41-42. Gene trees

[the misprint in the problem number and the answer to the first problem are corrected]

- In humans and other vertebrates, the mitochondrial genome (**mtDNA**) is inherited from the mother. Two individuals with the same mother have the same mtDNA genotype, except for any mutations that occurred. So do two individuals with the same maternal grandmother. And so on.
- Any two individuals have mtDNAs whose ancestry can be **traced to a common ancestor** at some time in the past. If they are closely related through their material ancestors, it will be only a few generations in the past. If they are unrelated, it might be much longer in the past. But eventually they have to have a common ancestor. How different the mtDNA sequences of two individual are depends on how many mutations have accumulated in each lineage since the **most recent common ancestor** (MRCA).
- A useful way to describe the ancestry of the maternal ancestry of two or more individuals is by a **gene genealogy** or **gene tree**. With two individuals, the tree is completely characterized by the time of the MRCA, which is the **coalescence time**.
- With more than two individuals, the gene tree has a **branching pattern** and a series of coalescence times. You do not know what the gene tree of mtDNA is unless you have a complete maternal pedigree of the population, but you know there has to be a gene tree.
- If you can sequence mtDNAs from different individuals, then you can infer the gene tree from the differences in DNA sequence. There are many ways to do this. I will discuss two which are chosen because they are simple and they represent different ways of inferring trees. With 4 individuals when one is known to be distantly related to the others (the **outgroup**). there are only 3 different branching patterns to compare. If you are inferring gene trees of human mtDNA, you can use the sequence of mtDNA from a Neanderthal as an outgroup.
- Here is a hypothetical data set.

	SNP																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
I	T	G	A	T	C	C	C	T	A	T	A	T	C	G	T	C	G
II	A	C	A	G	T	A	C	T	A	A	G	T	C	G	T	C	T
III	A	G	G	G	C	A	T	T	A	T	G	T	C	G	T	C	T
N	T	C	G	T	T	C	C	G	G	T	G	G	A	A	G	A	T

- One way to infer the gene tree is to use **parsimony**, which assumes that mutations are rare. You pick the branching pattern that requires the fewest mutations. With only four sequences, each site requires 1 or 2 mutations. For site 1, only one change is required on the tree that groups II and III together but two changes are required for the other two trees. For site 2, one change is required for the tree that groups I and III together, and so on. For sites 7-17, only a single change is required on any of the three trees, so those sites provide no information about the branching pattern when parsimony is used. The parsimony count for I+II tree is 22: for the I+III tree the score is 21; and for the II+III tree, the score is 20. Therefore parsimony favors the II+III tree.

- Another way to infer trees is to assign a **distance** between each pair by counting the number of differences in sequence. Then you compare the ways that distances add up on each possible tree. The distance matrix for the sample data is as follows.

	II	III	N
I	8	7	12
II		5	12
III			13

  With 4 species and an outgroup, the best tree is the one in which the two most similar sequences are the most closely related. In the table, the distance between II and III is smallest so you would favor the II+III tree using a distance method. For this data set, the two methods favor the same tree, but that is not always true.
- You can also estimate the time of the MRCA by counting the substitutions between the outgroup and the average of the other sequences. In this case, the average is d=12 1/3. If you know the substitution rate K per site, and the total number L of sites sequenced, you estimate the time from the formula d=2TKL. For mtDNA, K in vertebrates is about 1x10^{-8} per site per year. If L=6200 sites, you would conclude that the time to the MRCA is 12.333/(2x10^{-8}x6200)≈ 100,000 years. In fact, the best estimate of the time to the MRCA of the mtDNA of modern humans is about 170,000 years and the time to the MRCA of human and Neanderthal mtDNA is about 660,000 years.
- The Y-chromosome in mammals, except for the two pseudoautosomal regions, is transmitted from father to son. That means there is a gene tree of the Y. The pattern of paternal ancestry is different from the pattern of maternal ancestry, so the Y tree differs from the mtDNA tree. The MRCA of the mtDNA was in a female who lived at some time in the past. The MRCA of the Y was in a male who lived at a different time in the past. The time to the MRCA of the Y in modern humans is about 70,000 years.
- When you compare mtDNAs from different species, there is still a gene tree describing their ancestry. In many cases, the gene tree of mtDNAs from each species form **monophyletic** groups, which are groups that contain all the descendents of the MRCA of each group. That is what is found for mtDNAs sampled from humans, chimps, bonobos, and Neanderthals.
- The mtDNA gene tree describes the history of each of the sites in the mitochondrial genome because there is no recombination in mtDNA. There is also no recombination in the Y-chromosome in mammals, except for the pseudoautosomal regions. Most of the human genome is autosomal and subject to recombination. Nevertheless, each small genomic region has a gene genealogy. But because of recombination, different regions have different gene trees.
- In a few cases, the coalescence times for some genomic regions predate the times of speciation and you see **trans-species polymorphism**, in which alleles in different species are more closely related to one another than they are to different alleles in the same species. An example is DQB1 gene of the human MHC complex. Trans-species polymorphism in species separated by long times is evidence of the continuing action of balancing selection.

Green RE, et al. (2008) A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing. 134:416-426.
http://www.cell.com/abstract/S0092-8674(08)00773-3#

Karafet TM et al. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Research 18:830-838.
http://genome.cshlp.org/content/18/5/830.abstract

Hughes AL, Yeager M (1998) Natural selection at major histocompatibility complex loci of vertebrates. Annual Review of Genetics 32:415-435
http://arjournals.annualreviews.org/doi/full/10.1146/annurev.genet.32.1.415

Additional problem

	SNP																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
I	T	G	A	T	C	C	C	T	A	T	G	T	C	G	T	C	T
II	A	C	A	G	T	A	C	T	A	T	G	T	C	G	T	C	T
III	A	G	G	G	C	A	T	T	A	A	A	T	C	G	T	C	G
N	T	C	G	T	T	C	C	G	G	T	G	G	A	A	G	A	T

41.1 The above table shows hypothetical nucleotide sequences of mtDNAs of 3 humans and one Neanderthal, which you can assume is the outgroup.

a. Infer the best tree using parsimony.

Ans. Only sites 1-6 affect the parsimony score. Sites 1, 4, and 6 group II and III together, 2 and 5 group I and III together, and 3 groups I and II together. Therefore the tree with II and III most closely related is favored by parsimony.

b. Infer the best tree using a distance using the number of nucleotide differences as your measure of distance.

Ans. You start by computing the number nucleotide differences between all pairs of sequences.

	II	III	IV
I	5	8	10
II		7	11
III			16

With these distances, you would group I and II together.

c. What is it about this data set that causes you to get different answers for a and b?

Ans. It appears that the data are not consistent with a molecular clock. Of the sites at which a single species is different, sites 7-17, four of them (7, 10, 11, 17) have only species III different and the rest have only species 4 different. No matter which tree you assume is correct, there are many more substations on the branch leading to III than there are on the branches leading to I and II. The difference is not statistically significant because the sample size is too small, but in general when different methods of inferring trees give different answers, it tells you that there is something odd about the pattern of substitution.

d. Which is the right tree?

You cannot decide based on the data given.