based on the sequencing technology in use? Considering these issues, estimate the empirical error rate from the provided (real) sequencing reads in `yeast_reads_aligned.txt`. Each line follows this format: `observed_sequence ref_sequence ref_position`

Note that the data has been simplified to have only one chromosome and one read direction (strand orientation). Assume the organism is haploid.

For extra credit, describe any higher-order factors affecting the error rate. For example, are there read position or base-specific differences? What biological or technical mechanisms could cause this, and how should they be handled in downstream computations?

(c) Design a method to identify SNPs from the aligned sequencing reads. Ideally, this method will use your estimated error rate in a principled manner. What other parameters are needed, and how might you obtain them? What assumptions does your method make, and how might they be violated in practice? Apply your method to the provided reads and submit your top (highest confidence) 50 SNP predictions.

(d) Assume the sampled DNA reads were from an individual with a particular disease. Outline a method to predict the biological impact of a particular SNP. What other data sources would be needed to improve this prediction? What problems do you foresee occurring with this predictive approach? What other experiments would increase the evidence that a particular SNP has a functional impact, possibly relating to the disease in question?

5. **(6.878) Wright-Fisher process and the coalescent**

In this problem, we will explore statistical models for population genetics.

(a) First, write a program to simulate the Wright-Fisher reproduction process. Your program should accept as input the population size and number of generations to simulate.

i. Recall that the basic Wright-Fisher model assumes clonal reproduction of haploid individuals and that the population size remains constant. In each generation, each possible parent produces an infinite number of progeny, and a new population is selected from these childen (so the sampling is with replacement).

ii. Now, extend your simulator to track the coalescence times of lineages of the most recent individuals. If you are observing k individuals, you should report the $k-1$ generations where coalescence occurred. Run your simulator for 1000 times with a population size of $N = 500$. Report the average and standard deviation of the first coalescence times (across 1000 trials) of $k = 2$, 3, and 4 individuals. Explain how you selected the number of generations to simulate. How would your results change if you simulated too few or too many generations?

iii. Do your results agree with the coalescent approximation? Justify your answer (agreement or disagreement) with a brief quantitative argument and explain why you think it does or does not agree.

(b) We will now extend the simulation to approximate sexual reproduction.

i. Adjust your simulation so that the originating ancestors have a known gender, assigned randomly. There will F females and $M = N - F$ males (and this ratio will be constant in all generations). We will implement a simplified model of sexual reproduction where each individual is haploid and selects its chromosome from one of its two parents (without recombination). In each generation, possible parental combinations are formed by pairing all female chromosomes with all male chromosomes (assume these are only autosomal chromosomes). These pairs are then split up into haploid choices for all possible individuals. The next generation is chosen by selecting from this set of (haploid) chromosomes. Of the N individuals in this next generation, F will be female and $M = N - F$ will be male. You might verify your simulation by testing with $F = M = 250$ and comparing your results to those from the previous section.

ii. Assume that there are $F = 100$ females and $M = 400$ males. As in the previous section, perform 1000 simulations for $k = 2$ individuals. Report the average and standard deviation of the coalescence times. Do your results agree with the coalescent approximation? Again, provide a quantitative justification and a brief explanation.

iii. **Extra credit:** If your results do not agree with the (standard) coalescent approximation, can you extend the coalescent approximation to incorporate this gender imbalance? You might approach an answer empirically by comparing simulation results for various values of N and M or F. Hint: an extension might incorporate the expression $\frac{4MF}{N^2}$.

6. **Project progress update reminder**

   Your midcourse progress report is also due on Monday, 11/22.