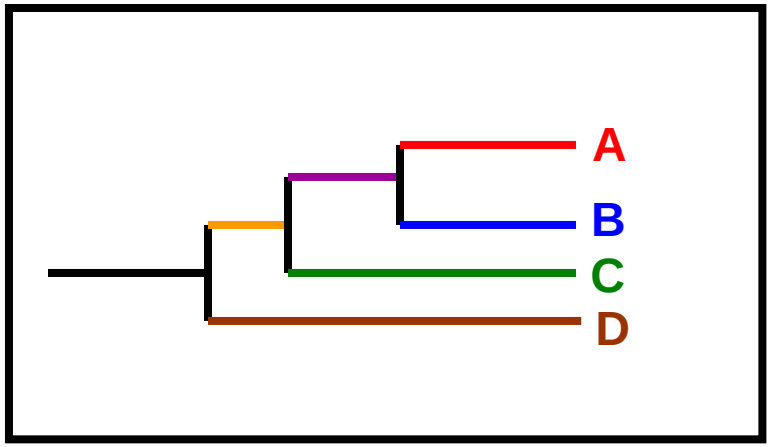
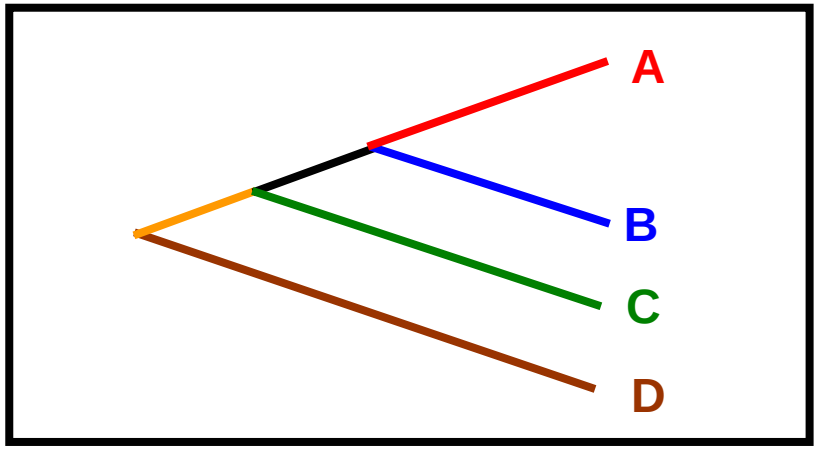
7.91 – Lecture #5 Michael Yaffe

Database Searching &
Molecular Phylogenetics





(((A,B)C)D)

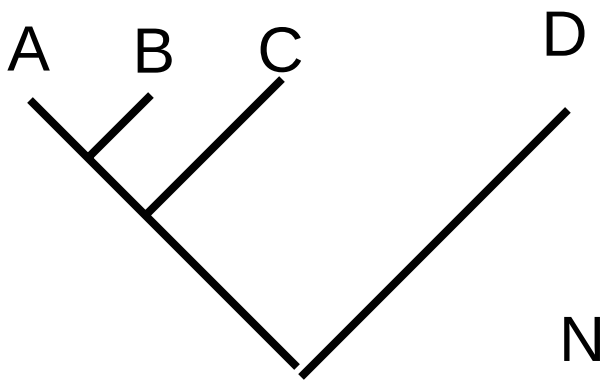
Outline

- Distance Matrix Methods
- Neighbor-Joining Method and Related Neighbor Methods
- Maximum Likelihood
- Parsimony
    - Branch and Bound
    - Heuristic Seaching
- Consensus Trees
- Software (PHYLIP, PAUP)
- The Tree of Life

Transformed Distance Method

UPGMA assumes constant rate
Of evolution across all lineages - can lead to wrong tree topologies

Can allow different rates of evolution across different lineages if you normalize using an external reference that diverged early…i.e. use an outgroup



Define $\bar{d}_D$ =average distance
Between outgroup and all ingroups

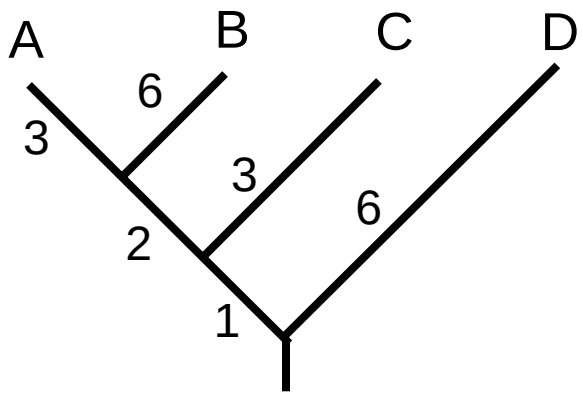
$$d'_{ij} = (d_{ij} - d_{iD} - d_{jD})/2 + \bar{d}_D$$

Now use d'_{ij} to do the clustering
..basically just comes from the insight that ingroups evolved separately from each other <u>ONLY AFTER</u> they diverged from outgroup

Example

Species	A	B	C
B	9		
C	8	11	
D	12	15	10

d_{AB} is distance Between A and B



$\overline{d_D} = 37/3$

Use D as outgroup

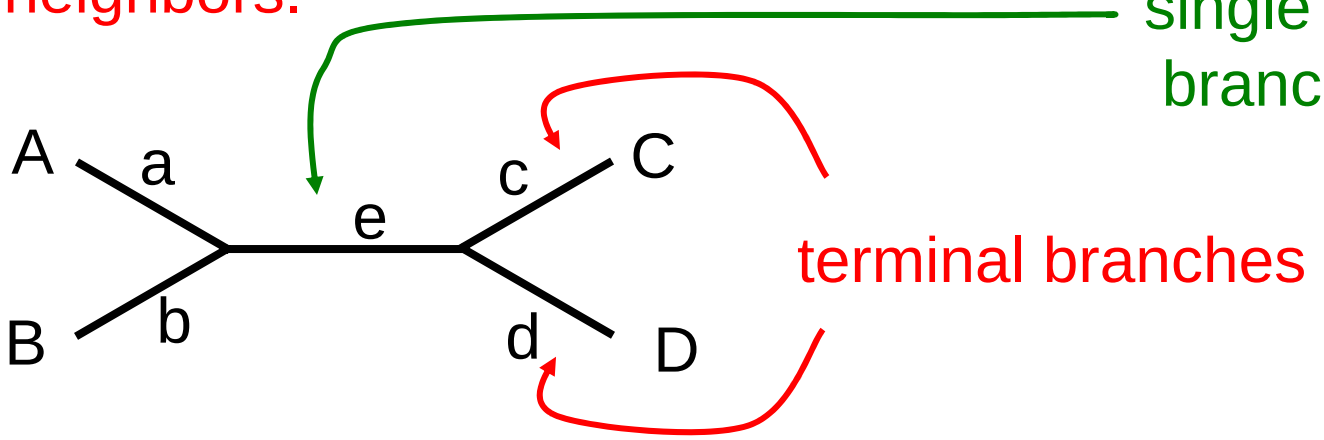
Species	A	B
B	10/3	
C	16/3	16/3

Now use UPGMA to build tree

Neighbor's Relation Method

Variant of UPGMA that pairs species in a way that creates a tree with minimal overall branch lengths.

Pairs of sequences separated by only 1 node are said to be neighbors.



For this tree topology

$d_{AC} + d_{BD} = d_{AD} + d_{BC} = a + b + c + d + 2e = d_{AB} + d_{CD} + 2e$

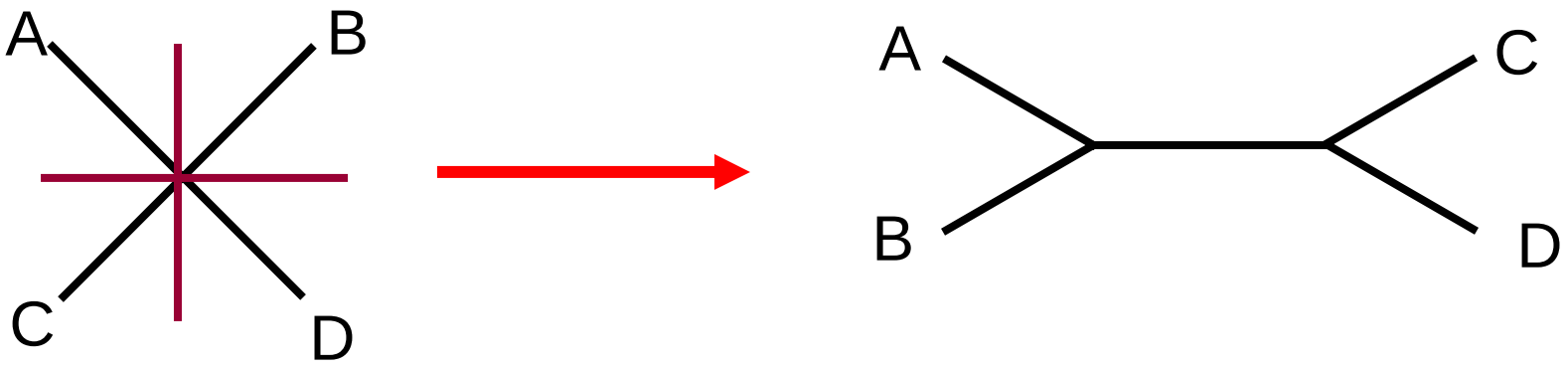
For neighbor relations, four-point condition will be true:

$d_{AB} + d_{CD} < d_{AC} + d_{BD}$ …and… $d_{AB} + d_{CD} < d_{AD} + d_{BC}$

So just have to consider all pairwise arrangements and determine which one satisfies the four-point condition.

Neighbor-Joining Methods

Start with star-like tree. Find neighbors sequentially to minimize total length of all branches



Studier & Kepler 1988:

$Q_{12}=(N-2)d_{12} - \Sigma d_{1i} - \Sigma d_{2i}$

Where any 2 sequences can be 1 and 2

Try all possible sequence combinations. Whichever combination of pairs gives the smallest Q_{12} is the final tree!

Maximum Likelihood

- A purely statistical method.
- Probablilities for every nucleotide substitution in a set of aligned sequences is considered.
- Calculation of probabilities is complex since ancestor is unknown
- Test all possible trees and calculate the aggregate probablility.
- Tree with single highest aggregate probablilty is the most likely to reflect the true phylogenetic tree.

VERY COMPUTATIONALLY INTENSE

Parsimony

Parsimony: a derogatory term from the 1930s and 1940s
To describe someone who was especially careful with
Spending money.

Biologically: Attach preference to an evolutionary pathway
That minimizes the number of mutational events since
(1) Mutations are rare events, and
(2) The more unlikely events a model postulates, the less l
likely the model is to be true.

Parsimony: a character-based method, NOT a distance-based
method.

Parsimony

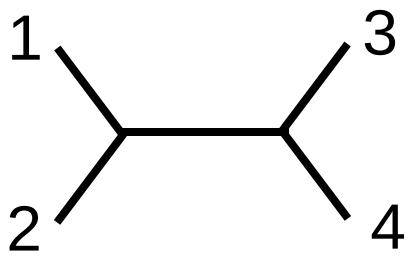
For parsimony analysis, positions in a sequence alignment fall into one of two categories: informative and uninformative.

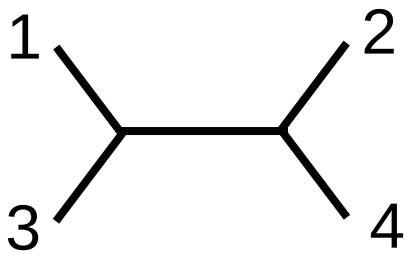
	Position					
Sequence	1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

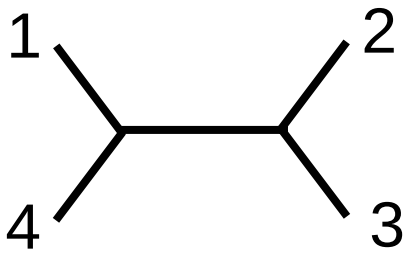
Only 3 possible unrooted trees you can make…

Parsimony

Sequence	Position 1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T





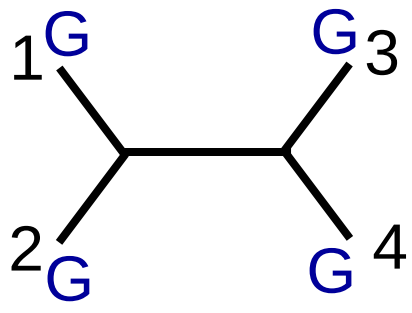


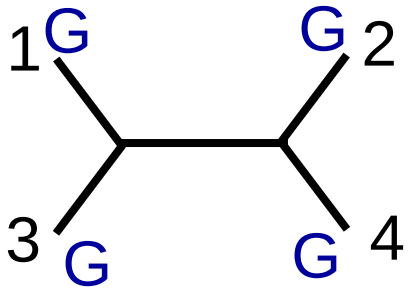
Which tree is the right one?

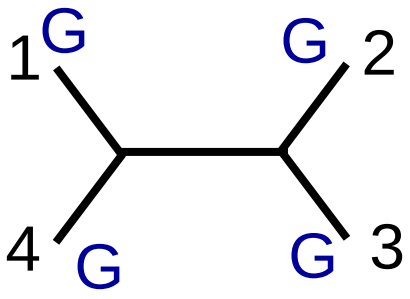
Parsimony

Position

Sequence	1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



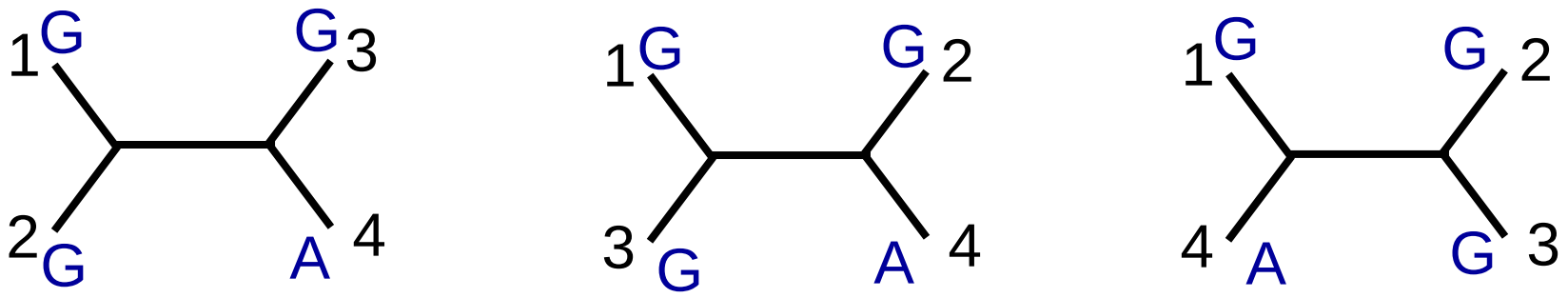




Invariant positions – contain NO INFORMATION → uninformative

Parsimony

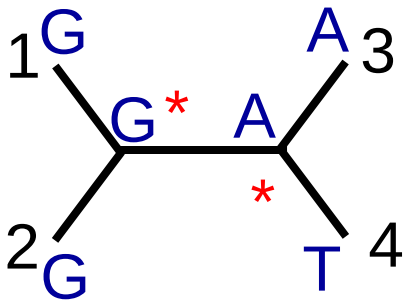
Sequence	Position 1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

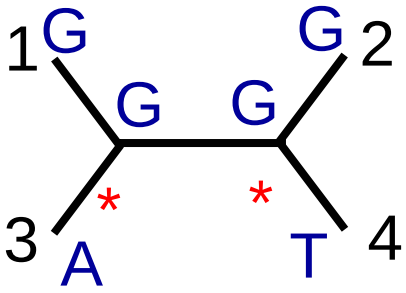


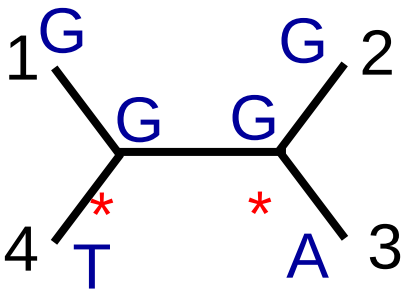
Equally uninformative – need one mutation in each tree

Parsimony

Sequence	Position 1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



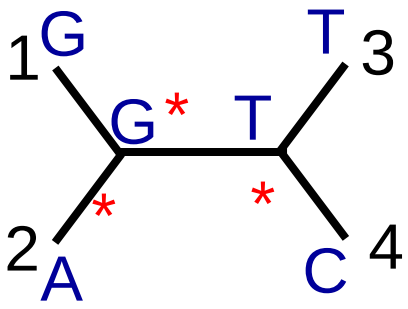


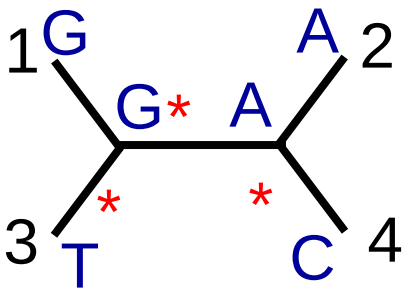


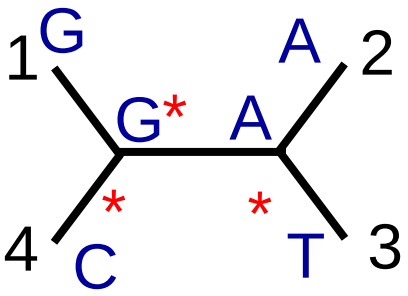
Also uninformative – need two mutations in each tree

Parsimony

Sequence	Position 1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T





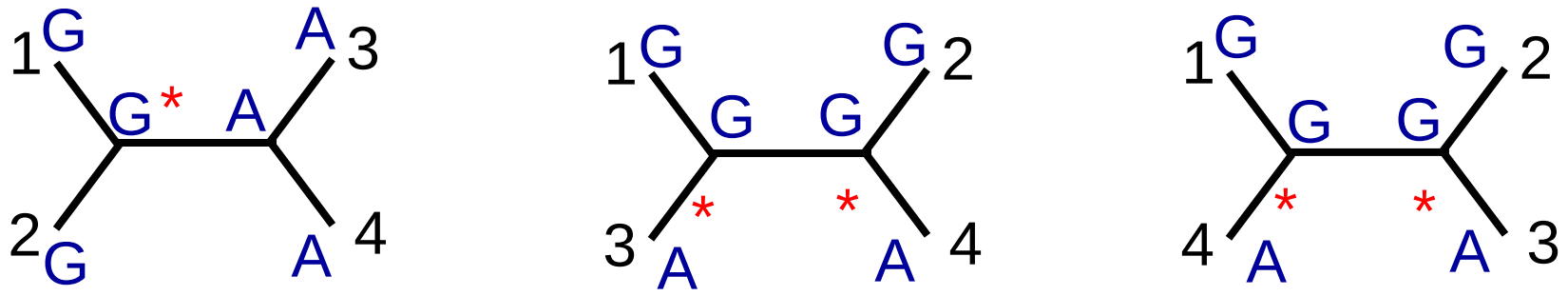


Also uninformative – need three mutations in each tree

Parsimony

Position

Sequence	1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

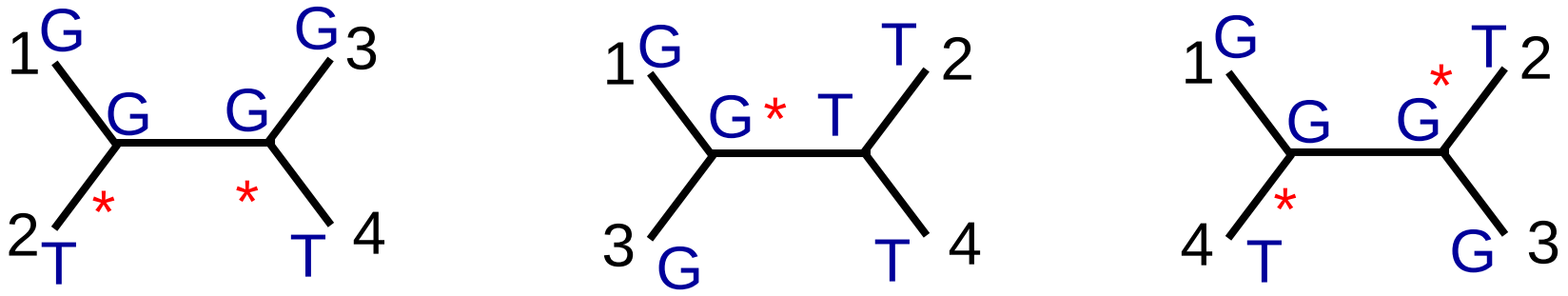


Informative! – need only one mutation in one tree but two In the other trees!

Parsimony

Position

Sequence	1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T



Informative! – need only one mutation in one tree but two In the other trees!

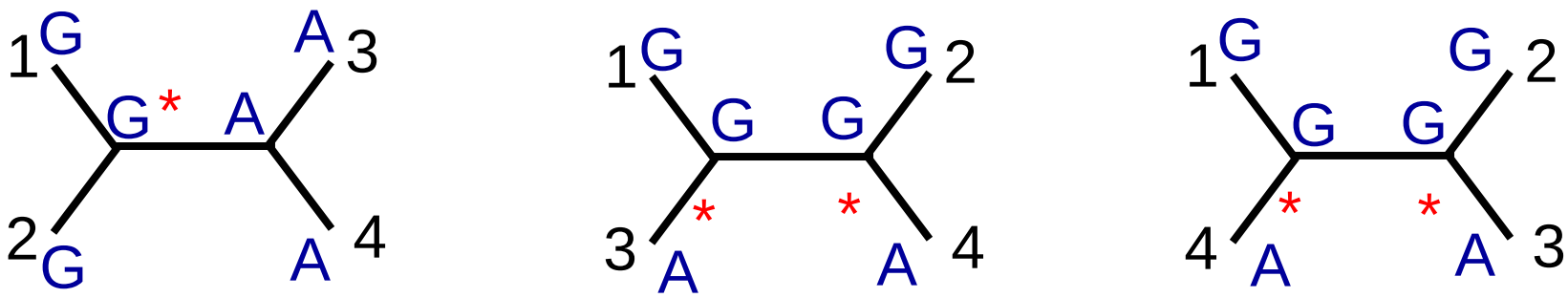
Parsimony

Sequence	Position 1	2	3	4	5	6
1	G	G	G	G	G	G
2	G	G	G	A	G	T
3	G	G	A	T	A	G
4	G	A	T	C	A	T

So to be Informative, need at least 2 different nucleotides And each has to be present at least twice.

Every tree is considered for every site, maintaining a running score of the number of mutations required. The tree with the smallest number of invoked mutations is the most parsimonious

Parsimony



Mathematically, most likely candidate nts at an Internal node are:

{descendent node 1} Ω {descendant node 2}

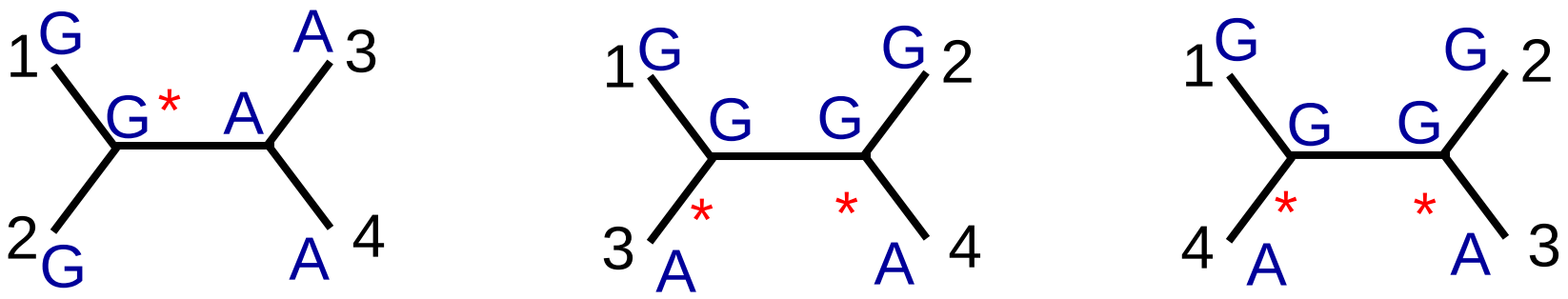
IF this is null set, then most likely candidate nts are:

{descendent node 1} U {descendant node 2}

Σ ***U*** = minimum number of substitutions required to account for nts at terminal nodes since they last shared common ancestor

Total number of substitutions, informative + uninformative = tree length

Parsimony



If you use parsimony but weigh the mutations by some kind of scoring system that accounts for the likelihood of each mutation →weighted parsimony

By-product of parsimony is inference of nt identity in the ancestral sequence

Parsimony

10 sequences: > 2 million possible trees…
Need a better way…

Branch and bound (Hardy and Penny, 1982)

Step 1: Determine an upper bound to the length of the most parsimonious tree = L - either chosen randomly, or else using a computationally fast way like UPGMA

Step 2: Starting from a simple tree with just some of the Sequences, grow trees incrementally by adding branches to a smaller tree that describes just some of the sequences.

Step 3: If at any point, the number of required substitutions is > L, abandon that tree.

Step 4: As soon as you get a tree with fewer substitutions than L, use that tree as the new upper bound to make remainder of the search even more efficient.

Works for <= 20 sequences

Parsimony

For > 20 sequences

Heuristic Searches

Assumption: Alternative trees are not all independent of each other. Most parsimonious trees have similar topologies.

Step 1: Construct an initial tree as a good guess: UPGMA, and use it as a starting point.

Step 2: Branch-swap subtrees and graft them onto the starting tree, keeping overall topology. See how many are shorter than the starting tree. The prune and re-graft, and see if it keeps getting better.

Step 3: Repeat until a round of branch swapping fails to generate any better trees

Parsimony

Often get tens or hundreds of equally parsimonious trees

Build a consensus tree – any internal node supported by At least half the trees becomes a simple bifurcation.

Phylogenetic Software

PHYLIP: Phylogenetics Inference Package
free at http://evolution.genetics.washington.edu

Includes many programs including various distance methods, maximum likelihood, parsimony, with many of the options we've discussed.

PAUP: Phylogenetic Analysis Using Parsimony
http://www.lms.si.edu/PAUP - NOT FREE
Now includes maximum likelihood and distance methods as well

Tree of Life

Carl Woese and colleagues, 1970s

Used 16S rRNA – all organisms possess.

Found 3 major evolutionary groups:
Bacteria
Eucarya
Archea (including thermophilic bacteria

Human Origins:
mt DNA sequences – huan populations differ by ~ 0.33% (very small). Greatest differences NOT between current populations on different continents, but between human populations residing in Africa – “out of Africa” theory

Mitochondrial “eve” and Y-chromosome “adam”