POPULATION GENETICS
Winter 2005
Lecture 17
Molecular phylogenetics

- in deriving a phylogeny our goal is simply to reconstruct the historical relationships between a group of taxa.
- before we review the principles of building phylogenetic trees, two things must be kept in mind:

1. Phylogenetic trees are hypotheses

- a phylogenetic tree is nothing more than an hypothesis.
- the tree may have very strong support, or it may have very little support.
- the former arises when there are a large number of characters supporting a specific topology.
- the latter arises when there are many possible trees that are difficult to exclude as possible alternatives.
- considerable caution must be exercised in the generation and testing of phylogenies, a point not appreciated by many researchers.

2. Gene trees are not the same as species trees

- species trees illustrate the evolutionary histories of a group of related species.
- in other words, species trees record the details of speciation for the group.
- gene trees show the evolutionary relationships among DNA sequences for a locus.
- gene trees may not be the same as species trees for one main reason – the existence of ancestral polymorphism.
- if this ancestral polymorphism is lost in some taxa but not in others, then one sequence isolated in species A may be more closely related to one in species B than to any other conspecific sequence.
- the gene tree will thus be different from the true species tree.
- the best way to guarantee that this will no occur is to use information provided by multiple independent loci!

Some terminology

- phylogenetic trees may be **rooted** or **unrooted**.
- trees are rooted by an **outgroup**, which is a taxon assumed (on the basis of fossil evidence) to have diverged prior to the group of taxa under study.
- the branching pattern of a tree is called its **topology**.
- the tree has both internal and external **branches**.
- as we learned from the lecture on coalescent theory, there is a considerable amount of information contained in a gene tree that extends far beyond the patterns of ancestry and descent.

- any type of data can be used to reconstruct phylogenetic trees.
- until recently, trees were constructed solely from morphological characters.
- now, the vast majority of researchers use molecular data.
- this molecular data can be in various forms:

1. Immunological distance.
2. DNA-DNA hybridization.
3. Allozyme data.
4. Restriction site data.
5. Amino acid sequences.
6. DNA sequences.

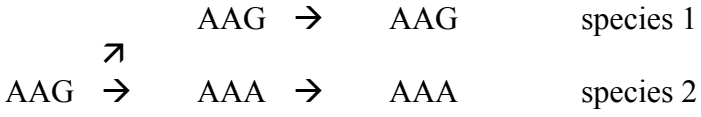
- characters may be binary (i.e., presence or absence of an isozyme allele) or multistate (i.e., ACGT)
- characters may also be ordered or unordered.
- when characters are ordered a certain directionality is implied among changes.
- nucleotide sequence data contains another important criteria: positional homology.
- the presence of insertions/deletions create problems in aligning sequences, especially for rRNA and non-coding regions.
- we will ignore this complication today.

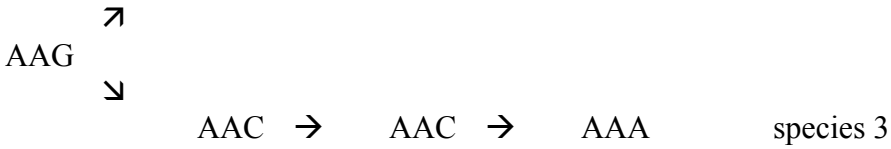
- there are two important assumptions about the characters used to build trees:

1. the characters are independent
2. the characters are homologous

- a homologous character is one shared by two species because it was inherited from a common ancestor.
- if a similar character or trait is possessed by two species but was not possessed by all the ancestors intervening between them, it is said to exhibit "**homoplasy**".
- homoplasy can result from **convergent or parallel evolution,** or from **evolutionary reversals.**
- reversals are very common at the DNA sequence level because there are only four nucleotide bases.

- the molecular characters used must also minimize **"homoplasy"**.
- homoplasy occurs when two taxa possess a certain character but that character was not present in all ancestors of the two taxa.
- homoplasy can occur by **convergent** or **parallel evolution**, or by **reversals**.
- for molecular data, evolutionary reversals are the main source of homoplasy.
- for example, consider the following series of substitutions:





- for this codon we would group species 2 and 3 together on the basis of sharing a derived character but this is an error.
- this problem is acute for third positions in codons.
- "multiple hits" at these positions are an important source of homoplasy in molecular sequence data.

How do we construct trees?

- there are three major types of phylogenetic methods:

1. Distance methods (e.g. UPGMA, NJ)
2. Maximum parsimony methods (MP)
3. Maximum likelihood methods (ML)

- MP and ML are both called "cladistic" methods.
- recently, there is a growing interest in using Bayesian tree-building methods (such as that used by the computer program MrBayes).

Distance Methods

- the general strategy behind distance methods is to cluster taxa (or OTUs) so that the most similar ones are found close together in the tree.
- this strategy is called a phenetic approach.
- the best tree, according to this approach, is to minimize the total distance among all taxa.

- all distance methods begin with an OTU x OTU matrix containing estimated distances between the taxa.
- these distances may be based on allozyme data, RS data or nucleotide sequence data.
- for allozymes, the two most common distances used are Rogers' or Nei's genetic distance.
- Nei's distance, for example, measures the probability of sampling the same allele at a locus from two species (or populations) relative to the probability of sampling the same allele twice from the same species (or population).

- for nucleotide sequences, a number of different distance measures have been proposed.
- these distances are primarily estimates of the number of nucleotide substitutions per site between two sequences.
- three common distances used are the p-distance, the Jukes-Cantor distance and the Kimura 2-parameter distance.

1. p-distance

- this is simply the proportion (p) of nucleotide sites at which the two sequences being compared differ.

n = total no. of nucleotides compared
n_d = number of nucleotide differences

$p = n_d/n$
$var = [p(1-p)]/n$

- this measure is only appropriate when the number of nucleotide differences are small, say < 0.10.
- if the divergence is greater than this, the p-distance gives an underestimate of the true distances.
- why?
- because of homoplasy.

2. Jukes-Cantor distance

- this method assumes that the rate of substitution is the same for all pairs of the four nucleotides A, T, C, and G.
- it gives a maximum likelihood estimate of the number of nucleotide substitutions between two sequences.

$d = -3/4 \ln (1-4/3\ p)$
$var\ (d) = p(1-p)[(1 - 4/3\ p)^2 n$

- where p is as above.
- the Jukes-Cantor distance gives a good estimate of the number of substitutions if the all four bases are equally frequent, there is no transition/transversion bias and d is not very large (say < 0.10).

3. Kimura 2-parameter (K2P) distance

- this distance assumes that the rate of transitional nucleotide substitution is higher than the rate of transversional nucleotide substitution.
- let α be the rate of transitional substitution and β be the rate of transversional substitution.
- the following matrix summarizes this model.

		Mutant		
	A	T	C	G
A	--	β	β	α
T	β	--	α	β

C	β	α	--	β
G	α	β	β	--

- to estimate the K2P distance we first need to know the proportion of transitional and transversional differences between the two sequences compared.
- let P be the proportion of transitional differences and Q be the proportion of transversional differences.

$$P = n_s/n \quad \text{and} \quad Q = n_v/n$$

- where n_s and n_v are the numbers transitional and transversional differences between the two sequences.
- the K2P distance is then given by:

$$d = -\frac{1}{2} \ln (1 - 2P - Q) - \frac{1}{4} \ln (1 - 2Q).$$
$$\text{var}(d) = [c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2]/n$$

- where $c_1 = 1/(1 - 2P - Q)$, $c_2 = 1/(1 - 2Q)$, and $c_3 = 1/2 (c_1 + c_2)$

Tree Construction

- let us use the following data matrix and construct a UPGMA tree.

	Character								
Taxon	1	2	3	4	5	6	7	8	9
A	A	G	C	C	T	A	C	A	G
B	A	G	C	C	T	A	G	T	C
C	T	C	C	C	T	A	C	A	G
D	T	C	T	G	A	T	C	A	G

2. estimate distances among taxa and construct matrix of pair-wise distances

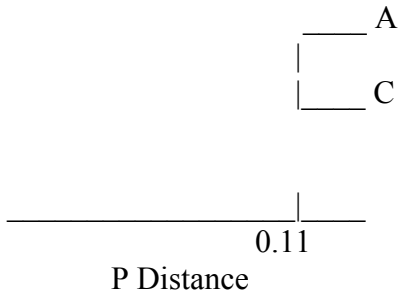
- let us use the simplest measure of distance, the “p distance”
- for taxa A and B, the p distance = 3/9 = 0.33.
- for A and C, the p distance = 2/9 = 0.22.
- we fill out the remaining entries.

	A	B	C	D
A	--	0.33	0.22	0.67
B		--	0.56	1.0
C			--	0.44
D				--

3. construct a tree based on the matrix - this is called a **phenogram.**

A. Join taxa with smallest distance

- here, we join taxa A and C



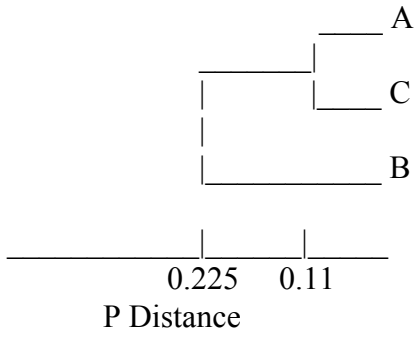
- the distance of the node separating taxa A and C is 0.11 (one half of 0.22).

B. Find taxon with lowest mean distance to (AC)

Mean distance
between (AC) and B = (0.33 + 0.56)/2 = 0.45

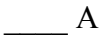
Mean distance
between (AC) and D = (0.67 + 0.44)/2 = 0.56

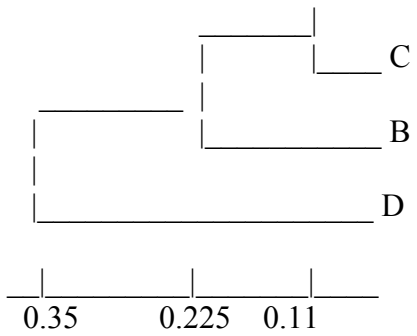
- thus we join taxon B to the (AC) group.



C. Add final taxon to the tree

Mean distance
between (ACB) and D = (0.67 + 1 + 0.44)/3
= 0.70





P Distance

- the branch lengths in phenograms carry important information about the degree of similarity between any pair of taxa.

- the closer two taxa resemble one another the higher they are positioned in the phenogram.
- phenograms may not represent the true phylogeny - in fact construction of the true phylogeny is not one of their goals.
- the branch lengths in phenograms carry information about the degree of similarity between any pair of taxa.
- an important assumption of UPGMA is that an equal rate of evolution occurs along all branches.
- although this assumption may be violated frequently, the UPGMA methods does quite well in simulation studies recovering true tree topographies.

- the principle behind NJ is to find neighbors sequentially that may minimize the total length of the tree.
- the total length of the tree is simply the sum of all branch lengths.
- an important advantage of the NJ approach over UPGMA is that it allows for unequal rates of evolution along different branches.

- distance methods are fast and efficient but all suffer from the same problem.
- this problem is that they are all based on an estimate derived from the data – they do not use the data itself to produce the tree.
- because information is lost in coverting the data into a distance matrix, distance methods fail to use all of the information present in the data.

Maximum parsimony (MP)

- according to the MP approach, the best tree is that which minimizes the number of evolutionary steps (i.e., changes among characters)
- this is the principle of parsimony - the least number of changes, required the better the tree.
- evolutionary change does not always obey laws of parsimony but it is a reasonable starting point.

- unlike distance methods that use information from all characters, MP trees are based exclusively on **synapomorphies.**
- synapomorphies are characters that are **shared** by two or more taxa that are **derived** (i.e., having changed) from some ancestral state.
- to establish whether a character is derived, it is essential to use one or more outgroup taxa that are hoped to possess the ancestral state of the character.
- an outgroup is ideally picked from fossil evidence - i.e., it belongs to a genus or family that existed prior to the **ingroup** upon which the phylogeny is based.

- in evaluating MP trees, a sizable problem faced is the large number of possible trees that need to be evaluated.
- the number of possible trees increases dramatically with the number of taxa:

No. of taxa	No. of possible trees
4	3
5	15
6	105
7	945
10	2×10^6
11	34×10^6
50	3×10^{74}

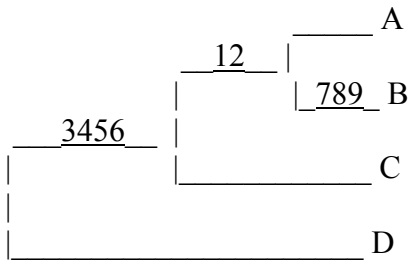
- how do we decide what is the best tree?
- two main approaches are used, both using sophisticated computer programs to evaluate alternative trees.

Types of parsimony

- a number of different types of parsimony have been described.
- in **Fitch** parsimony, one assumes that all types of changes among characters are equally likely and free reversibility is allowed
- **Dollo** parsimony can only be used when characters are coded as present or absent.
- it differs from the Wagner/Fitch parsimony in assuming that all nonancestral characters are uniquely derived.
- in other words, a character can change from being present to absent once, and only once, in the entire tree topology.
- this form of parsimony might appear to be unreasonably restrictive but it is the preferred type when one works with restriction site data.
- **Camin-Sokal** parsimony carries the stringent requirement that all evolutionary change is irreversible.
- finally, **generalized** parsimony allows for flexibility in assigning the "costs" of transformation among character states.
- to do this, one needs independent evidence about the relative frequencies of the different types of changes.

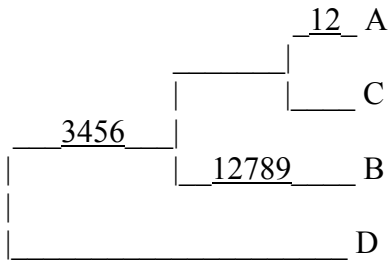
- let us use the parsimony principle to select the best tree.
- let us assume that taxon D is the outgroup.
- for the remaining three taxa, there are only three possible trees to evaluate.

Tree 1:



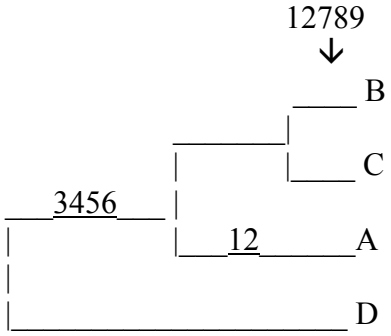
- this tree involves 9 character step changes (or steps).

Tree 2:



- this tree involves 11 character step changes (or steps).

Tree 3:



- this tree also involves 11 character step changes (or steps).
- therefore, tree 1 is the most parsimonious and is thus preferred over trees 2 or 3.

Maximium likelihood

- given a certain model of base substitution and a specific tree, what is the probability of obtaining this set of DNA sequences?
- this probability is estimated by a tree's likelihood score.
- the best tree is that which has the highest likelihood, or probability of being produced.

$$L = Pr(\text{data} \mid \text{hypothesis})$$

- here, L is the likelihood (probability) of obtaining the data given a substitution model (with certain parameter values) and a specific tree.
- the objective is to find the tree that maximizes L.
- note that unlike MP, we need to specify an explicit model of substitution.
- ML methods are computationally intensive and thus have not been easy to use until recently

Evaluating trees

- ideally, one would want to compare the trees produced by different methods.
- if the same topology is found using distance, MP, and ML methods, one can be reasonably sure that the tree is optimal.
- new methods are being developed (for example log-likelihood ratio tests) for testing whether or not two trees differ significantly from each other.
- another more common approach is to use a statistical technique called bootstrapping.
- bootstrapping allows us to evaluate the degree of support for branches in a tree.

- the bootstrapping procedure works by randomly re-sampling the nucleotide sequence data (with replacement), constructing a tree from this data and counting the number of times a particular branch is found out of say, 100, 500, or 1,000 replicate pseudosamples.
- consider an example of 300 bases.
- the bootstrapping process begins by randomly sampling one site and assigning it as the first entry in the new dataset.
- it then randomly selects another site which becomes the second data point in the new dataset (there is a 1/300 chance that this is the same as the first site).
- resampling continues until 300 bases of data are obtained.
- a tree is made form the data and the procedure is repeated, say, 100 times.
- the result is a measure of bootstrap support for each branch.
- bootstrap support values of 70% or higher are usually taken as strong support for that particular branch.
- values approaching 50% should be viewed with concern.

- for distance methods, various techniques are available to test the reliability of a given tree in addition to bootstrapping.
- for example, one can test whether internodal distances are significantly greater than 0.
- if all such distances are significant, then the inferred tree is judged significant.
- another approach is to use a minimum evolution criteria.
- here, if the total length of an observed tree is significantly shorter than every other alternative tree, then it is considered significant.

Comparison of methods

- the performance of different methods of tree construction can be tested in two ways.
- the first is by empirical studies in which the evolutionary history is known.
- for example, Atchley and Fitch (1991) used genetic data to test whether phylogenetic methods could correctly recover the known relations among inbred strains of laboratory mice.
- Hillis et al (1992) mutagenized bacteriophage T7 through a large number of generations and tested the accuracy of methods to reconstruct a known history.
- from the methods considered today, UPGMA may be thought to be the worst of the group in assuming that rates of evolution are equal along all branches.
- surprisingly, it performs rather well even if this assumption is violated.
- NJ is an improvement, but is prone to considerable error if the distances are small, or if there is considerable variation in rates of evolution among sites.
- MP makes no explicit assumptions (it has a number of implicit assumptions) and generally performs well when sequence divergence is low.
- with larger distances (and more homoplasy), it does not do well.
- furthermore, if some sequences have evolved faster than others (and homoplastic events occurred more often in these sequences) then the MP criterion can be misleading.
- one way out of this of problem is to assign different weights for different characters - typically giving transversions much more weight than transitions.
- ML methods are the most flexible and potentially accurate.
- their biggest drawback is that they are computationally intensive.