BIO4102/BIO6102/MSB315

Evolutionary Ecology
(Varsha 2023)

Ullasa Kodandaramaiah

MODULE: PHYLOGENETICS

Acknowledgments: Some content taken and modified from slides used by others in teaching. Caro-Beth Stewart, Niklas Wahlberg, Arthur Chou, Robert Cox, etc

Saturniidae

Photo: Arthur Anker

Gregariousness & Aposematism

What came first?

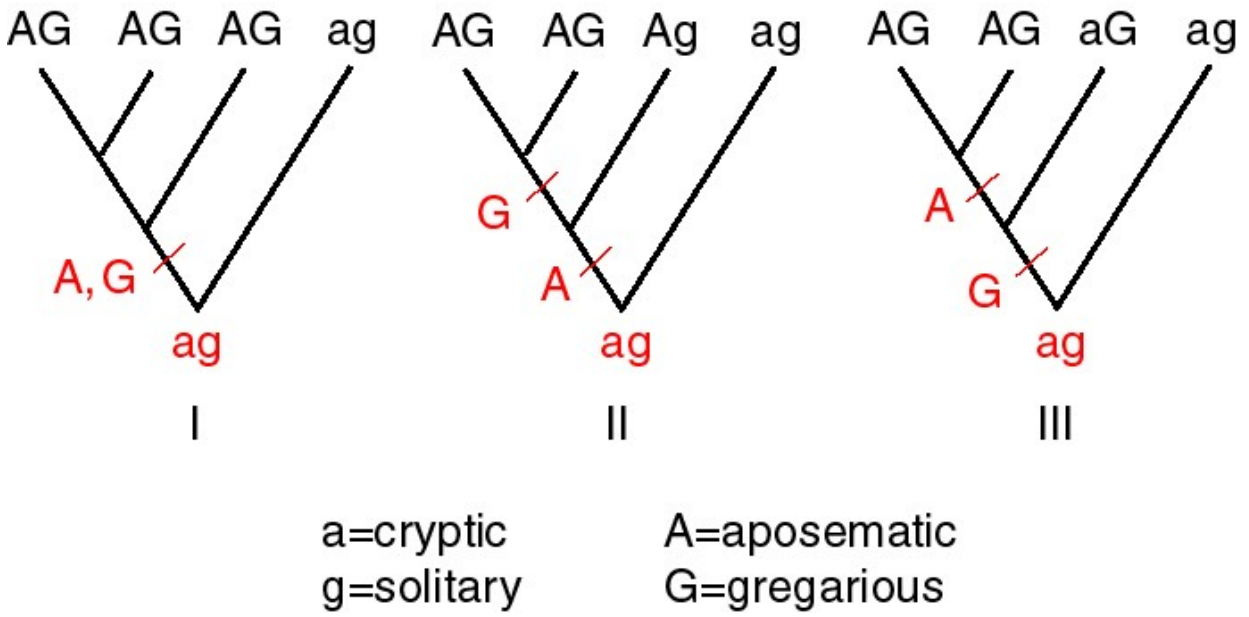


Figure taken from www.biology.duke.edu, Mark Rausher

Sillen-Tullberg 1988 *Evolution* 42(2): 293-305

Used bright colouration as a proxy for unpalatability

potential problems? (batesian mimicry, unpalatable cryptic species)

Definition of conspicuous colors: combinations of black with yellow, white, red & orange

Definition of gregariousness: at least 10

In her butterfly dataset, she found that bright colouration always preceded gregariousness

Phylogeny

What we see today in nature is the outcome of what has happened in the past

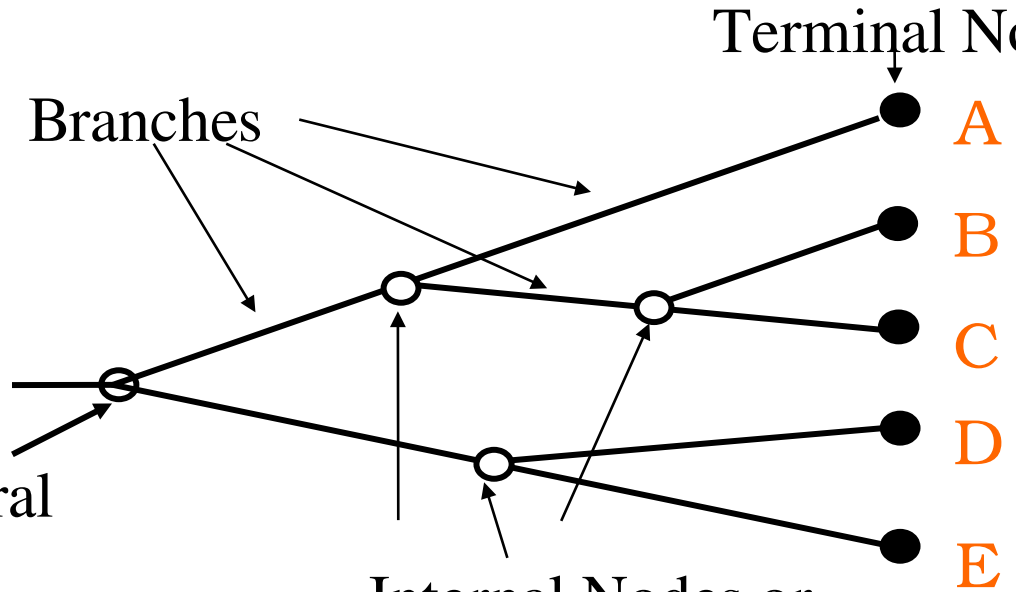
An ancestral species gives rise to two or more daughter species through speciation, all of which are potential ancestral species that can further undergo speciation

A species-level phylogeny is a reconstruction of historical speciation events, depicted in the form a tree

A phylogeny can also represent relationships among lineages other than species, e.g. individuals within a species

Tree Terminology

Monophyletic group=clade

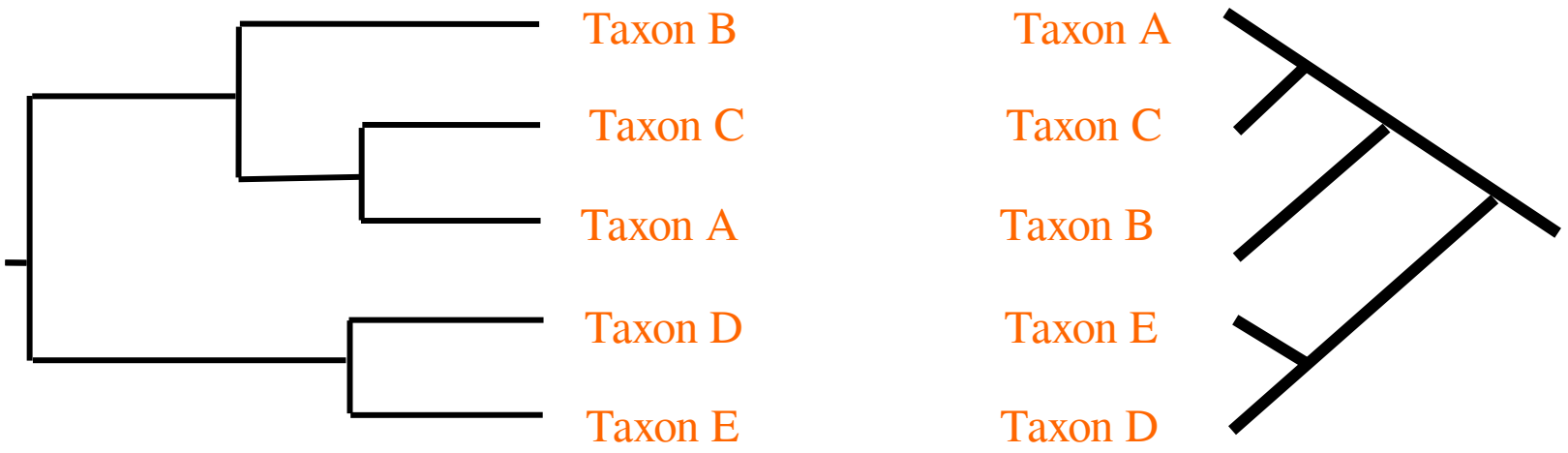


Represent the TAXA (species, populations, etc.)

sister clades (sister taxa)
C, B
A, BC
ABC, DE
D, E

Trees can be flipped at nodes

Can be depicted in different ways: rectangular, slanted, etc



How do we find a tree for a given set of taxa?

Important !! We dont know what the true phylogeny is.
We can only estimate - *phylogenetic hypothesis*

Collect data on character states of multiple characters for all taxa of interest, and analyze the data.

Character

A feature of an organism that can be observed or measured. Part or attribute.

Heritable

Character-state

One of the alternate conditions of a character

Morphological data

Character	Character states
Wings	presence, absence
Mouth part	absense, chewing, sucking, piercing, etc.
No. of petals in a flower	0, 3, 5, 8, 13, 21, 34, 55

Exercise

Identify 5 characters and their character states in these taxa. Use information about *similarity of character states* to reconstruct the phylogeny of these animals

- Lion
- Domestic Cat
- Zebra
- Zebrafish
- Common Mormon butterfly

- **Molecular data (molecular phylogenetics)**

- Most commonly DNA sequences

  Character – Position in sequence

  Character state – A, T, G, C, Gap

Taxa	Characters
Species A	A T G G C T A T T C T
Species B	A T C G C T A G T C T
Species C	T T C A - - - G A C C
Species D	T T G A C C A G A C C
Species E	T T G A C C A G T T C

- Today, phylogenies are usually reconstructed using DNA sequence data, and rarely using other types of data

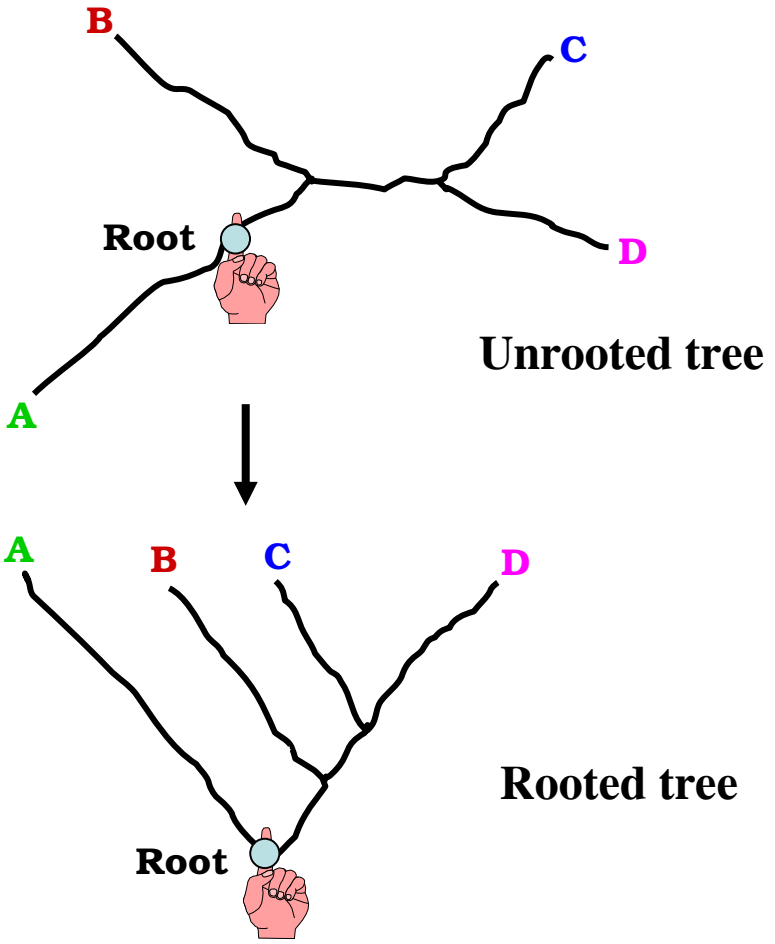
Exercise

Reconstruct the phylogeny of the 5 species

Rooted and unrooted trees

To root a tree mentally, imagine that the tree is made of string. Grab the string at the root and tug on it until the ends of the string (the taxa) fall opposite the root.

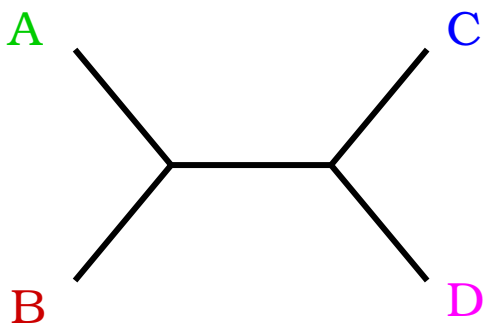
Unrooted trees have information about relationships. Rooted trees have information about relationships and direction of evolution



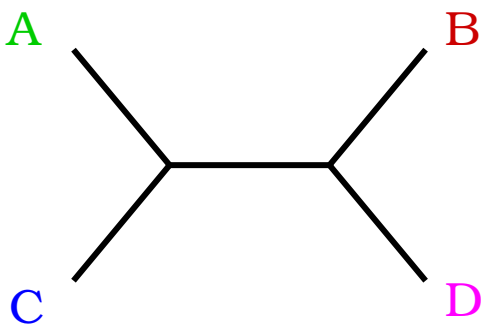
This and next few slides based on slides by Caro-Beth Stewart

No. of possible unrooted trees for four taxa (A, B, C, D)

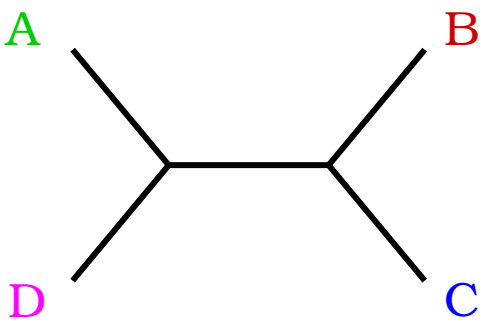
Tree 1



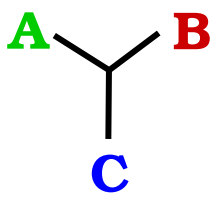
Tree 2

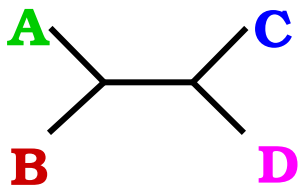


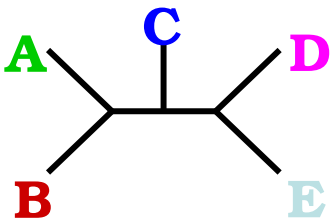
Tree 3

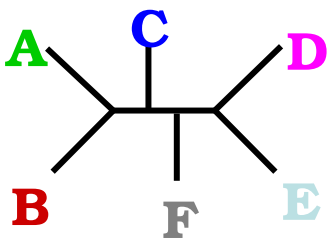


The number of unrooted trees increases in a greater than exponential manner with number of taxa









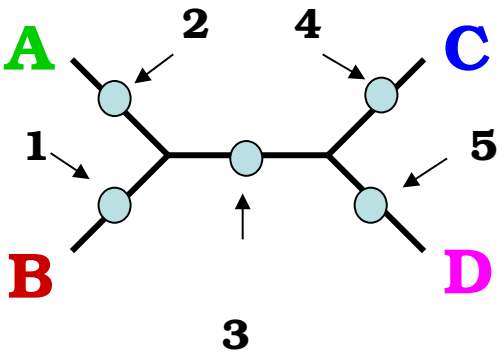
# Taxa (N)	# Unrooted trees
3	1
4	3
5	15
6	105
7	945
8	10,935
9	135,135
10	2,027,025
.	.
.	.
.	.

$$\frac{(2n-5)!}{[2^{n-3}\,(n-3)!]}$$

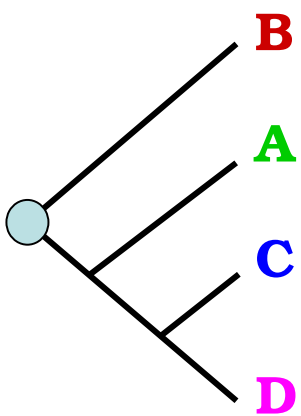
= no. of unrooted trees for N taxa

An unrooted, 4-taxon tree can be rooted in 5 places to produce 5 rooted trees

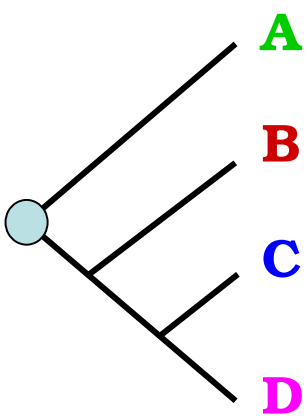
The unrooted tree 1:



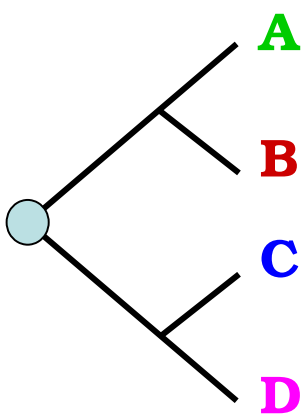
Rooted 1a



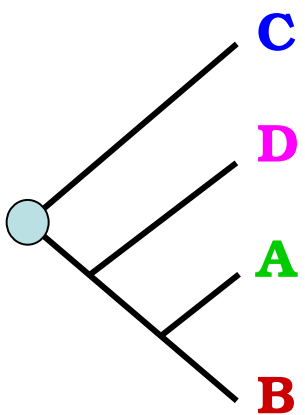
Rooted 1b



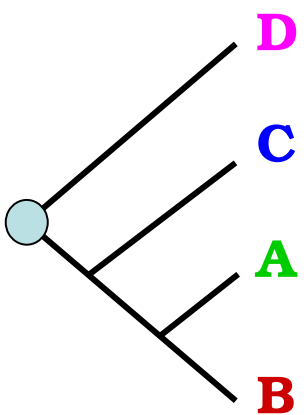
Rooted 1c



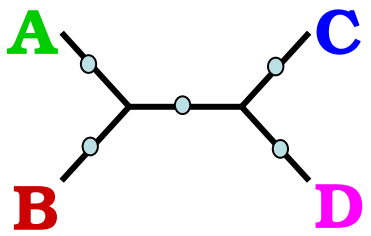
Rooted 1d

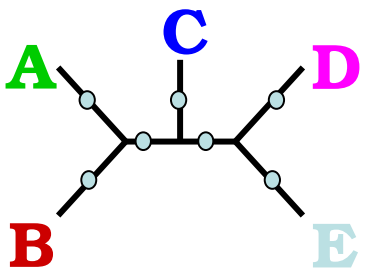


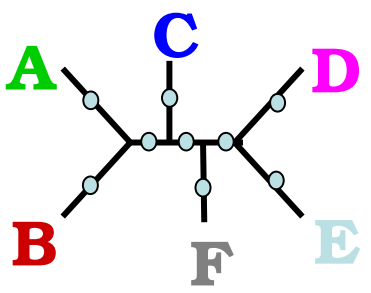
Rooted 1e



Each unrooted tree can be rooted anywhere along any of its branches







# Taxa	# Unrooted Trees	# Roots	# Rooted Trees
3	1	3	3
4	3	5	15
5	15	7	105
6	105	9	945
7	945	11	10,395
8	10,935	13	135,135
9	135,135	15	2,027,025
.	.	.	.
.	.	.	.
.	.	.	.

$$\frac{(2n-3)!}{[2^{n-2}\ (n-2)!]}$$ = no. of rooted trees for N taxa

The total number of rooted trees is much higher than that of unrooted trees.

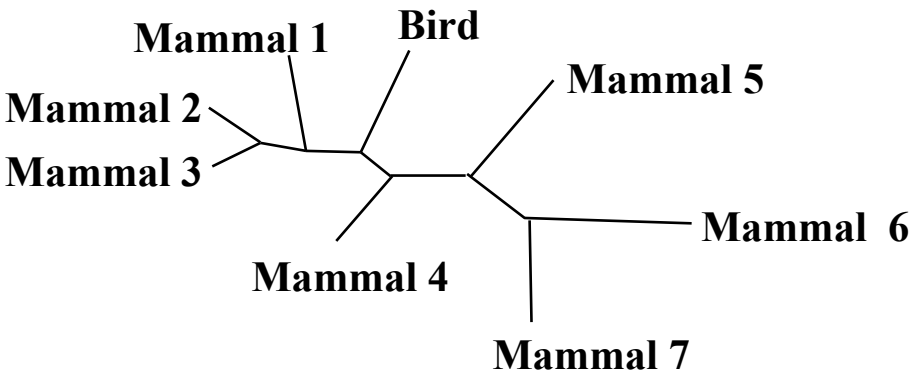
We start by examining unrooted trees. We choose a tree that is we think is the best tree and root it with an *outgroup* - a taxon that does not belong to the group of interest.

Assumption: The divergence of the outgroup from the ingroup (i.e., group of interest) happened before the first divergence within the ingroup

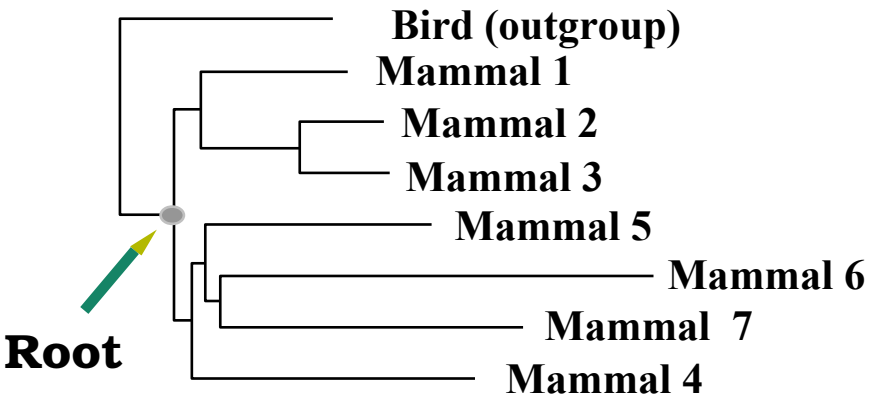
The outgroup has to be included in the analysis for us to know where to place it.

Rooting with outgroups

Unrooted tree



Rooted by Outgroup



A bird was used an outgroup to root a tree of mammals. Similarly, a mammal can be used to root a tree of birds

Which is the best tree?

Different methods of phylogenetic analysis differ in their '*optimality criterion*'.

Most popular

1) Maximum Parsimony

2) Maximum Likelihood

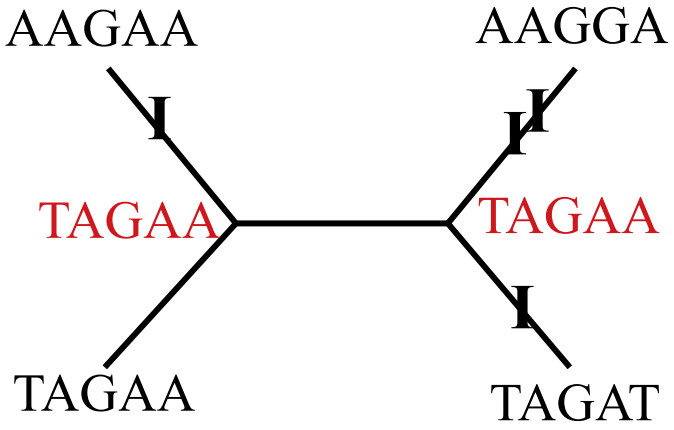
3) Bayesian

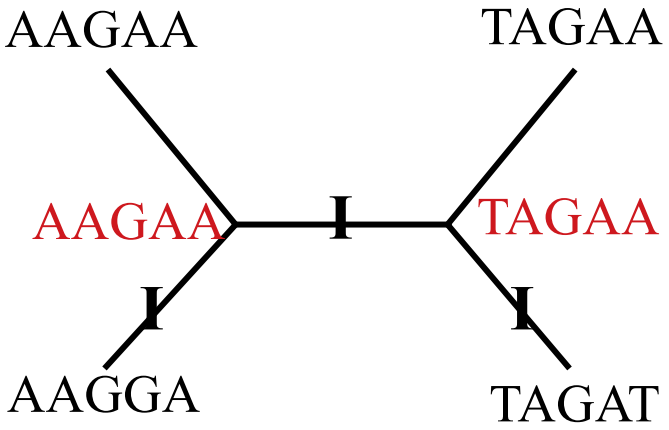
Maximum parsimony

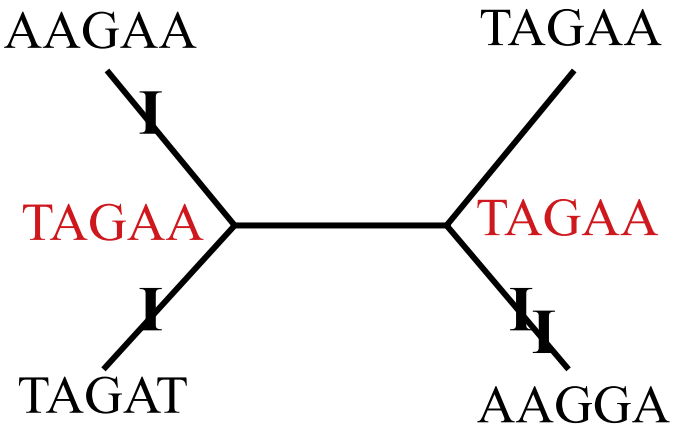
***Principle of parsimony*:** All things being equal, the simplest solution (explanation) tends to be the best one

- The best hypothesis is the one with the fewest assumptions.
- The best tree is the one with the fewest number of evolutionary changes (the shortest tree). For DNA sequence data, the best tree is the one with fewest substitutions

Species *Alpha*: AAGAA
Beta: AAGGA
Gamma: TAGAT
Delta: TAGAA







Ancestral sequences (internal nodes) are in red.
Which tree is the most parsimonious?

If all possible trees are investigated for length – exhaustive search (simplest algorithm).

Guaranteed to find the best tree.

Computationally difficult. Impossible for large number of taxa.

Heuristic methods: short-cuts.

- Heuristic: Technique to solve a problem that ignores whether the solution can be proven to be correct, but which usually produces a good solution.
- Intended to gain computational performance, potentially at the cost of precision
- Involves sets of algorithms

In phylogenetics, heuristic methods usually involve two steps.

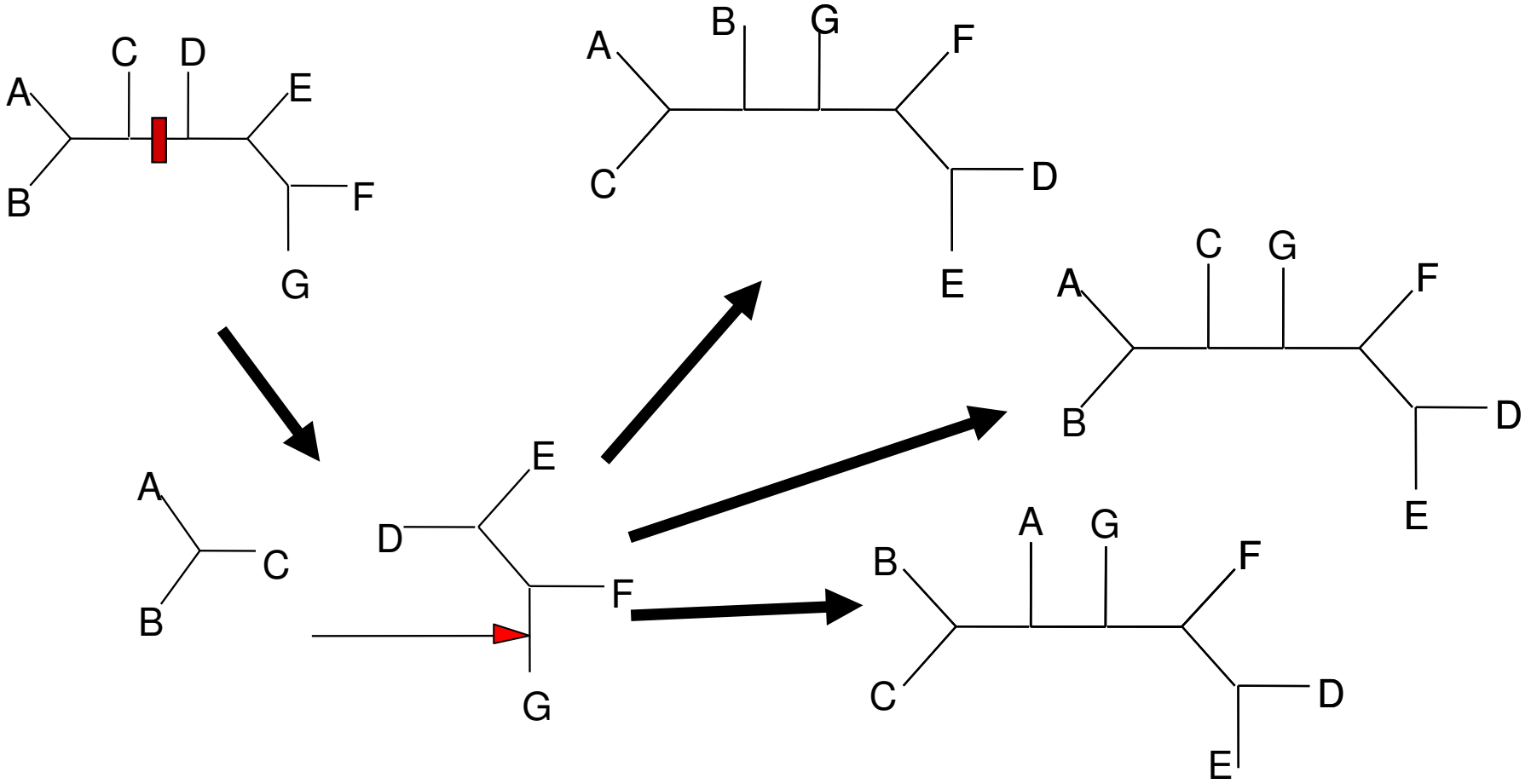
1. Build a 'starting' tree using a fast (potentially inaccurate) method (e.g., by adding taxa in random order, with each taxon placed on the tree such that the resulting tree is the most optimal for that set of taxa)

2. Try different rearrangements to improve upon the tree. Keep trying until you cannot find an improvement. (e.g. branch swapping)

These two steps are usually repeated many times over – replications.

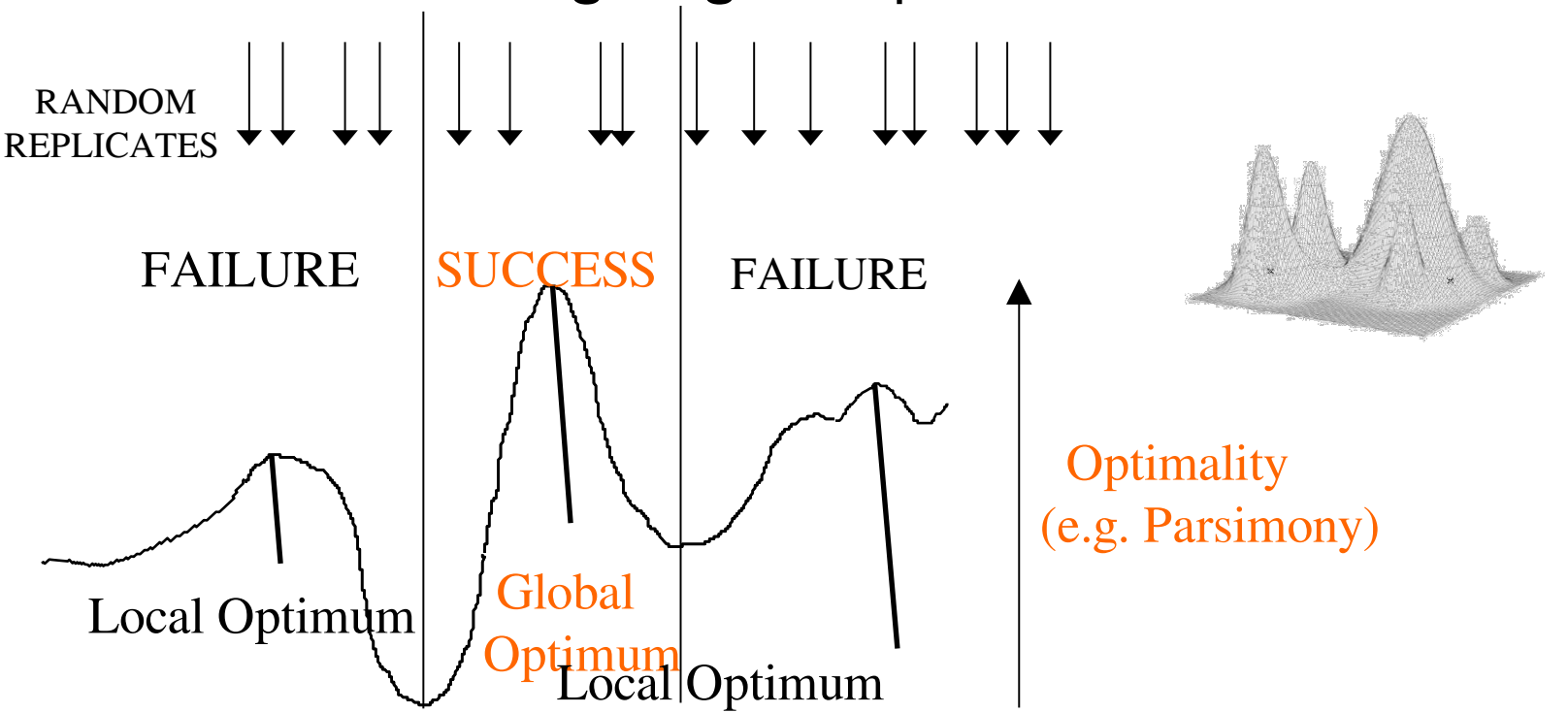
Rerrangements: Branch-swapping

Popular: Tree Bisection and Reconnection



Heuristic methods use 'greedy algorithms' – hill climbing

Tree space may be populated by local optima (smaller peaks) and a global optimum (the highest peak). The global optimum is what we want to reach. By using a large number of random replicates (i.e., starting points) we maximize the chances of reaching the global optimum



How confident can we be about the inferred relationships?

There are multiple clades (i.e, multiple sets of relationships in a phylogeny) and not all of these are equally well supported by the data.

Measures of clade support give us an idea of how much confidence we can have that the grouping represents the grouping in the true phylogeny

Most popular: **Bootstrapping**

Bootstrapping phylogenies

Characters are re-sampled with replacement to create many bootstrap pseudo-replicate datasets.

Original data matrix

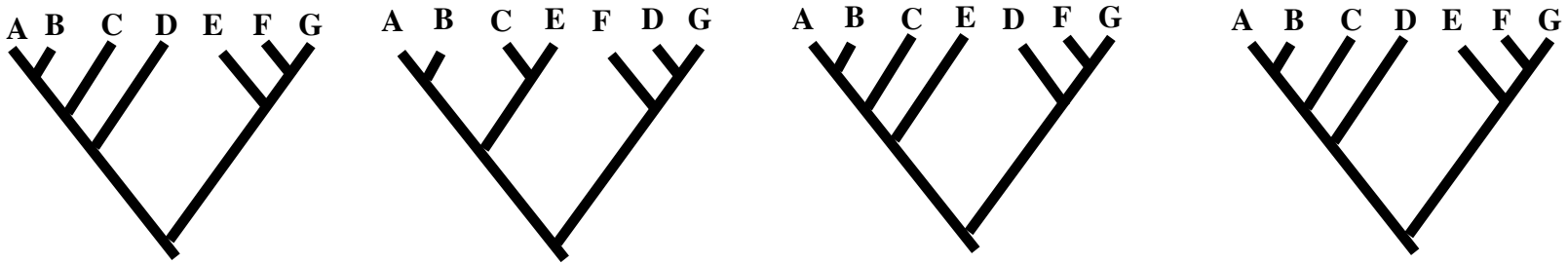
Characters							
Taxa	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Alpha	A	G	G	T	C	G	G
Beta	A	G	G	T	T	T	T
Gamma	G	G	T	G	A	C	A
Delta	A	T	G	A	A	G	T

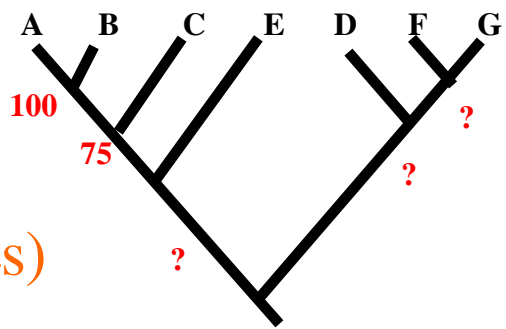
Resampled data matrix (*pseudo-replicate*)

Characters							
Taxa	**5**	**7**	**7**	**3**	**2**	**6**	**5**
Alpha	C	G	G	G	G	G	C
Beta	T	T	T	G	G	T	T
Gamma	A	A	A	T	G	C	A
Delta	A	T	T	G	T	G	A

Each bootstrap psuedo-replicate data set is analyzed using similar heuristic searches as those used to find the most parsimonious tree.

Bootstrap proportions are the frequencies with which groups are encountered in analyses of replicate datasets. Values closer to 100 indicate high support, values <50 indicate poor (weak) support





e.g. with 4 bootstrap pseudo-replicates
(in practice, 100 to 1000 psuedo-replicates)

Model based methods
(Maximum Likelihood and Bayesian)

MP performs poorly in some cases. Model-based methods have been shown to be more reliable. They use models of DNA substitution that incorporate information about the rates at which each nucleotide substitution. E.gs. of models:

- Jukes Cantor (simplest model, mutation occurs at a constant rate, each nucleotide is equally likely to mutate into any other nucleotide with rate)
- Kimura 2 parameter (transitions and transversions have different rates of mutation)

…….

- Generalised Time Reversible GTR (most complex, each pair of nucleotide substitutions has a different rate)

Maximum Likelihood method

Likelihood (Tree) = Probability (Data | Tree)

- Data -> set of sequences
- Tree -> topology *and* branch lengths

Topology is the set of relationships. Therefore, two trees with the same set of relationships but different combinations of branch lengths are considered different trees (whereas the maximum parsimony approach considers only topology and would treat them as a single tree)

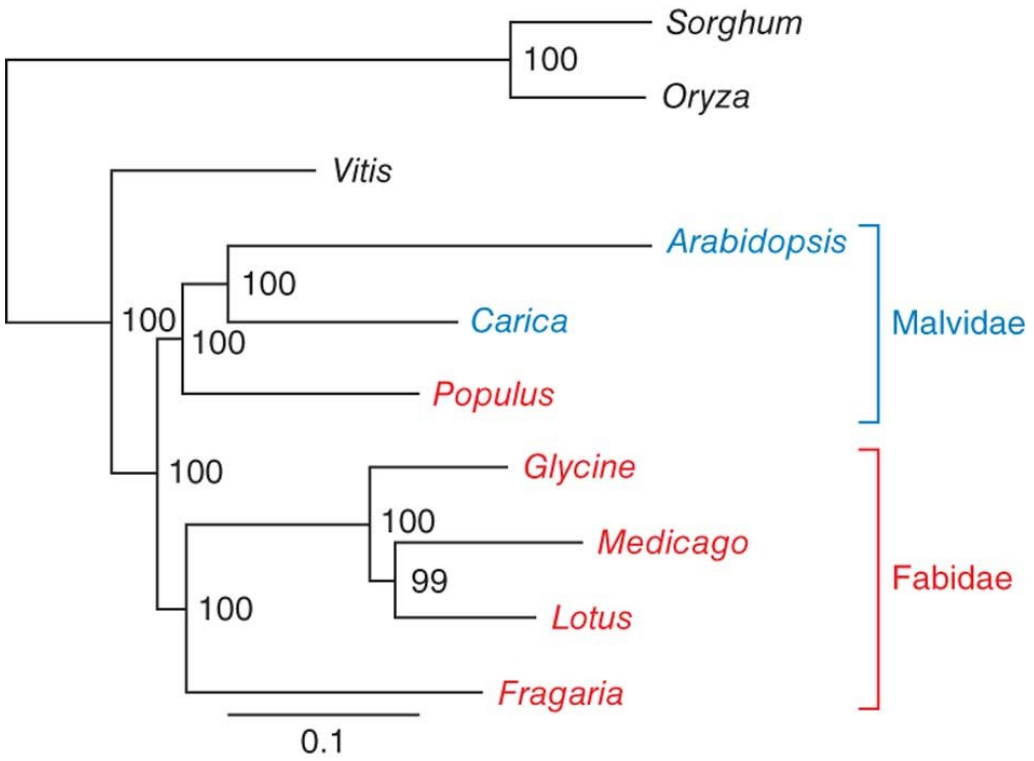
In the ML approach, the ML tree is the one that has the highest likelihood score.

Simulations have shown that ML performs better than MP

Clade support - bootstrapping.

Rooting - outgroup.

An ML analysis results in a phylogeny where branch lengths are proportional to evolutionary change (i.e,. no of substitutions)

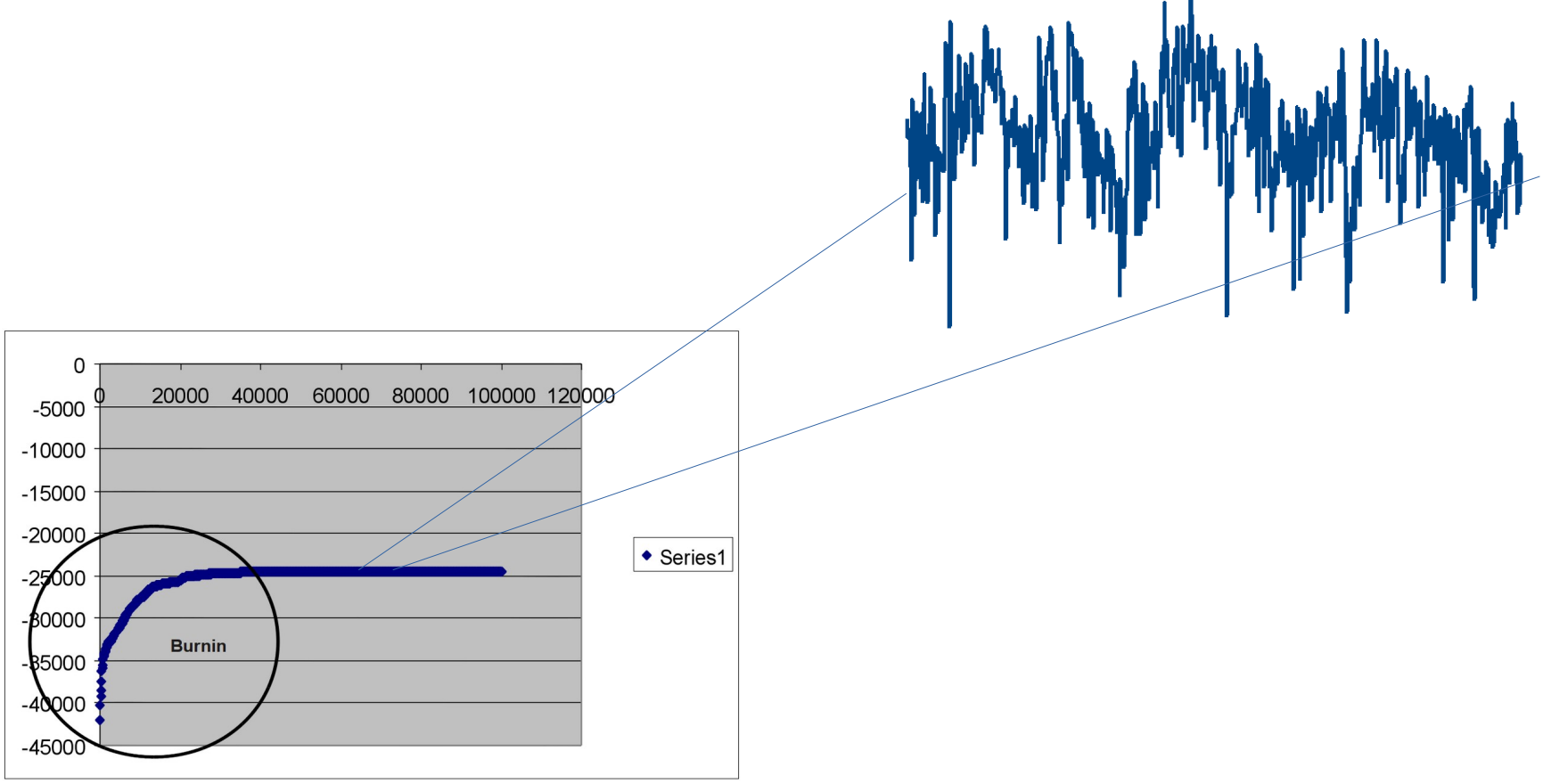


Shulaev *et al* (2011) *Nature Genetics* 43, 109–116

Bayesian Inference

Rather than trying to trying to find one ‘best’ tree, BI uses a Markov chain Monte Carlo (MCMC) approach to sample a large number of trees

- Begin with a random tree.
- Propose a new tree (e.g. using branch swapping ,or, simply by modifying the length of branches)
- Accept new tree if it has better likelihood.
- If the new tree is much worse, reject it.
- If new tree is almost as good, accept it, with a certain probability.
- Build up a chain of trees.



Above: Example of an MCMC run with log likelihoods plotted on the Y axis and the generation on the X. The 'burnin' represents the number of generations until the log likelihood curve stabilizes. The 'burnin' trees are discarded before summarizing.

Top right: Close-up of MCMC chain

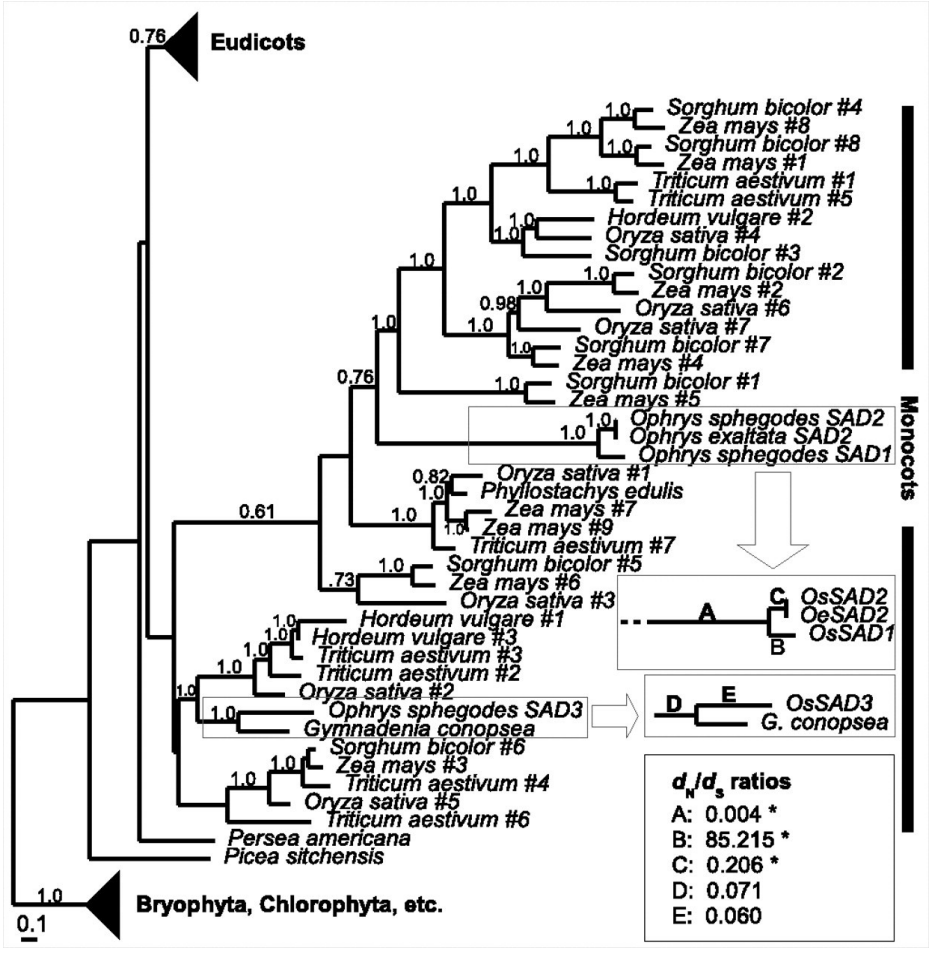
Clades with higher support should be present in more trees (and clades with low support should be present in fewer trees). Thus, the proportion of trees with a particular clade is an estimate of support for the clade (***posterior probability***).

After running the chain for a large no. of times, delete the initial trees where likelihood values have not stabilized (i.e. burnin).

The Bayesian phylogeny summarizes information in the remaining trees.

Rooting done using an outgroup.

Phylogenetic analysis of SAD homologs, showing monocot clade.



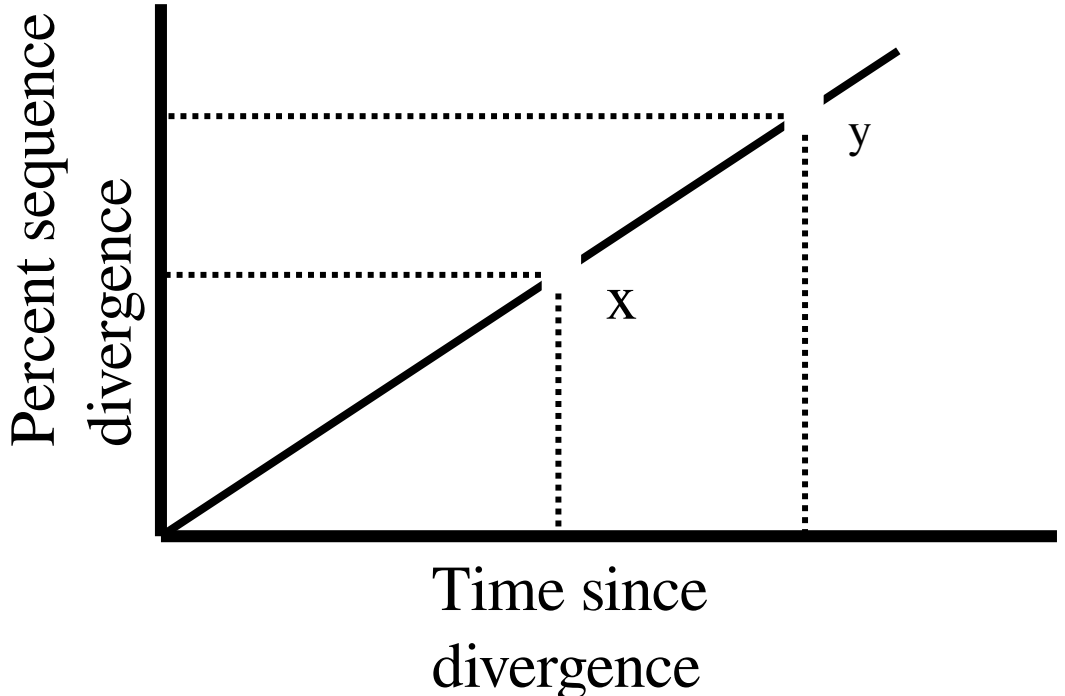
Schlüter *et al* (2011) PNAS 108:5696-5701

Molecular dating

- Result of ML or Bayesian analysis = phylogenies with branch length information
- Where, *Branch length* α *Time of evolution and Rate of evolution.*
- Molecular dating analyses try to tease them apart to estimate times of divergences.
  - Methods assuming molecular clocks
  - Methods not assuming molecular clocks

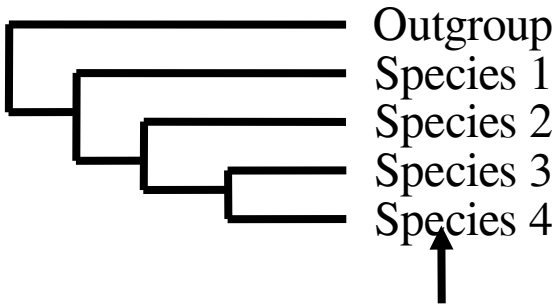
Molecular clocks

- The mutation rate for some genes may be relatively constant across species i.e follow a molecular clock.
- Branch lengths α time



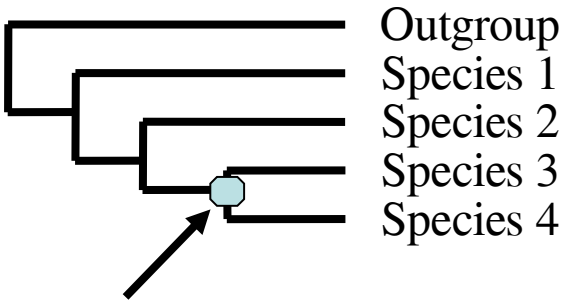
Molecular Clocks (simplistic example)

1) Reconstruct the phylogeny using a model based method



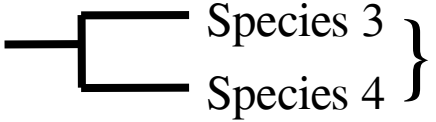
Fossil for Species 4
~1 MY

2) Date a Node in the Tree



You know that the most recent possible divergence between 3 and 4 is at least 1 MY

3) Calculate Divergence

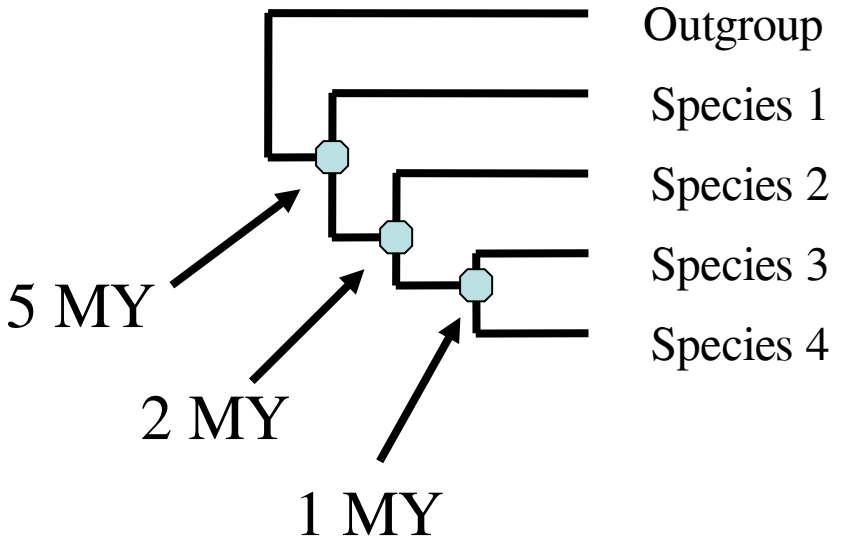


2%
Sequence
Divergence

4) Calculate a Rate

R= 2%/1MY

5) Extrapolate Rate to Other Nodes in Tree



Applicable only in a molecular clock like scenario of evolution

Methods that do not assume molecular clocks.

Rely on complex models (e.g. rate autocorrelation)

All molecular dating methods use fossil or other calibrations.

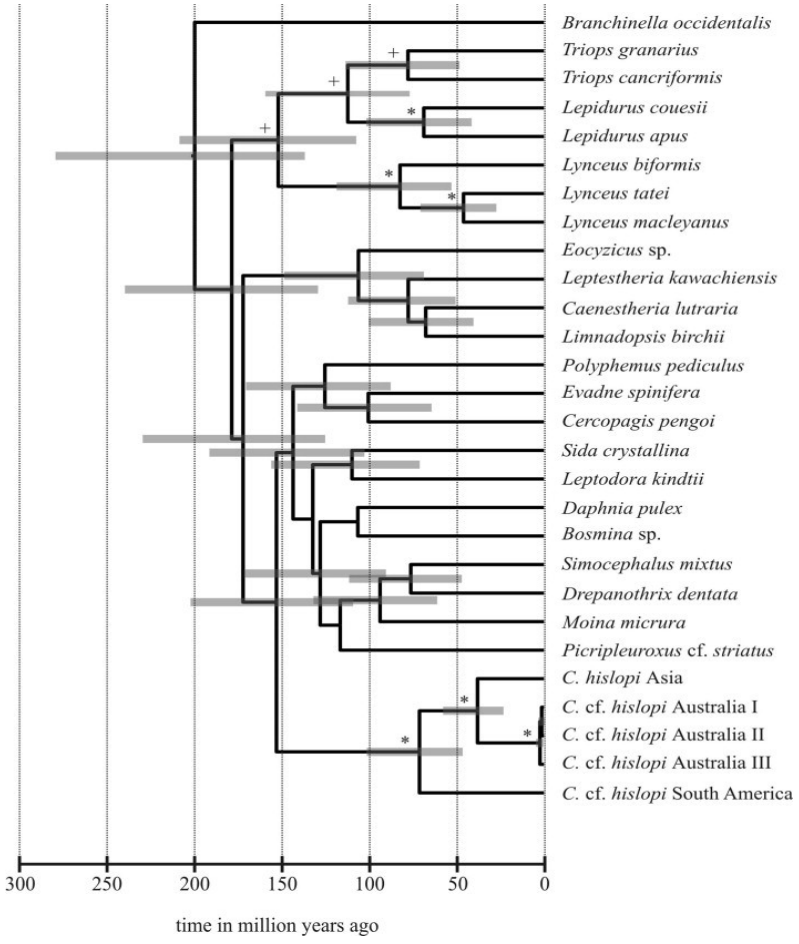


Fig. S1. Phylogenetic analyses of the COI data set with relaxed molecular clock estimates based on fossil calibration points. The following fossil calibration points were used: Phyllopoda 400 million ago years (mya), Spinic...

Schwentner *et al* (2013) *Molecular Phylogenetics & Evolution* 66: 800-810

Divergence events in the lab (e.g. viral strains) can be accurately reconstructed using phylogenetic methods

Phylogenetic hypotheses are not 'fixed in stone'.

Phylogeny as end v/s means.

Opens the door to many questions in evolutionary biology