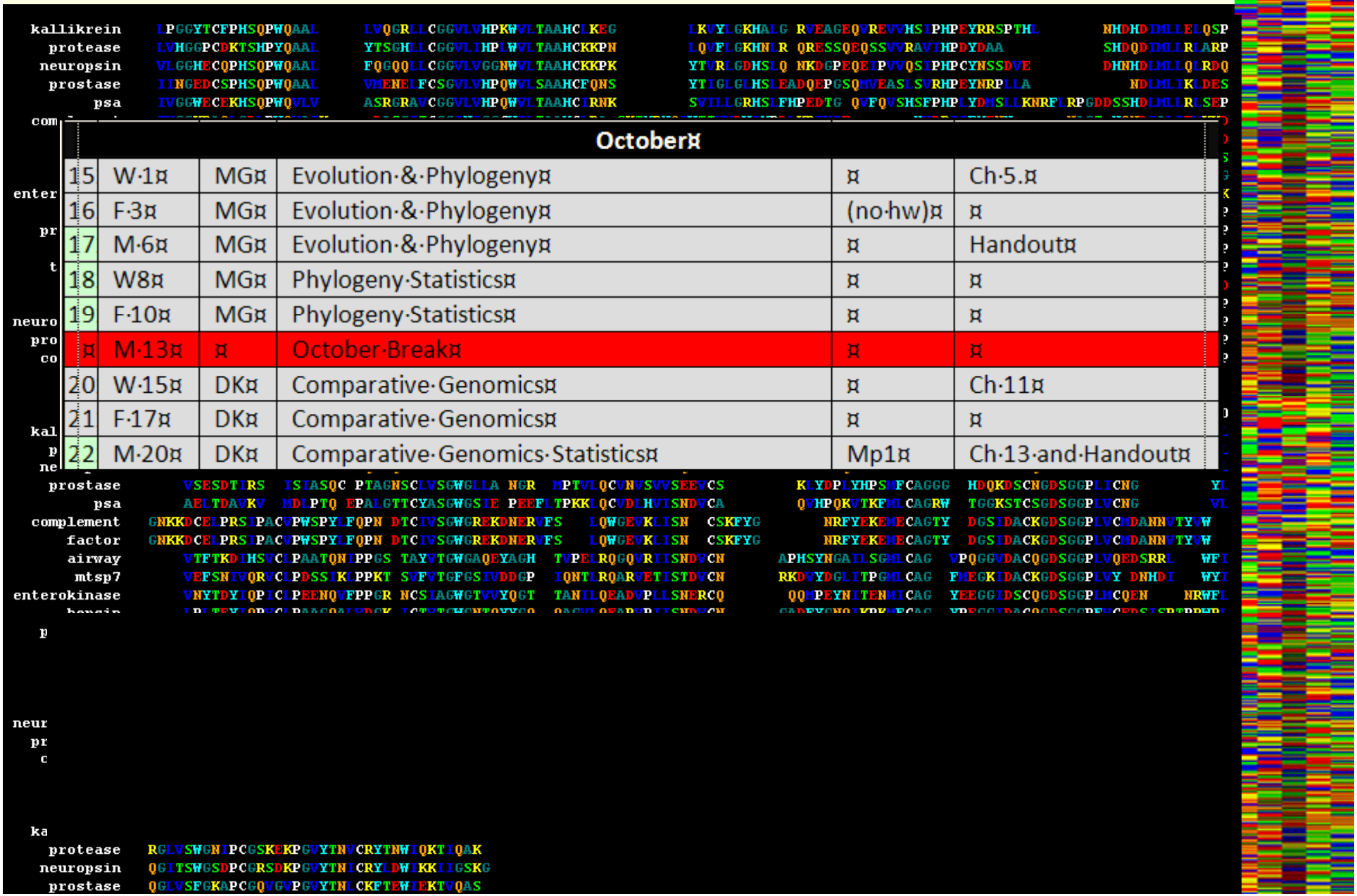
Biol 478/595 Intro to Bioinformatics

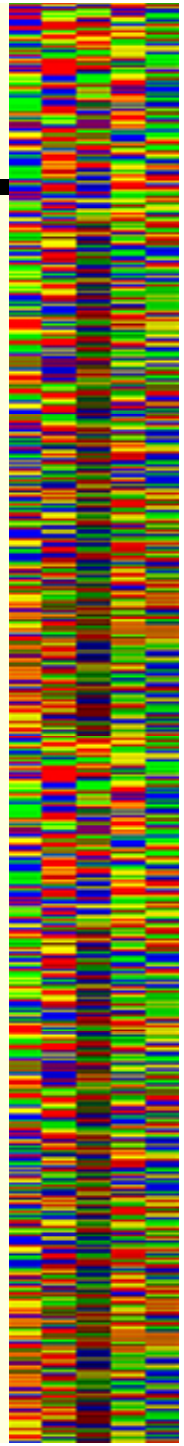


October¤					
15	W·1¤	MG¤	Evolution·&·Phylogeny¤	¤	Ch·5.¤
16	F·3¤	MG¤	Evolution·&·Phylogeny¤	(no·hw)¤	¤
17	M·6¤	MG¤	Evolution·&·Phylogeny¤	¤	Handout¤
18	W8¤	MG¤	Phylogeny·Statistics¤	¤	¤
19	F·10¤	MG¤	Phylogeny·Statistics¤	¤	¤
¤	M·13¤	¤	October·Break¤	¤	¤
20	W·15¤	DK¤	Comparative·Genomics¤	¤	Ch·11¤
21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

Multiple Alignment and Trees

Rooting trees

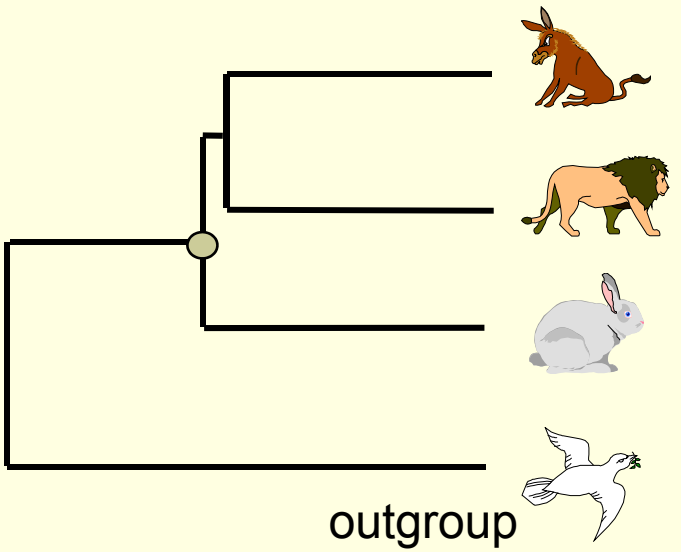
- **Rooting the tree lets you unambiguously decide what is ancestral and what is derived**
- **Trees are implicitly unrooted. That is, you can't tell from the data used to construct the tree where the ancestral node lies. You must have additional data to find the root of a tree.**
- **Most common procedure 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.**
  - **for best result, the outgroup should be as close as possible to the taxa of interest while still clearly outside them**
  - **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)**



There are two major ways to root trees:

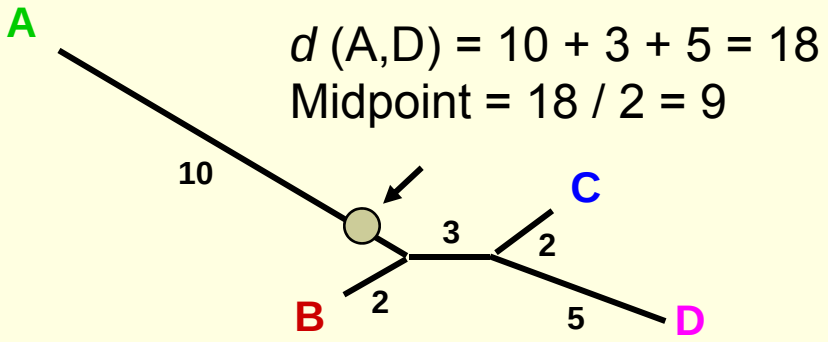
By outgroup:

Uses taxa (the "outgroup") that are known to fall outside of the group of interest (the "ingroup"). Requires some prior knowledge about the relationships among the taxa. The outgroup can either be species (*e.g.*, birds to root a mammalian tree) or previous gene duplicates (*e.g.*, α-globins to root β-globins).



By midpoint or distance:

Roots the tree at the midway point between the two most distant taxa in the tree, as determined by branch lengths. Assumes that the taxa are evolving in a clock-like manner. This assumption is built into some of the distance-based tree building methods.

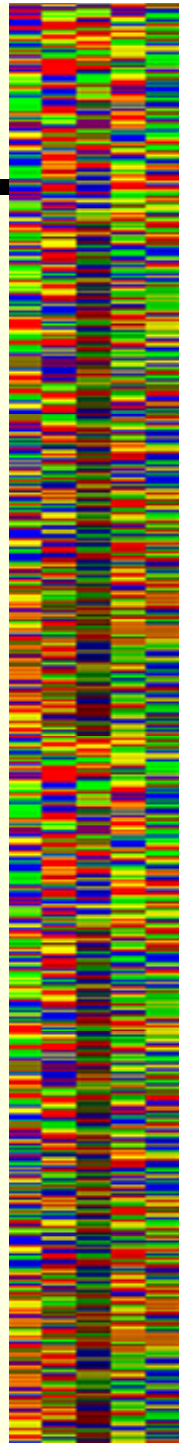


d (A,D) = 10 + 3 + 5 = 18
Midpoint = 18 / 2 = 9

Multiple Alignment and Trees

Take home messages

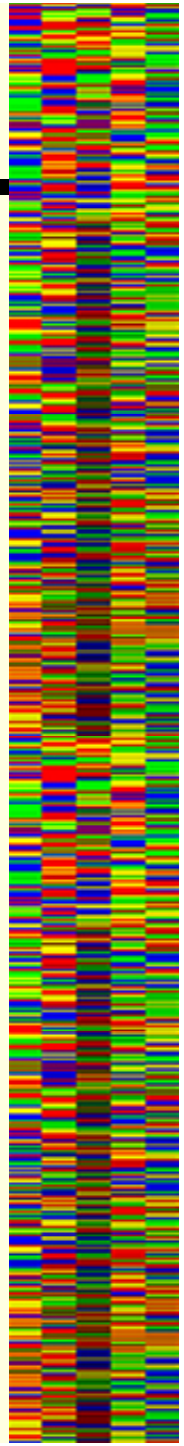
- **The number of tree topologies grows factorially with the number of taxa - it is generally impossible to examine all tree topologies**
- **Trees based on molecular sequences are much more straightforward to calculate, and more reliable than those based on morphological characters**
- **Calculating 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**



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**

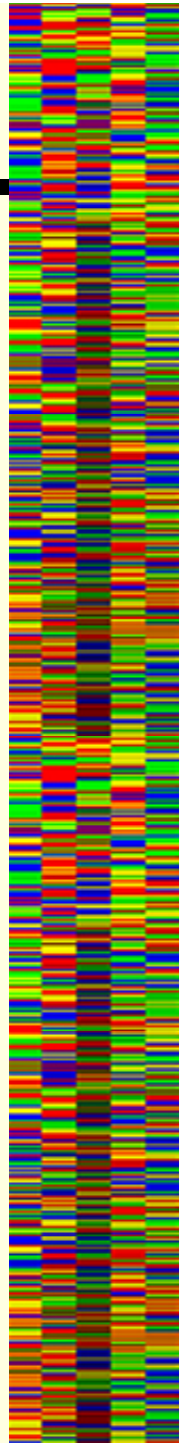


Multiple Alignment and Trees

Clocks

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

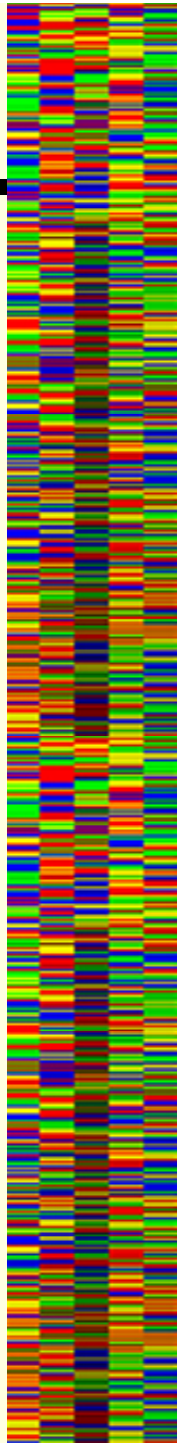
- **Most of our clocks are deterministic, they make "ticks" at precise intervals**
- **A stochastic clock is probabilistic, it makes ticks at a certain probability in each unit of time. Ticks may not be evenly spaced.**
- **Stochastic clocks are not necessarily inaccurate – atomic clocks are stochastic 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.**



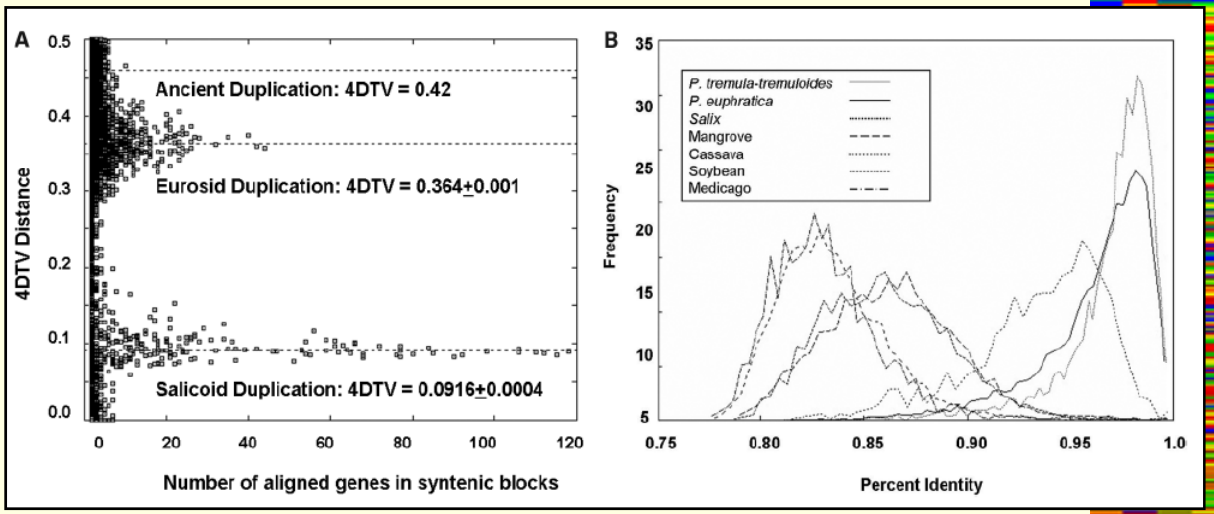
Multiple Alignment and Trees

Timing gene duplications

4DTV distance

fourfold synonymous third-codon transversion

- Identify possibly duplicated genes (BLAST)
- Calculate 4DTV based on alignments
- third codon position is not usually selected (should be clocklike)
- Find pairs of genes with same distances

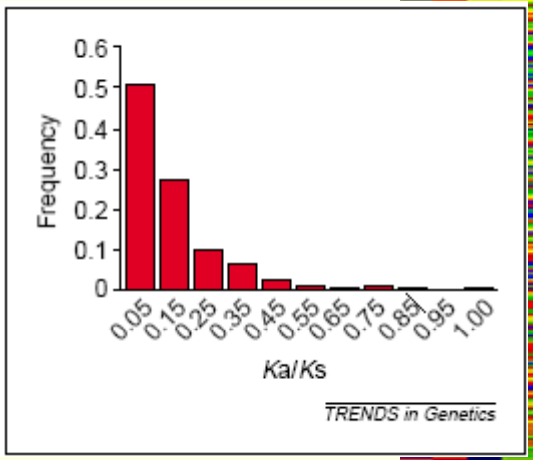


Multiple Alignments and Trees

Identifying selection

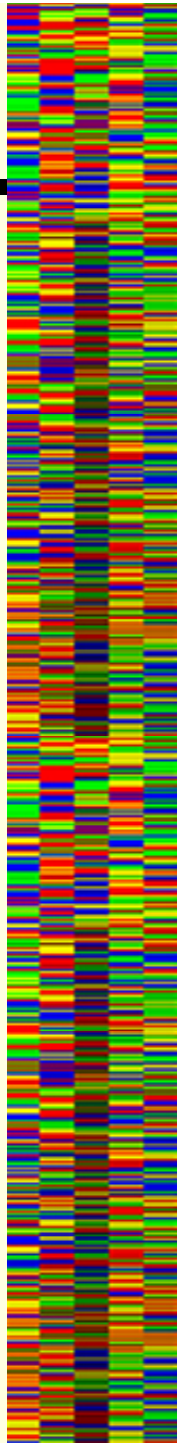
Ka/Ks ratio

- ratio of non-synonymous (change amino acid) to synonymous (don't change amino acid) changes in coding gene DNA
- Ks measures clocklike behavior
- If Ka is much larger, it is changing in a not clock-like way – it is being selected
- Has to be adjusted for multiple hits if time period is long
- Has to be adjusted for 2, 4, and 6 codon families
- Usually Ka<<Ks (most non-synonomous changes are bad)



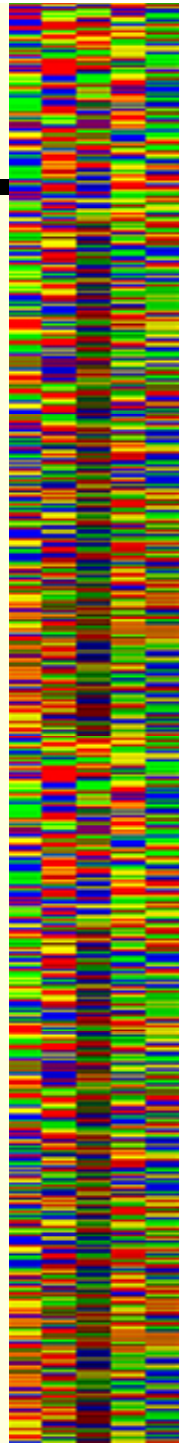
Tests for Trees

- **Non-parametric bootstrap and jacknife**
- **Tree length**
- **Winning sites tests (and other psir tests)**
- **Bayesian posterior probabilities**
- **Likelihood ratio tests**



Bootstrap - interpretation

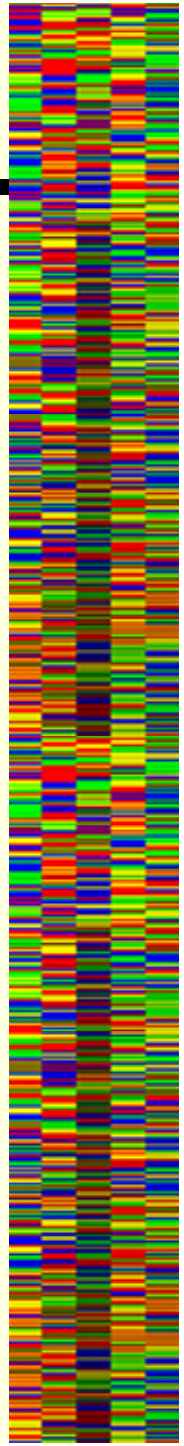
- **Bootstrapping is a very valuable and widely used technique (it is demanded by some journals), but requires a pragmatic interpretation:**
- **BPs depend on two aspects of the support for a group - the numbers of characters supporting a group and the level of support for incongruent groups**
- **BPs thus provides a reasonable index of the relative support for groups provided by a set of data**
- **Bootstrap confidence applies to single nodes and not trees**
- **In simulations BP>70% is typical for correct topologies**



Tests for Trees

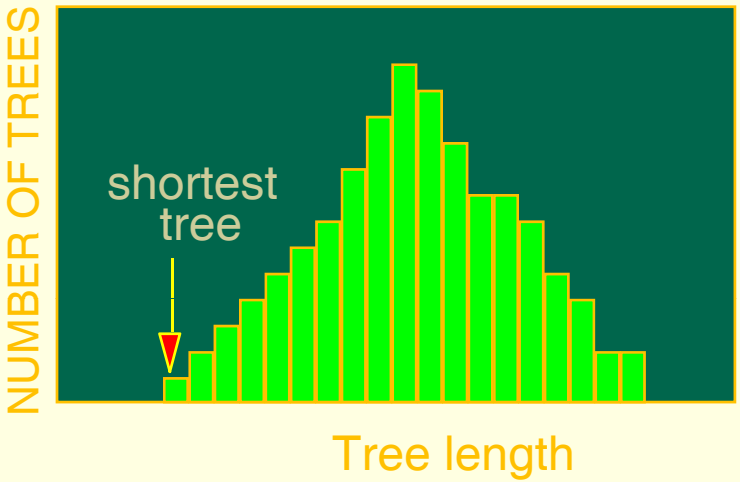
Jackknifing

- very similar to bootstrapping and differs only in the character resampling strategy
- Some proportion of characters (e.g. 50%) are randomly selected and deleted
- Replicate data sets are analysed and the results summarised with a majority-rule consensus tree
- Jackknifing and bootstrapping tend to produce broadly similar results and have similar interpretations

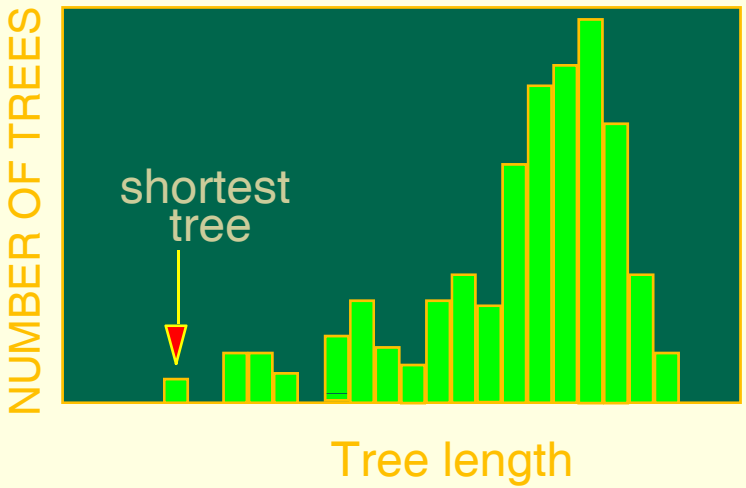


Tests for Trees

Skewness of Tree Length Distributions



- Studies with random (and phylogenetically uninformative) data showed that the distribution of tree lengths tends to be normal



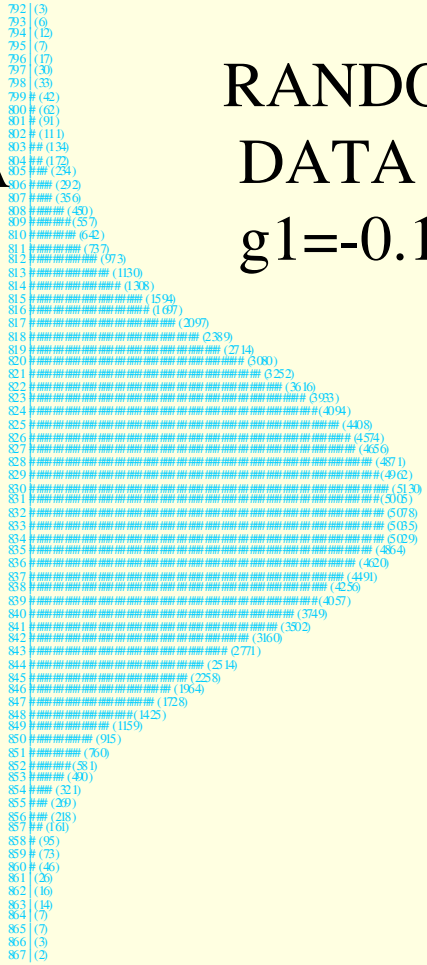
- In contrast, phylogenetically informative data is expected to have a strongly skewed distribution with few shortest trees and few trees nearly as short

Skewness - example

REAL DATA
Ciliate SSUrDNA
g1=-0.951947

Frequency distribution of tree lengths

RANDOMLY PERMUTED DATA
g1=-0.100478

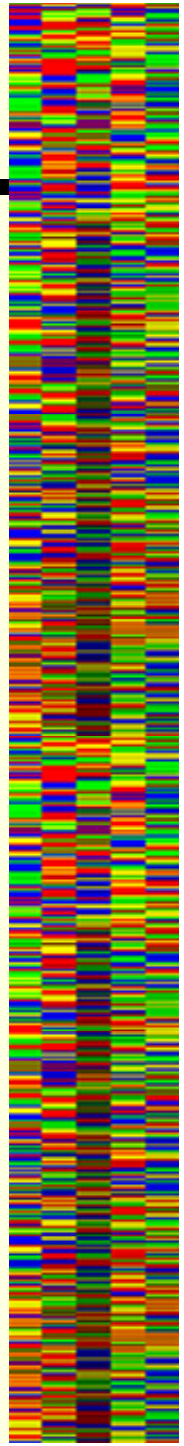


Frequency distribution of tree lengths

Decay analysis

Parsimony

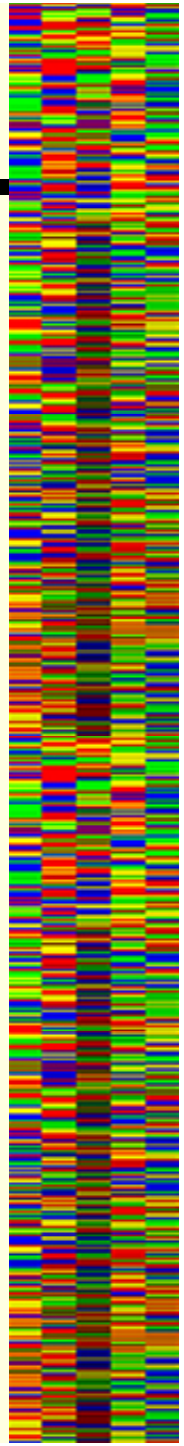
- **In parsimony analysis, a way to assess support for a group is to see if the group occurs in slightly less parsimonious trees also**
- **The length difference between the shortest trees including the group and the shortest trees that exclude the group (the extra steps required to overturn a group) is the decay index or Bremer support**
- **Total support (for a tree) is the sum of all clade decay indices - this has been advocated as a measure for an as yet unavailable matrix randomization test**



Paired-sites tests

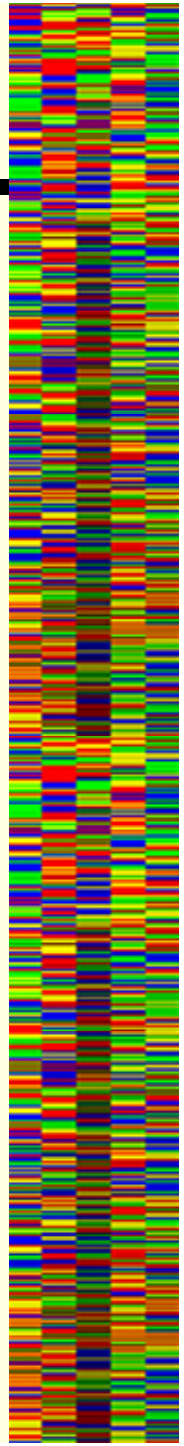
Some sites support one tree some another

- **The basic question:**
  - **Is tree A significantly better than tree B or are the differences (in parsimony or likelihood scores) within the expectations of random error?**
- **Compare two trees for either parsimony or likelihood scores**
  - **Number of steps or likelihood score at each site**
- **Decide whether the difference is significant by comparing to an assumed distribution (binomial or normal), or creating a null distribution using bootstrap sampling techniques**



Paired-sites tests: parsimony

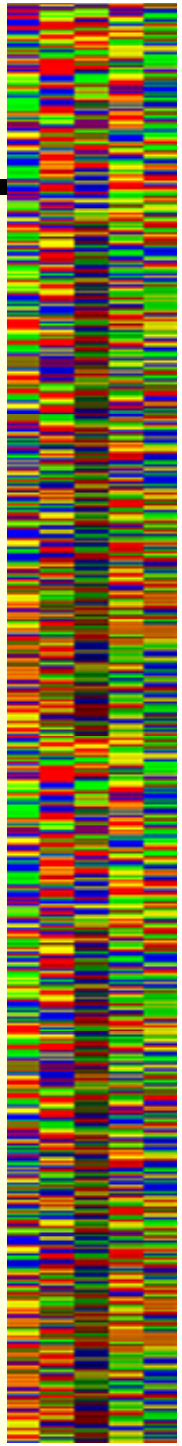
- **Developed by Alan Templeton (originally for restriction site data; 1983)**
- **Winning sites test (Prager and Wilson 1988)**
  - **a simpler version of Templeton**
  - **for each site, score which tree is better so that each site is assigned either + or - (or 0 if the are equivalent). Use a binomial distribution to test whether the fraction of + versus - is significantly different from 0.5**
- **Kishino-Hasegawa test (1989)**
  - **Assume all sites are i.i.d**
  - **Test statistic T = ΣT(i)**
    - *Where T(i) is the difference in the minimum number of substitutions on the two trees at the ith informative site.*
    - *Expectation for T (under the null hypothesis that the two trees are not significantly different) is zero. The sample variance can be obtained with an equation and tested with a t-test with n-1 degrees of freedom (n = number of informative sites)*
  - **If no a priori reason to suspect one tree better than other (two-tailed)**
  - **If a priori reason (such as one tree being the most parsimonious), test is not valid**



Paired-sites tests: parsimony

Example

- **One dataset**
- **Two trees:**
  - **Tree 1: 1153 steps (Better)**
  - **Tree 2: 1279 steps (Worse)**
- **Question: Are the two trees significantly different?**



Winning sites: PAUP* output

Comparison of tree 1 (best) to tree 2:

	------ Changes ------			----- Ranks -----
Character	*Tree 1*	*Tree 2*	*Difference*	*Positive Negative*
7	*2*	*3*	*1*	*81.5*
8	*1*	*2*	*1*	*81.5*
19	*4*	*3*	*-1*	*-81.5*
25	*5*	*6*	*1*	*81.5*
28	*2*	*1*	*-1*	*-81.5*
34	*3*	*4*	*1*	*81.5*
43	*4*	*3*	*-1*	*-81.5*
.				
82	*2*	*3*	*1*	*81.5*
88	*2*	*4*	*2*	*164*
96	*1*	*2*	*1*	*81.5*
.				
888	*1*	*2*	*1*	*81.5*
890	*4*	*3*	*-1*	*-81.5*
Totals		*126*	*12068*	*-1793*

Winning-sites (sign) test:

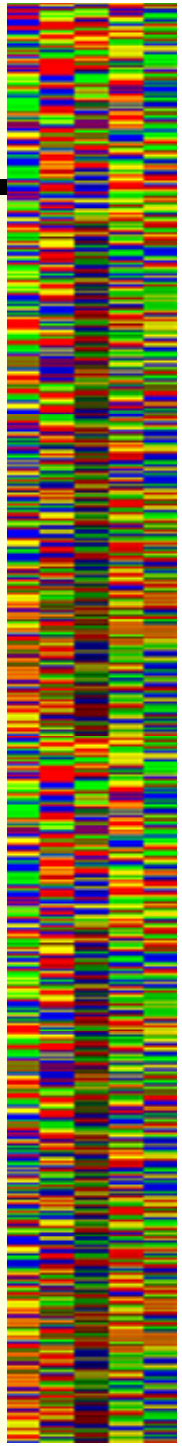
N = 166

Number of positive differences = 144

Number of negative differences = 22

Test statistic = 0.867470

P < 0.0001 (normal approximation with z = 9.391421)*



Wilcoxon signed-ranks (Templeton) test

Comparison of tree 1 (best) to tree 2:

Character	Changes: Tree 1	Changes: Tree 2	Difference	Ranks: Positive	Ranks: Negative
7	2	3	1	81.5	
8	1	2	1	81.5	
19	4	3	-1		-81.5
25	5	6	1	81.5	
28	2	1	-1		-81.5
34	3	4	1	81.5	
43	4	3	-1		-81.5
.					
82	2	3	1	81.5	
88	2	4	2	164	
96	1	2	1	81.5	
.					
888	1	2	1	81.5	
890	4	3	-1		-81.5
Totals			126	12068	-1793

Templeton (Wilcoxon signed-ranks) test:

N = 162

Test statistic (T) = 1467

P < 0.0001 (normal approximation with z = -9.899495)*