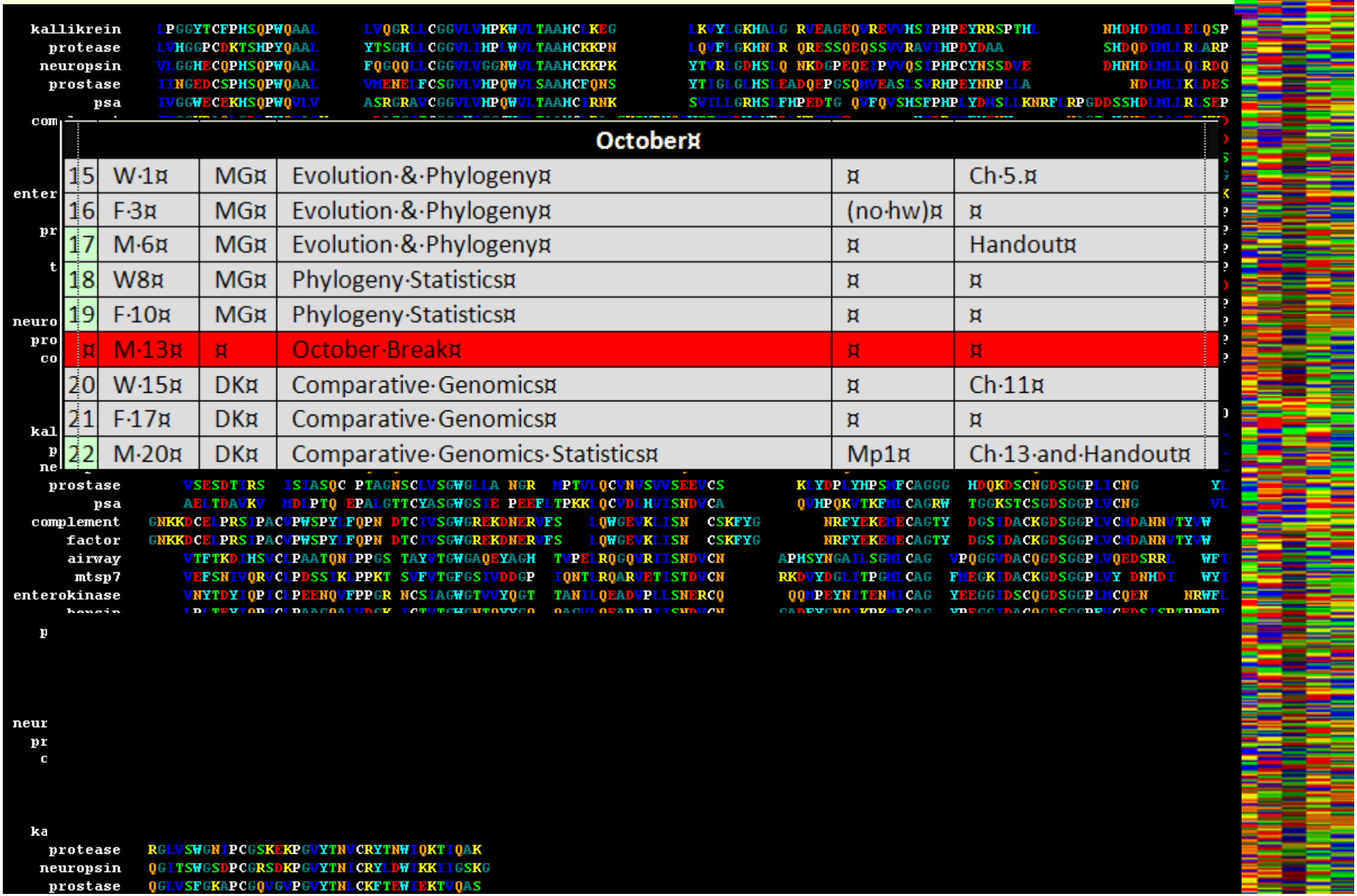
Biol 478/595 Intro to Bioinformatics

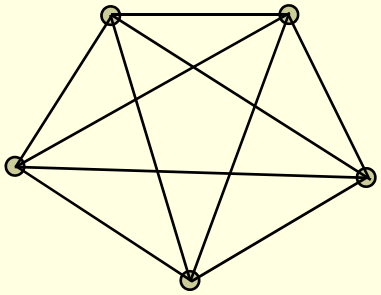


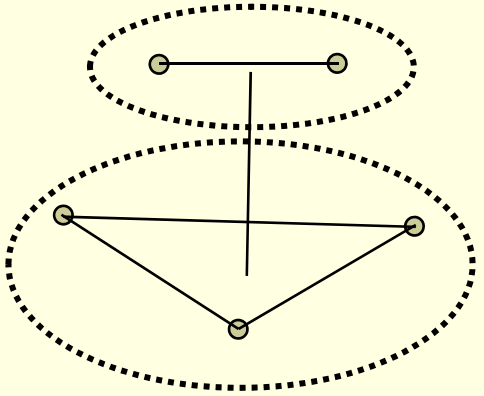
October¤					
15	W·1¤	MG¤	Evolution·&·Phylogeny¤	¤	Ch·5.¤
16	F·3¤	MG¤	Evolution·&·Phylogeny¤	(no·hw)¤	¤
17	M·6¤	MG¤	Evolution·&·Phylogeny¤	¤	Handout¤
18	W8¤	MG¤	Phylogeny·Statistics¤	¤	¤
19	F·10¤	MG¤	Phylogeny·Statistics¤	¤	¤
¤	M·13¤	¤	October·Break¤	¤	¤
20	W·15¤	DK¤	Comparative·Genomics¤	¤	Ch·11¤
21	F·17¤	DK¤	Comparative·Genomics¤	¤	¤
22	M·20¤	DK¤	Comparative·Genomics·Statistics¤	Mp1¤	Ch·13·and·Handout¤

Multiple Alignment and Trees

Distance Methods – Neighbor Joining

- **Consider a group of taxa and all of the distances between them.**
- **We want to find the two taxa that are closer to each other than to anyone else, I.e., two split off two taxa so that the the two groups are both as compact as possible**
- **Use all distances as in Fitch-Margoliash**

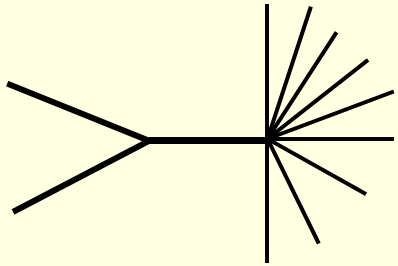




Multiple Alignment and Trees

Distance methods - Neighbor Joining

- **Find pair of sequences i,j that minimize S**
  $S_{ij} = D_{ij}/2 + [2Q - R_i - R_j] / 2(n-2)$
  where
  $Q = \Sigma_{ij} D_{ij}$ $R_j = \Sigma_i D_{ij}$ $R_i = \Sigma_j D_{ij}$
- **Replace distances in matrix by average values**
- **Iterate as in UPGMA, finding best pair to link at each stage until all are linked.**
- **Determine branch lengths by Fitch-Margoliash procedure**

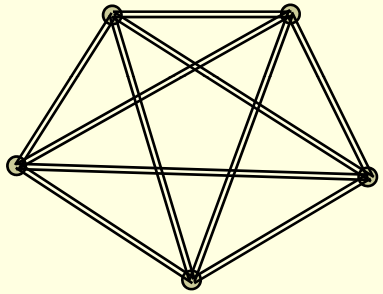


Multiple Alignment and Trees

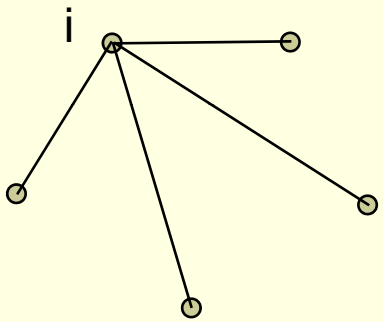
Neighbor Joining

$S_{ij} = D_{ij}/2 + [2Q - R_i - R_j] / 2(n-2)$
where
$Q = \Sigma_{ij} D_{ij}$ $R_j = \Sigma_i D_{ij}$ $R_i = \Sigma_j D_{ij}$

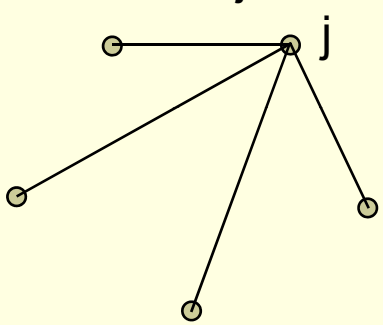
2Q



R_i

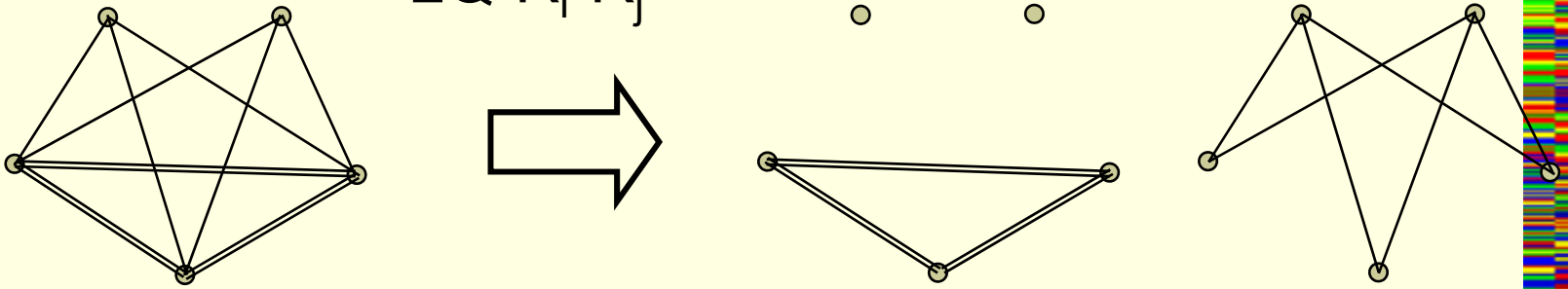


R_j



2(n-2) distances from ij to non-ij

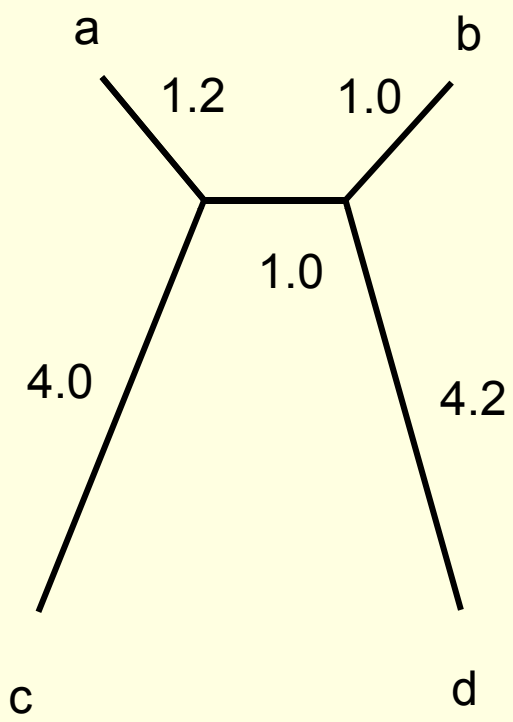
$2Q-R_i-R_j$



2(n-2) distances in non ij group

Multiple Alignment and Trees

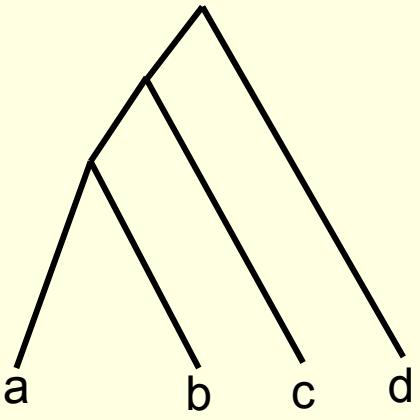
Long Branches Attract



Distance Matrix

(perfect agreement with tree)

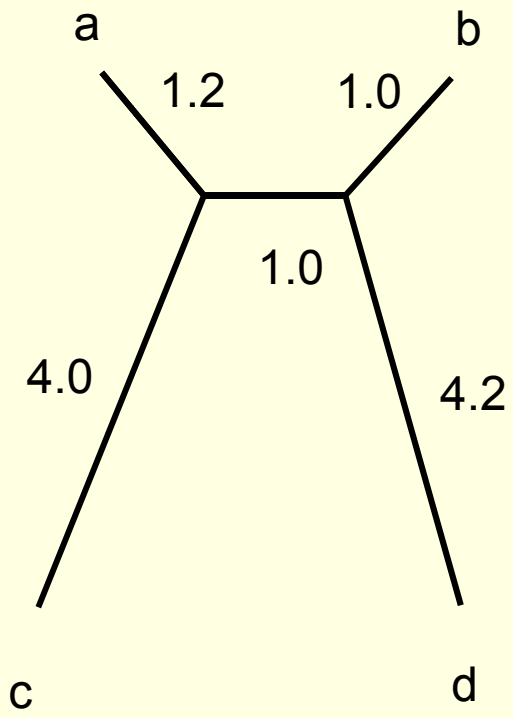
	a	b	c	d
a	-	3.2	5.2	6.4
b		-	6.0	5.2
c			-	9.2
d				-



UPGMA Tree

Multiple Alignment and Trees

Neighbor Joining and long branches



Distance Matrix

	a	b	c	d
a	-	3.2	5.2	6.4
b		-	6.0	5.2
c			-	9.2
d				-

Neighbor joining:

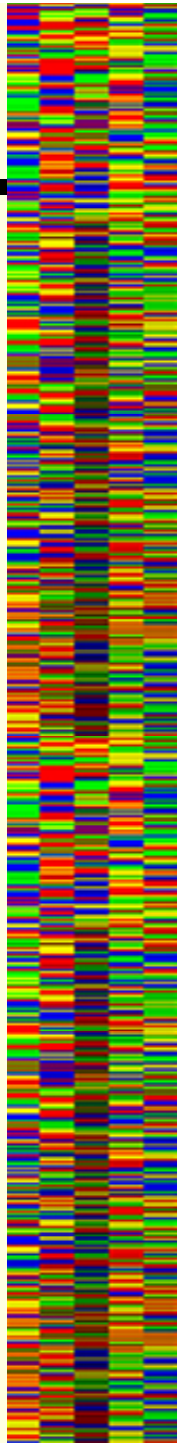
$Q = 35.2$ $R_a = 14.8$ $R_b = 14.4$
$R_c = 20.4$ $R_d = 20.8$

$S_{ab} = 3.2/2 + (70.4-14.8-14.4)/4 = 14.5$
$S_{ac} = 5.2/2 + (70.4 -14.8-20.4)/4 = 11.5$
$S_{ad} = 6.4/2 + (70.4 -14.8-20.8)/4 = 11.7$
$S_{bc} = 6/2 + (70.4 -14.4-20.4)/4 = 11.9$
$S_{bd} = 5.2/2 + (70.4 -14.4-20.8)/4 = 11.4$
$S_{cd} = 9.2/2 + (70.4 -20.4-20.8)/4 = 11.9$

Multiple Alignment and Trees

Parsimony Methods

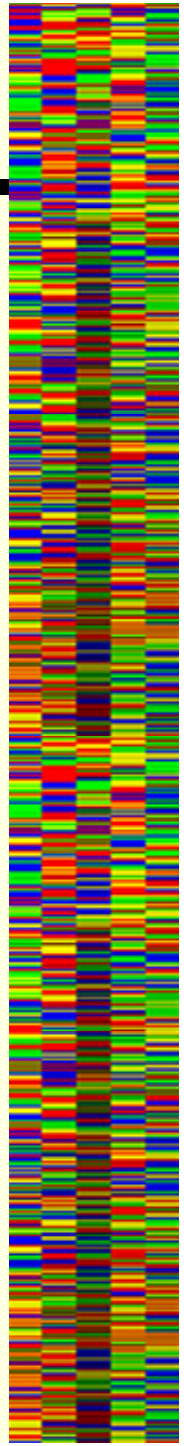
- **Based on minimum evolution principal - the tree that would require the fewest inferred mutations is the "correct" one**
  - **Sometimes called a version of "Occam's razor" – the simplest explanation should be accepted as correct**
- **Does not provide a method to construct the tree topology, only a principle for deciding if a topology is best**
- **Looks at individual residue or base positions**
- **Only some positions are "informative"**



Multiple Alignment and Trees

Parsimony Methods

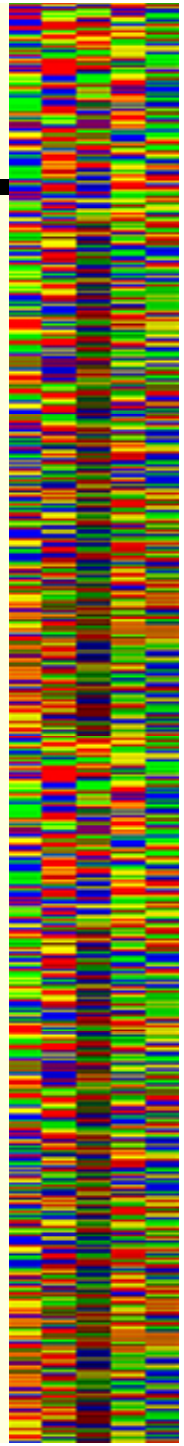
- **The best tree has the fewest mutations**
- **Fitch method - any state (*e.g.*, amino acid residue) can convert to any other**
- **Works out the actual sequence of the ancestral organisms, as well as minimum number of changes.**
- **The set of residues (or bases) possible for the ancestor is the "state set"**
  - **ifa node has two descendents with sequences A and G respectively, the state set is [AG] (can't tell which is ancestral)**
- **Two stages**
  - **First traverse the tree from leaves to root to determine state sets of ancestral nodes**
  - **Second, traverse from root to leaves to assign ancestral states**



Multiple Alignment and Trees

Parsimony methods

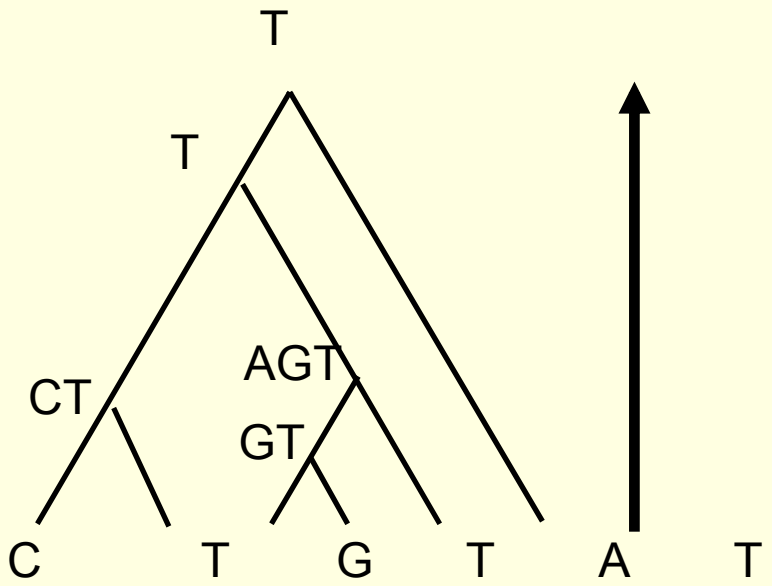
- **Compare ancestral node to descendent nodes of a tree**
  - **If the intersection of the residues seen in the descendent nodes is not empty, the ancestral state set is the intersection and the tree length remains the same**
  - **If the intersection of the residues seen in the descendent nodes is empty, the ancestral state set is the union of the descendent state sets and the tree length is incremented by 1**
  - **If the state of the root is not included in the next most distal internal node, the tree length is incremented by 1**
  - **When assigning final states, if a nodes state set includes the state of its ancestor, it is assigned that state. Otherwise you pick arbitrarily.**



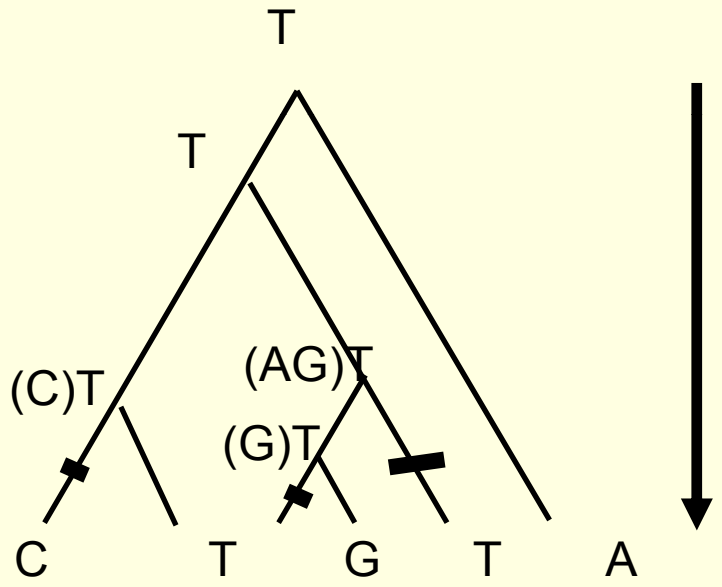
Multiple Alignment and Trees

Parsimony Methods

Step 1: Ancestral state sets



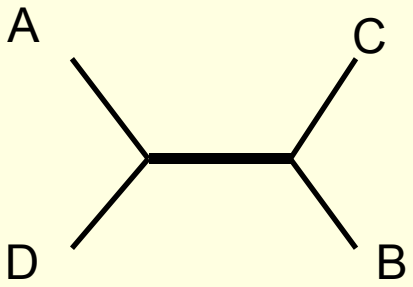
Step 2: Assigned ancestral sequence

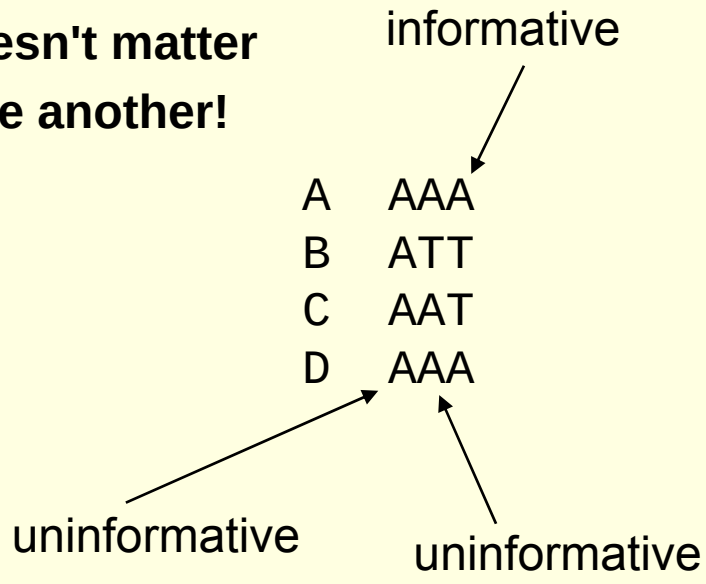


Multiple Alignment and Trees

Parsimony Methods

- **Not all sites are "informative" in parsimony, i.e. they don't let you discriminate between two different tree topologies**
- **To be informative, a position must have**
  - **at least two kinds of bases/residues**
  - **each occurring at least two times.**
  - **any base/residue occuring in only one sequence is NOT informative,because it is compatible with any tree.**
- **Order of positions in the sequence doesn't matter**
- **Some positions support one tree, some another!**

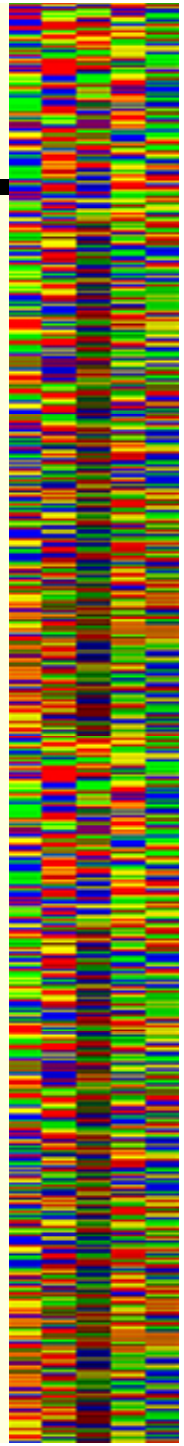




Multiple Alignment and Trees

Parsimony Methods

- Since parsimony methods don't define the tree topology, you must search for the best tree.
- Exhaustive search is only possible for small numbers of sequences
- For larger numbers of sequences, a branch and bound method can be used.
- In any case one must start with a "good" tree, probably a UPGMA or neighbor joining tree.
- Can also optimize the tree using heuristic methods
  - Random topology
  - Branch swapping
- Use PAUP (Swofford)
- Parsimony methods are prone to problems when there are multiple substitutions because you cannot then see all of the mutational changes



Multiple Alignment and Trees

Observable vs. Hidden Evolution in Nucleotide Sequences (Mount, p. 268)

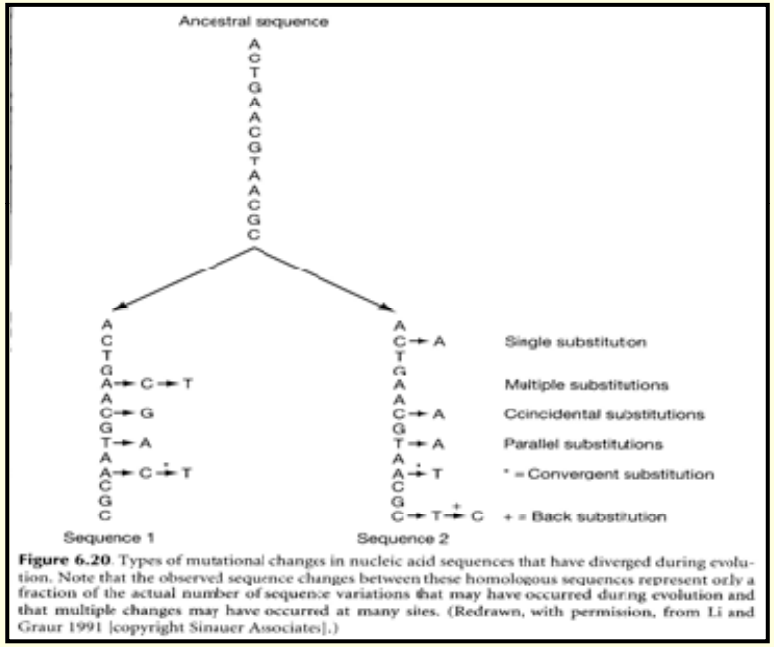
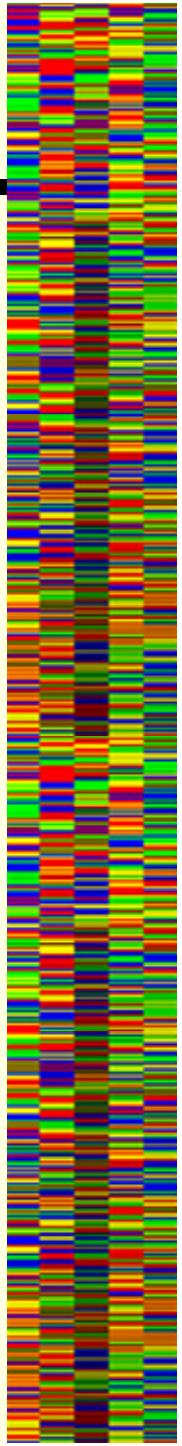


Figure 6.20. Types of mutational changes in nucleic acid sequences that have diverged during evolution. Note that the observed sequence changes between these homologous sequences represent only a fraction of the actual number of sequence variations that may have occurred during evolution and that multiple changes may have occurred at many sites. (Redrawn, with permission, from Li and Graur 1991 [copyright Sinauer Associates].)

Multiple Alignment and Trees

Confidence

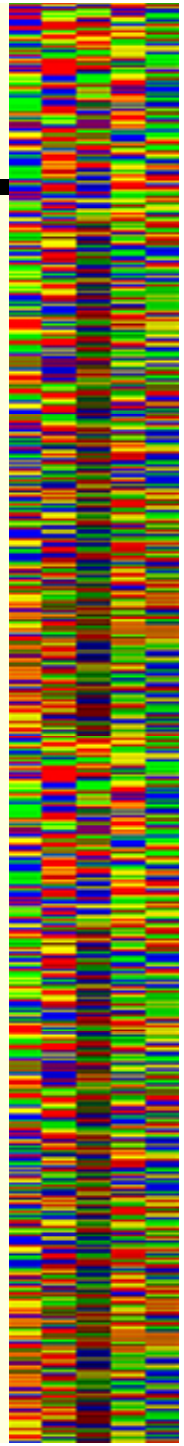
- How sure can you be that your tree is a "good" tree?
- Bootstrap Method
- Remember that aligned positions are essentially independent from the point of view of trees.
- Choose subset of positions randomly and recalculate tree
  - Repeat perhaps 1000 times
  - How many times do you see the same branching pattern?
- Gives confidence that the data indicates the tree - NOT confidence that the tree is correct



Multiple Alignment and Trees

Considerations

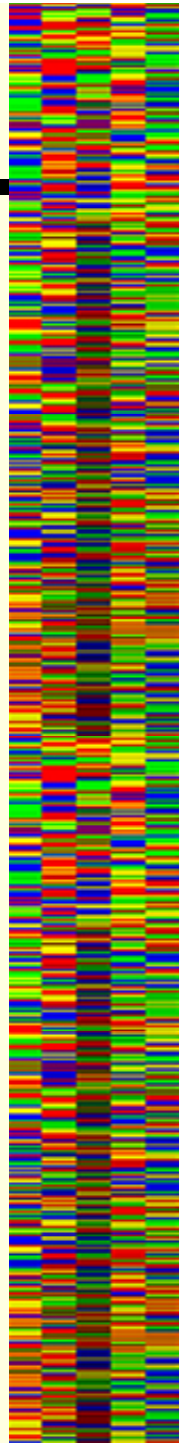
- **Is your alignment correct?**
- **Should you throw out some positions because you are unsure of the alignment?**
- **Rapidly changing positions give more information about leaf nodes**
- **Slowly changing positions give more information about internal nodes**
- **What is your criterion of "goodness"?**
- **Is the tree you have really the best?**
  - **All greedy methods are likely to be suboptimal sometimes**



Multiple Alignment and Trees

Maximum Likelihood

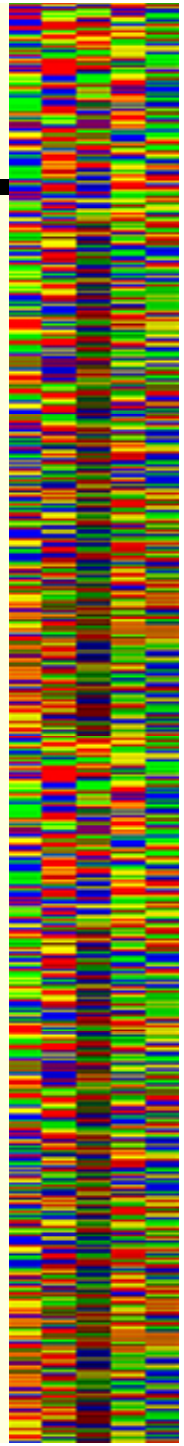
- **Does not generate topology**
- **Estimates the probability of the observed sequences given the phylogenetic model (i.e. the topology and branch lengths)**
- **Most common method is Felsenstein's as implemented in the PHYLIP package**
- **Very time consuming because of the many calculations over many tree topologies**



Multiple Alignment and Trees

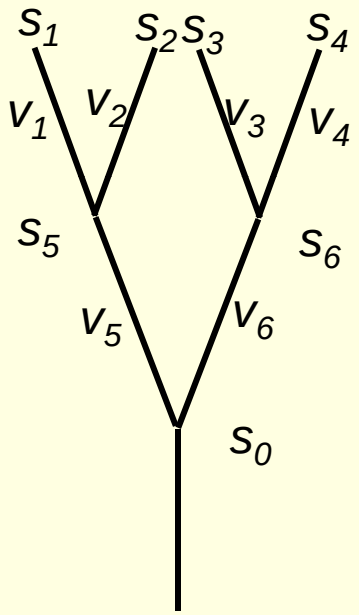
Maximum Likelihood Methods

- **Estimate topology and branch length viewing evolution as a random process**
- **Requires a probability model of evolution as a function of time.**
  - **For DNA one can use Jukes-Cantor model (all nucleotides have same substitution rates), or Kimura model (different rates for transitions, R→R or Y →Y, and transversion, R→Y or Y →R).**
  - **For proteins one can use Dayhoff, but in the probability form not the log-odds form.**



Multiple Alignment and Trees

Maximum Likelihood Methods



S_1 etc are the bases or residues observed in the extant and ancestral taxa

$v = \lambda t$ where λ is the substitution rate and t is absolute time

$P_{i,j}(v)$ is the probability that the residue at node s_i becomes residue at node s_j in time v

g_0 is the prior probability of the bases or nucleotides at any position

$$L = g_0 P_{0,5}(v_5)\, P_{5,1}(v_1)\, P_{5,2}(v_2)\, P_{0,6}(v_6)\, P_{6,3}(v_3)\, P_{6,4}(v_4)$$

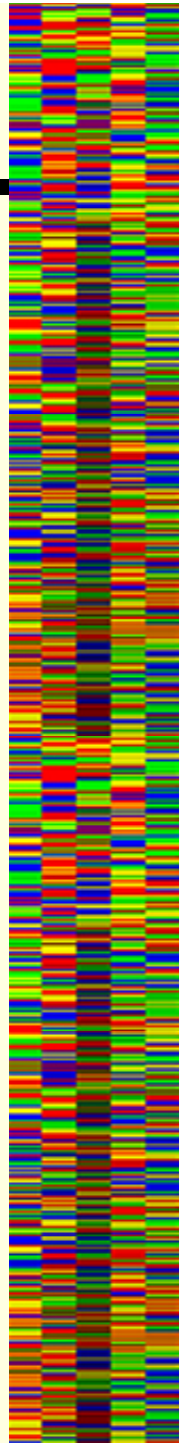
Multiple Alignment and Trees

Maximum Likelihood

- $L = g_0 P_{0,5}(v_5)\, P_{5,1}(v_1)\, P_{5,2}(v_2)\, P_{0,6}(v_6)\, P_{6,3}(v_3)\, P_{6,4}(v_4)$
- **We don't know the identity of the bases or residues at s5, s6, or s0 so the likelihood must be summed over all possibilities:**

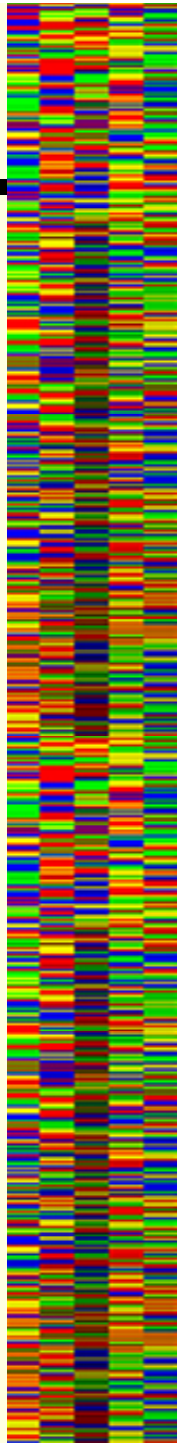
$$L = \sum_{s_0} \sum_{s_5} \sum_{s_6} g_0 P_{0,5}(v_5)\, P_{5,1}(v_1)\, P_{5,2}(v_2)\, P_{0,6}(v_6)\, P_{6,3}(v_3)\, P_{6,4}(v_4)$$

- **Felsenstein suggests a simple model:**
  $P_{i,j} = e^{-v\delta i,j} + (1-e^{-v})g_j$ where $\delta_{i,j}$ is 1 when i=j and 0 otherwise
- **Must vary all v to maximize likelihood and the try numerous topologies to find the highest likelihood tree**



Maximum Likelihood

- **Problems**
  - **The value of *v* varies with position due to conserved and unconserved regions**
  - **The probability of a substitution $P_{i.i}$ is also position specific**



Multiple Alignment and Trees

Distance methods (UPGMA & N-J)

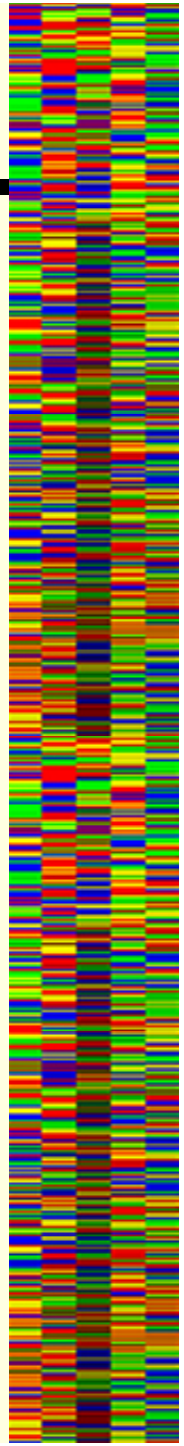
Optimality criterion: NONE. The algorithm itself builds 'the' tree.

Advantages:

- **Can be used on indirectly-measured distances (immunological, hybridization).**
- **Distances can be 'corrected' for unseen events.**
- **The fastest of the methods available (N-J is screamingly fast!).**
- **Can therefore analyze very large datasets quickly (needed for HIV, etc.).**
- **Can be used for some types of rate and date analysis.**

Disadvantages:

- **Similarity and relationship are not necessarily the same thing**
- **clustering by similarity does not necessarily give an evolutionary tree.**
- **cannot be used for character analysis!**
- **Have no explicit optimization criteria, so one cannot even know if the program worked properly to find the correct tree for the method.**



Multiple Alignment and Trees

Parsimony methods

Optimality criterion

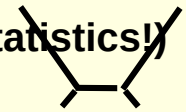
- best tree is the one that requires the fewest number of evolutionary events (e.g., nucleotide substitutions, amino acid replacements) to explain the sequences.

Advantages:

- Are simple, intuitive, and logical (many possible by 'pencil-and-paper').
- Can be used on molecular and non-molecular (e.g., morphological) data.
- Can tease apart types of similarity (shared-derived, shared-ancestral, homoplasy)
- Can be used for character (can infer the exact substitutions) and rate analysis.
- Can be used to infer the sequences of the extinct (hypothetical) ancestors.

Disadvantages:

- Are simple, intuitive, and logical (derived from "Medieval logic", not statistics!)
- Can be fooled by high levels of homoplasy ('same' events).
- Can become positively misleading in the "Felsenstein Zone":



Multiple Alignment and Trees

Maximum likelihood (ML) methods

Optimality criterion

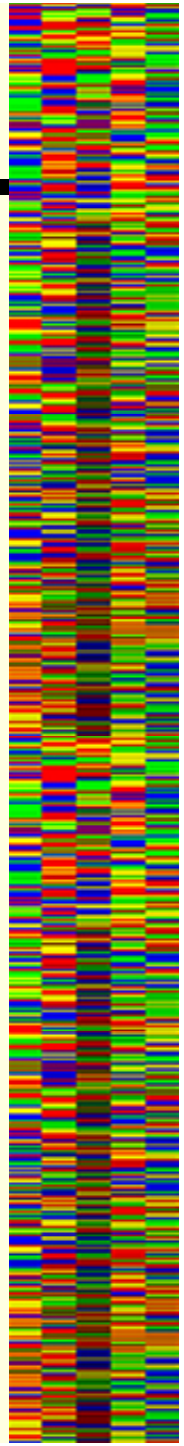
- **probability that a proposed model of the evolutionary process and the proposed unrooted tree would give rise to the observed data.**
- **The tree found to have the highest ML value is considered to be the preferred tree.**

Advantages:

- **Inherently statistical and evolutionary model-based.**
- **Usually the most 'consistent' of the methods available.**
- **Can be used for character (can infer the exact substitutions) and rate analysis.**
- **Can be used to infer the sequences of the extinct (hypothetical) ancestors.**
- **Can help account for branch-length effects in unbalanced trees.**
- **Can be applied to nucleotide or amino acid sequences, and other types of data.**

Disadvantages:

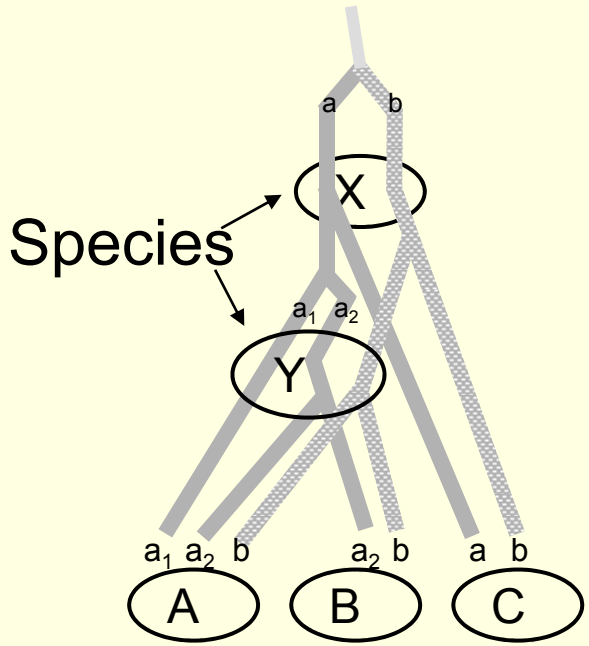
- **Not as simple and intuitive as many other methods.**
- **Computationally very intense (limits number of taxa and length of sequence).**
- **Like parsimony, can be fooled by high levels of homoplasy.**
- **Violations of the assumed model can lead to incorrect trees.**



Multiple Alignment and Trees

Gene Trees vs Species Trees

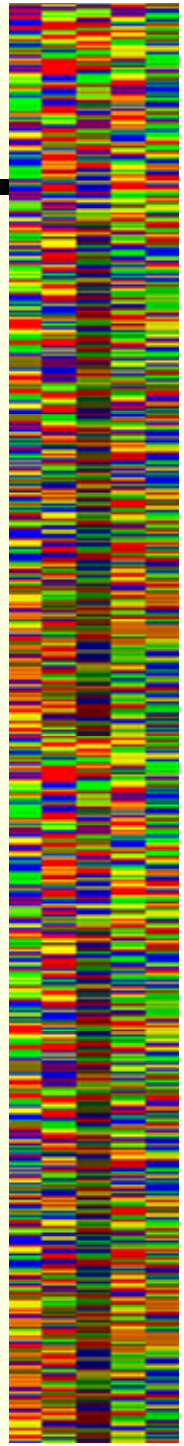
- **Genes usually diverge before species, *i.e.*, gene trees overestimate distance to common ancestor**
- **In speciation, two species would be expected to get different populations of alleles**
- **Alleles are copies of genes that are already diverging in a species**
- **species X has a and b**
- **species Y has a_1, a_2, and b**
- **if some alleles are lost, A or B may not have all three**



Multiple Alignment and Trees

Orthology and Paralogy

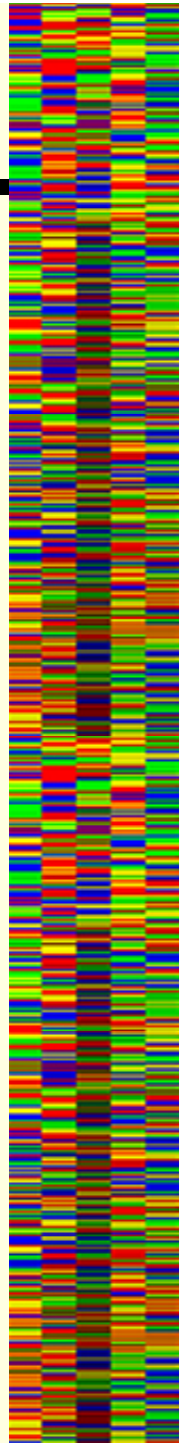
- **Tree construction seeks to understand the evolutionary relationship between taxa (groups of organisms).**
- **In sequence based methods, one must distinguish between gene trees and species trees due to the presence of orthologous and paralogous genes**
  - **Orthologous - homologous genes that are truly the same, e.g. myoglobin in sharks and humans (arise from speciation).**
  - **Paralogous - homologous genes that resulted from a gene duplication, e.g. hemoglobin and myoglobin.**
  - **Xenologous - horizontally transferred genes**
- **These concepts are often used, perhaps overused, in making inferences about gene function – i.e., that orthologous genes have the same function and paralogous genes have different functions. This is an oversimplification at best due to continuous creation of paralogous genes by duplication and stochastic loss of genes by deletion.**



Multiple Alignment and Trees

Rooting trees

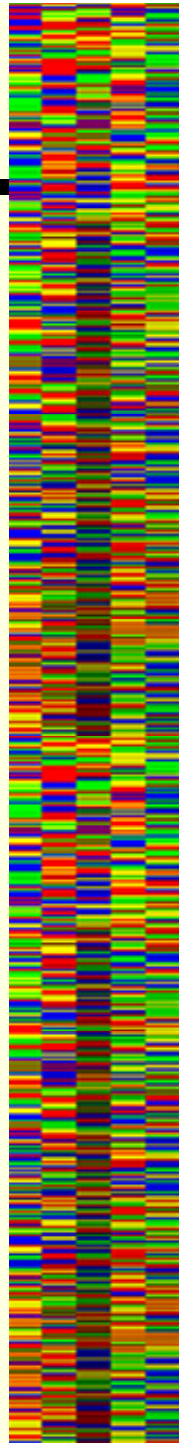
- **Rooting the tree lets you unambiguously decide what is ancestral and what is derived**
- **Trees are implicitly unrooted. That is, you can't tell from the data used to construct the tree where the ancestral node lies. You must have additional data to find the root of a tree.**
- **Most common procedure is to use an outgroup, i.e. a taxon that is guaranteed to be more distant from all of the taxa of interest than any of them are from each other.**
  - **Orangutan can be used as outgroup for human, chimp, gorilla**
  - **Alligator can be used as outgroup for human, rat, dog, cow, horse**
- **Midpoint between farthest pair can be used as root (assumes a clock)**



Multiple Alignment and Trees

Dawkin's Parable – Why are there trees?

- **"Methinks it is like a weasel" (target phrase = 28 bases)**
  - **28 chars. long with 27 options per char. (letter or space)**
- **Random Genetic Drift (Monkey at Keyboard) expectations:**
  - **$(1/27)^{28}$ = avg. 1040 generations**
- **Tree process: expectations**
  - **1. generate random sequence of 28 chars**
  - **2. replicate (1000) with random error (= breeding with mutation)**
  - **3. screen "progeny" for most accurate and keep them (= selection of**
  - **most "fit")**
  - **4. repeat steps 2-3 until the phrase is correct (= adaptation)**
- **Results (3 independent runs): 43 / 64 / 41 generations !!**
- **Natural selection is the filter of variation, a potent process that produces change in much less than random time – this strong filter is seen in biological trees**

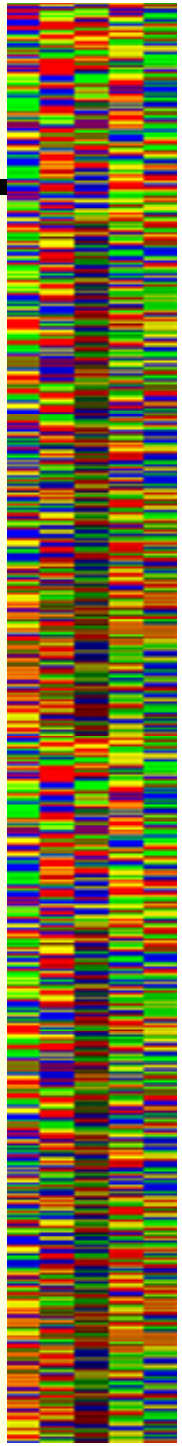


Multiple Alignment and Trees

Clocks

Some trees assume or try to identify a clock – when did these species diverge?

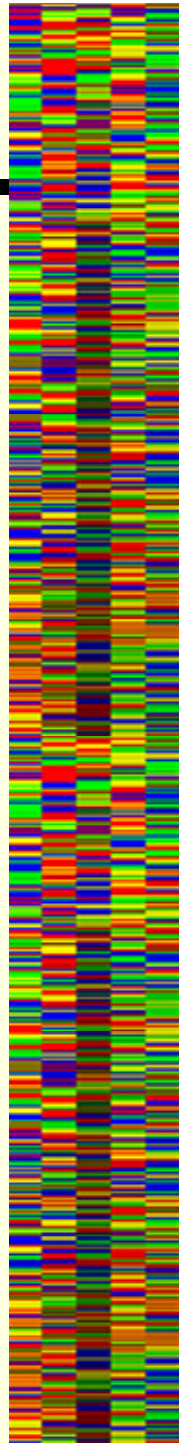
- **Most of our clocks are deterministic, they make "ticks" at precise intervals**
- **A stochastic clock is probabilistic, it makes ticks at a certain probability in each unit of time. Ticks may not be evenly spaced.**
- **Stochastic clocks are not necessarily inaccurate – atomic clocks are stochastic clocks, however for them to be accurate you must have a single process underlying the "ticking"**



Multiple Alignment and Trees

Neutral Mutations

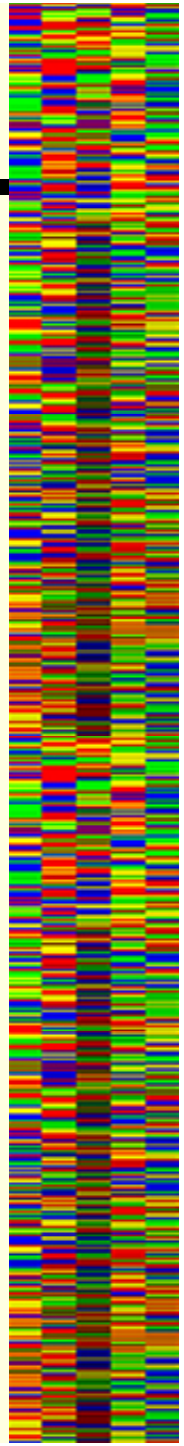
- **Neutral mutations - Most mutations are neither highly advantageous or deleterious - they are effectively neutral (Kimura theory).**
- **Neutral mutations should be the most clocklike because the represent the random accumulation of changes over time.**
- **Correcting distances - distances are often corrected for multiple mutational events so that they have a linear relationship to time.**
- **One of the reasons behind the original formulation of the Dayhoff mutational distance matrix was to provide a mapping from amino acid residue changes to a clock – 1 PAM is a time unit in this sense.**



Multiple Alignment and Trees

Take home messages

- **The number of tree topologies grows factorially with the number of taxa - it is generally impossible to examine all tree topologies**
- **Trees based on molecular sequences are much more straightforward to calculate, and more reliable than those based on morphological characters**
- **Calculating real divergence times and accurate branch lengths depends on mutations acting like a molecular clock. In turn, the molecular clock assumption is only appropriate when looking at neutral mutations (Kimura's hypothesis)**
- **Often analyses will focus on apparently neutral differences such as synonomous codon changes or third position of codon changes in order to get the most clock-like data**



Statistical Tests Comparing Trees

Tests of one overall hypothesis (tree) against other hypotheses

- Wilson's "winning sites" test
- Templeton's test
- Kishino-Hasegawa ML test

Tests of strength of support for lineages within trees:

- Bootstrap
- Jack-knife
- Decay index

These are implemented for numerous phylogenetic methods in *PAUP**.