Bioinformatics Algorithms
COS-BIOL-530/630
Lecture05

Days & Times	Room	Meeting Dates
Tu 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025
Th 2:00PM - 3:50PM	Thomas Gosnell Hall (GOS)-2178	01/13/2025 - 04/28/2025

Instructor:
Fernando Rodriguez
email: frvsbi@rit.edu
Office: Orange Hall 1311

Phylogenetics
- Lecture04 & Lecture05 -

Announcements

Week 4
Lecture04
Lab04
- Discussion 4
- Activity 4

Week 5
Lecture05
Lab05
Activity 5
Discussion 5

Quiz 4 (week 4-5)

Perusal: Rodriguez *et al*. 2009

Phylogenetics
- Lecture04 & Lecture05 -

Announcements

Oedipus is still down.

Run PHYLIP locally on your computer!

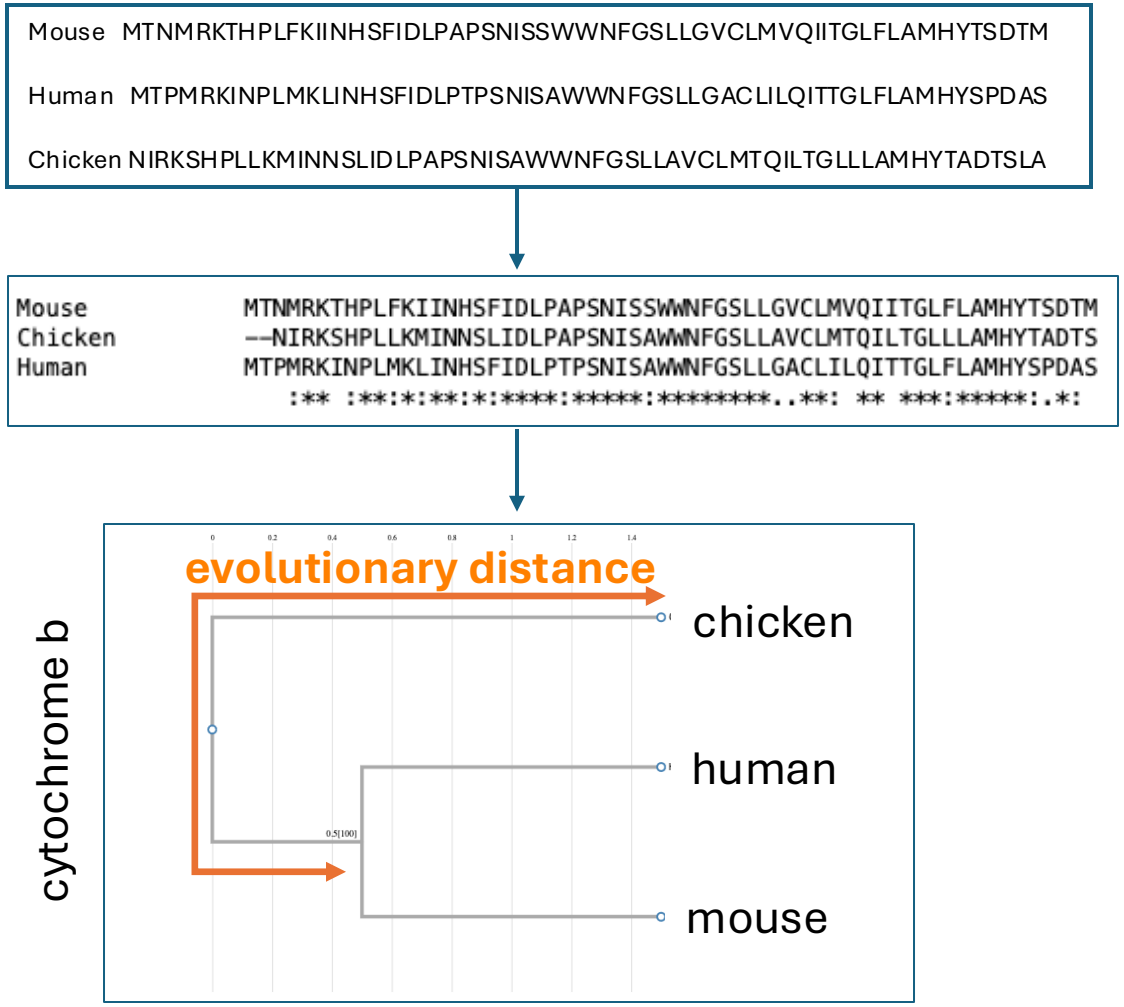
In MyCourses > Content > Labs
- Lab04_Rup_cytb.aln.phy
- Lab04_Rup_cytb_outgroup.aln.phy

Exam 2: Thursday, April 17th

Phylogenetics
- Lecture04 & Lecture05 -

Topics:

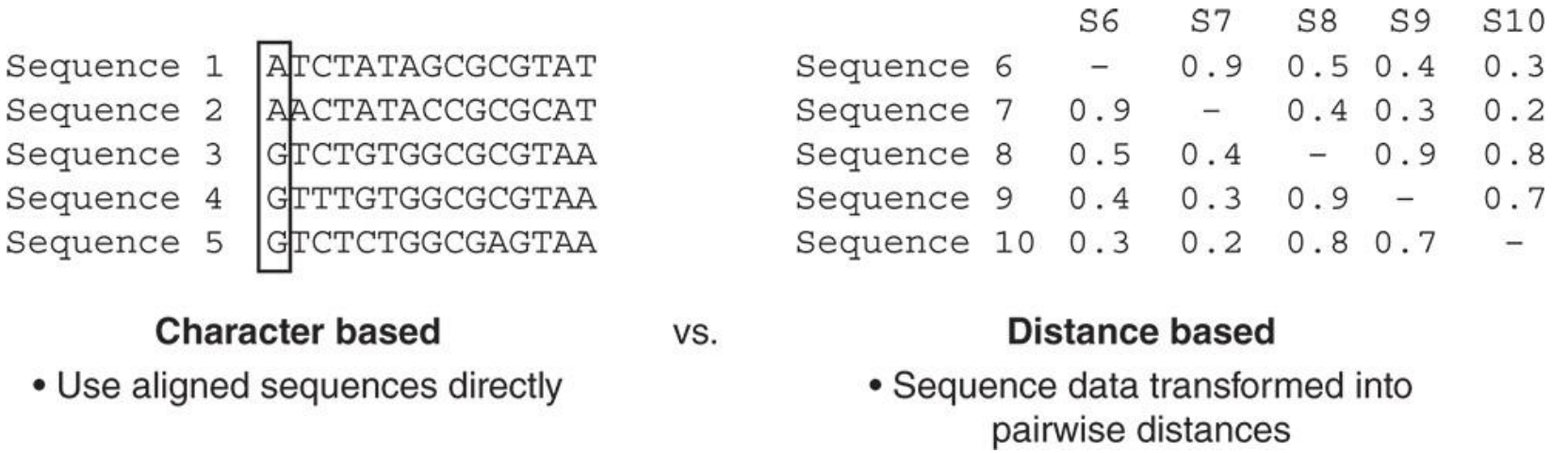
- Substitution matrices
- Phylogenetic trees
  - Distance based
  - Character based
- Bootstrap
- Consensus tree
- Labs 04 &05: PHYLIP!



Tree Building Methods

- Designed to fit species into related branches and nodes, based on evolutionary models.
- Tree-building methods can be sorted into
  - **Distance-based**: transform the sequence data into pairwise distances (calculated values which link the most similar sequences together); they then use these derived values rather than the characters directly to build trees. They are much less computationally intensive.
  - **Character-based** methods: use the aligned sequences directly during tree building . They are much more computationally intensive.

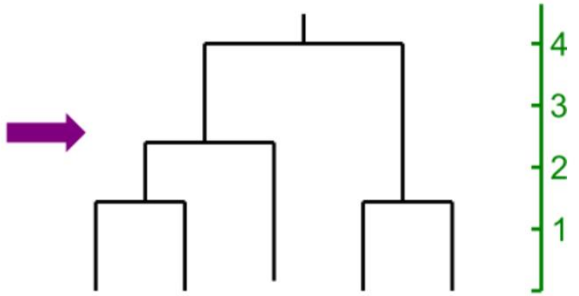
Source: *Bioinformatics*, by Baxevanis



Construction Phylogenetic Trees

UPGMA

	A	B	C	D	E
A	0	8	8	5	3
B		0	3	8	8
C			0	8	8
D				0	5
E					0

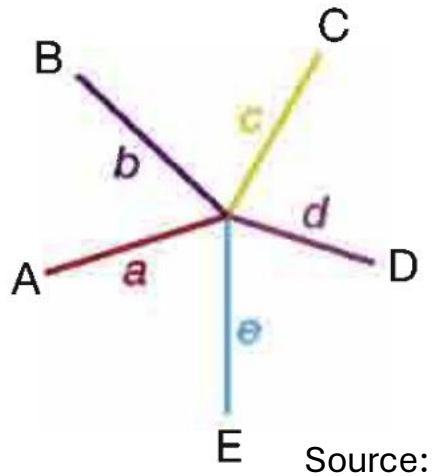


- Once we have a distance matrix, there are several ways to generate a tree.
- Unweighted Pair Group Method using arithmetic Averages (UPGMA):
  - UPGMA assumes that all lineages evolve at the same rate (per the molecular clock hypothesis).
  - It creates a tree where all leaves are equidistant from the root.
  - If the lineages evolve at different rates (which they do in reality), the UPGMA tree may not fit the distance data well.

Construction Phylogenetic Trees

Neighbor Joining

- Neighbor Joining starts with a “star-like” tree.



Source: *Bioinformatics,* Mount.

- All terminal nodes are joined to a single internal node
- The tree is then modified, splitting off neighbors until all nodes are bifurcating.

Construction Phylogenetic Trees

UPGMA

Neighbor Joining

Construction Phylogenetic Trees

UPGMA

- Distance - based method.
- Construct a tree by iteratively joining subtrees
- Assumes that all lineages evolve at the same rate.
- Produces ”midpoint” rooted trees.
- Ultrametric property, and so often not the best choice.

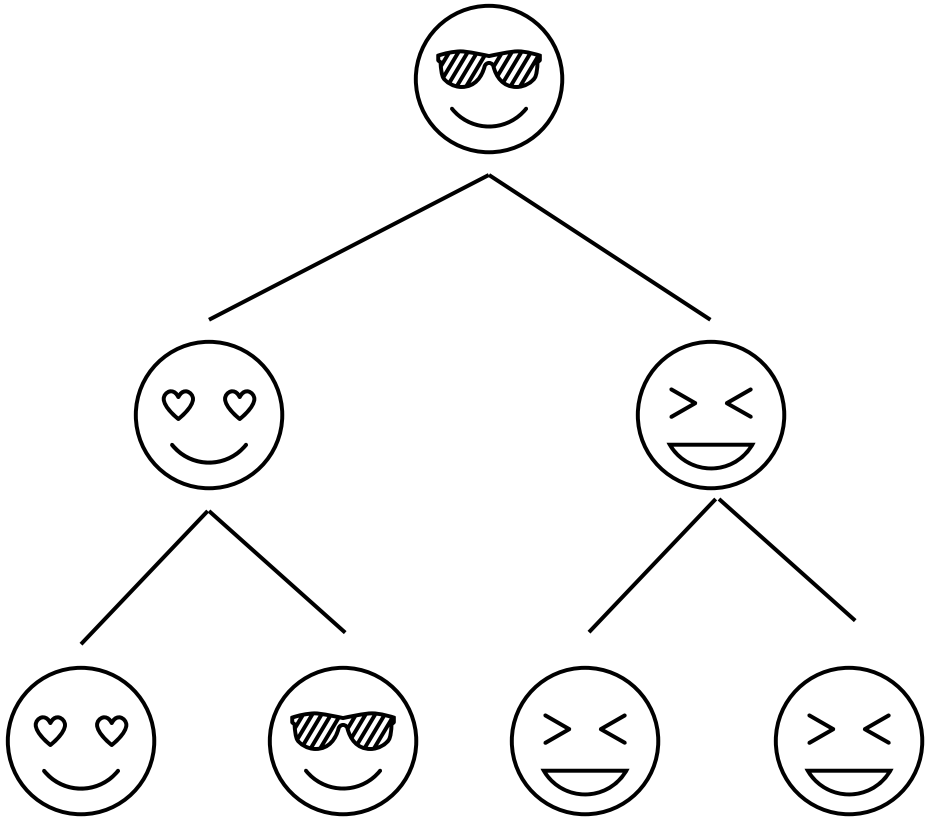
Neighbor Joining

- Distance-based method.
- Construct a tree by iteratively joining subtrees.
- But doesn’t make molecular clock assumption.
- Produces unrooted trees. An unrooted tree can be rooted with an outgroup.
- Distance of clusters from each other is solved geometrically from component distances, assuming the data came from a tree.

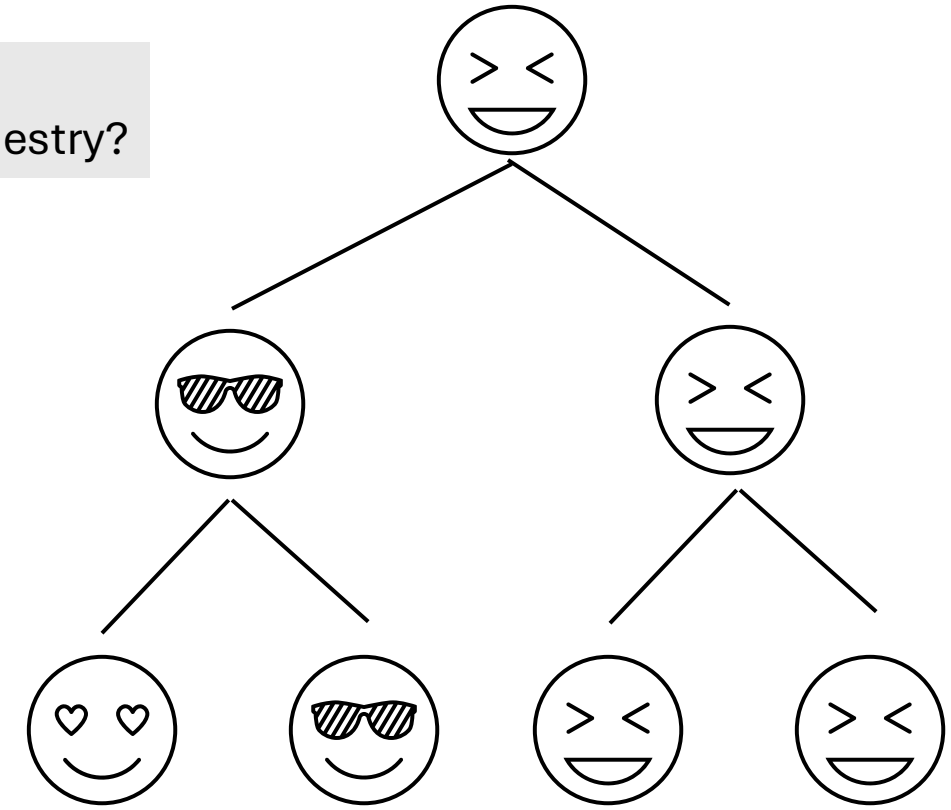
Character-based Methods
Parsimony

- Distance-based tree methods work well in general.
- But it is all reduced to measure distance.
- And some significant information can be thrown away.
- Parsimony-bases methods don’t use distance matrices:
    - Character-based
    - Try to explain the data with a minimal number of changes
    - Focus on finding the right tree topology, not estimating branch lengths

Parsimony

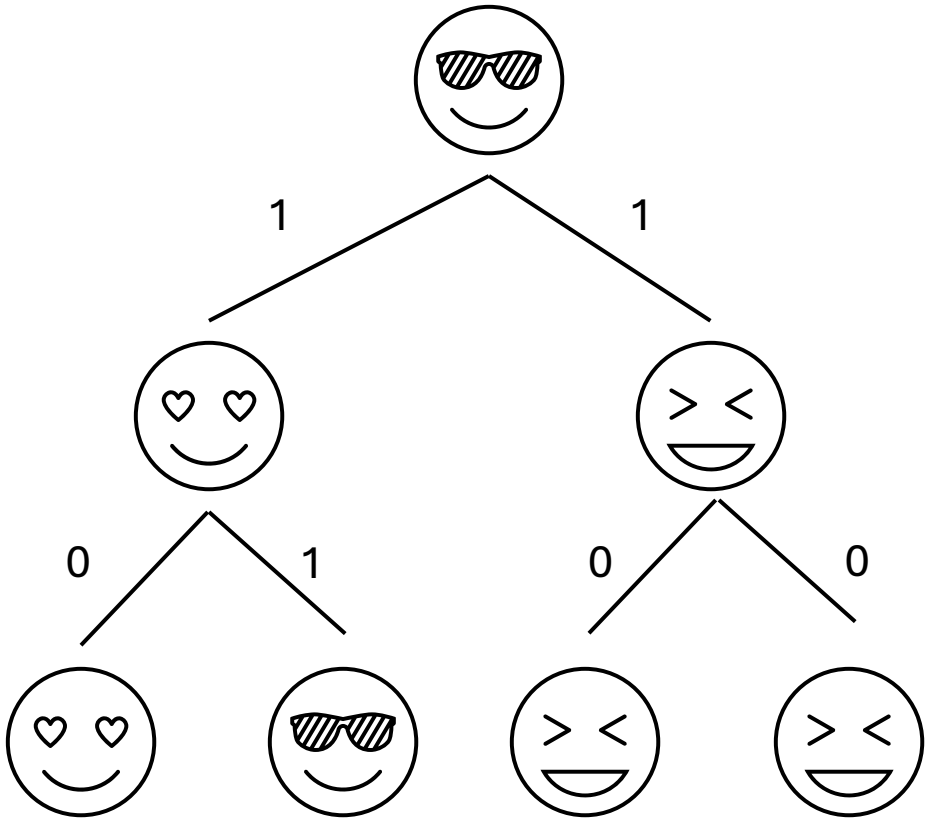


Which tree better explains their ancestry?



Parsimony principle:
If all things are equal, the simplest possible explanation is the best.

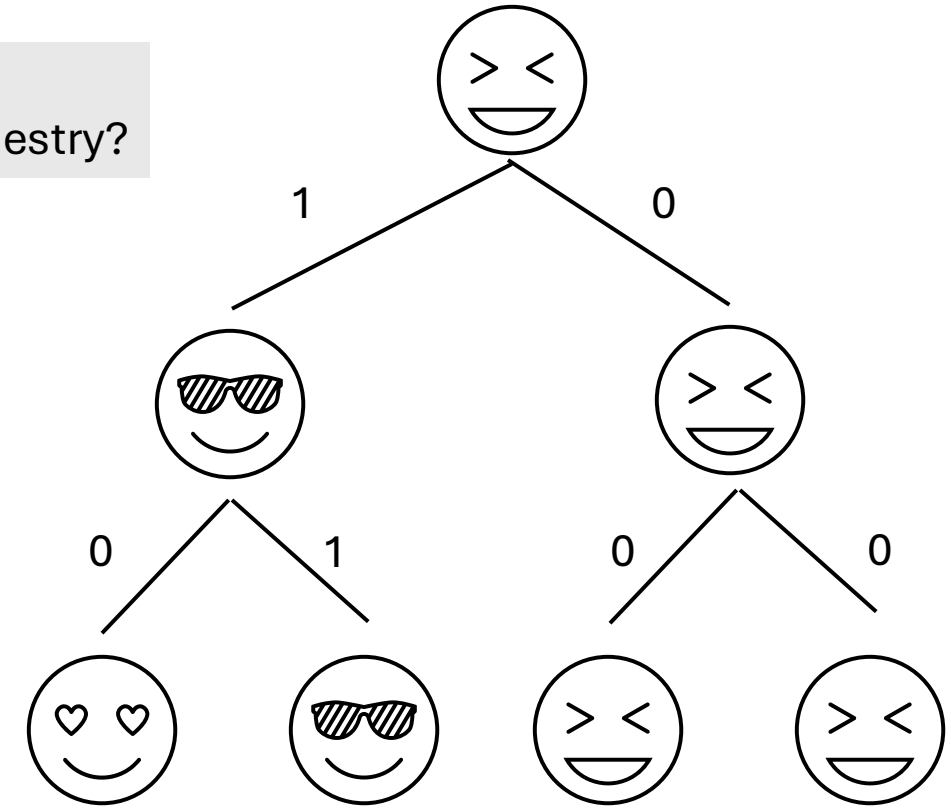
Parsimony



Parsimony score: 3

Which tree better explains their ancestry?

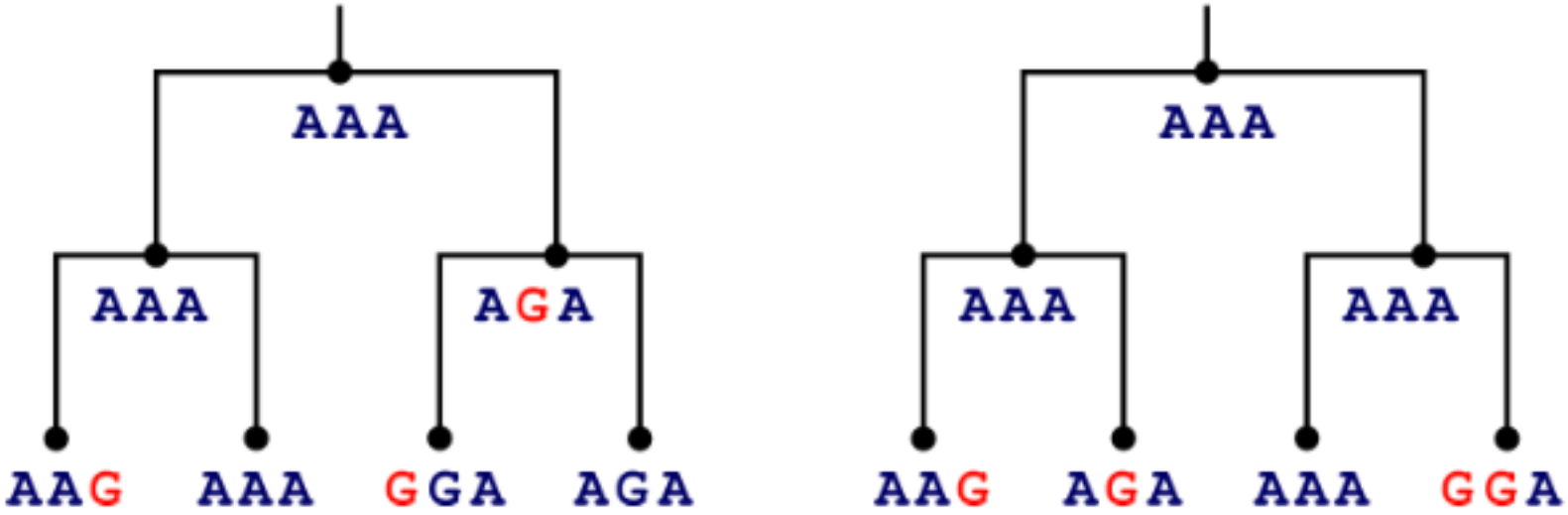
Both trees could be correct. But parsimony says the one with the fewer changes (right) is the more likely tree.



Parsimony score: 2

Parsimony

Which tree is more parsimonious?



In parsimony:

- Every feature (nucleotide/amino acid) is independent of all the others.
- Problems can be split into components.

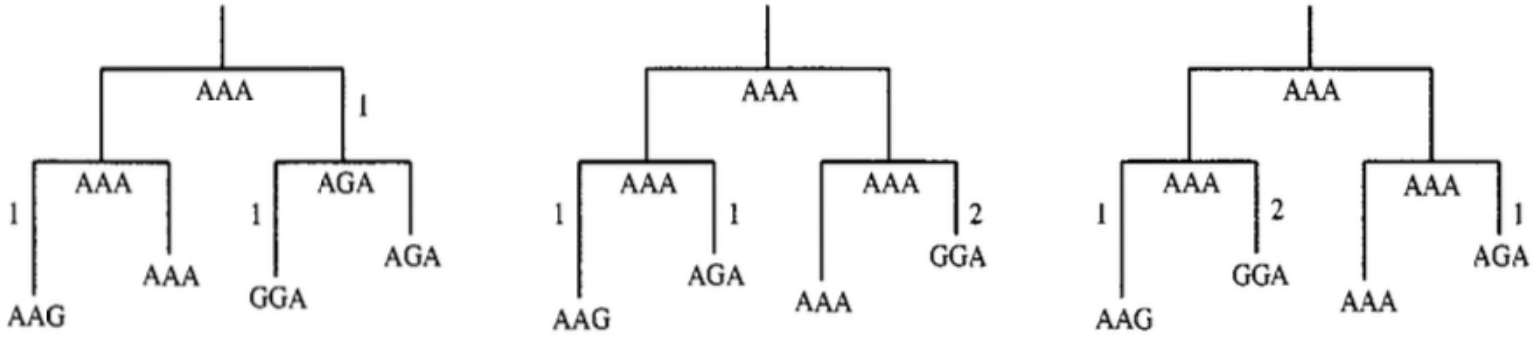
Parsimony approach

Parsimony approaches usually have three different components:

1. Search through many different tree topologies to find the best.
2. Note the number of substitution events by each possible tree.
3. Find the minimum number of changes needed to explain the data, for a given tree topology. Find the most parsimonious tree.

Sequences:

AAG
AAA
GGA
AGA

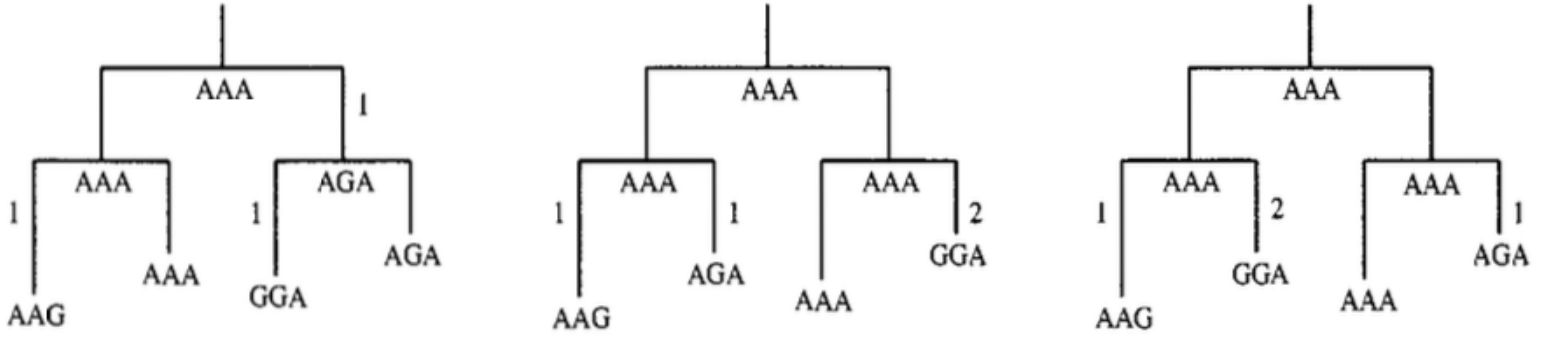


Maximum Parsimony (MP)

- In terms of tree building, the MP method requires the fewest evolutionary changes or the fewest number of character-state changes. All of the character data are used in the analysis.
- However, no branch lengths are calculated while the relationships between sequences are determined.
- Although scoring a phylogenetic tree by counting the number of character-state changes is easy, there is no algorithm to generate the most parsimonious tree quickly.
- The most parsimonious tree must be found in what is commonly referred to as “tree space,” meaning among all possible trees.
- MP analyses tend to yield numerous trees, which have the same score but different topologies.

Sequences:

AAG
AAA
GGA
AGA

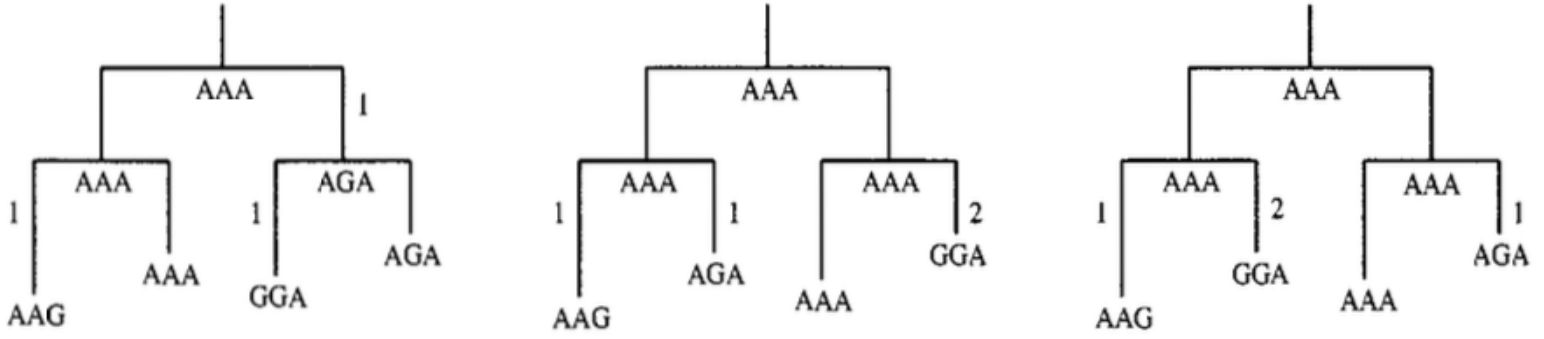


Maximum Parsimony (MP)

- The tree with the topology containing the most nodes in consensus with all equally likely trees is considered to be the one that best supports the data.
- It is possible to do an exhaustive search in which every possible tree is scored, and the best one is then selected when a small number of taxa are considered.
- For greater numbers of taxa, a **heuristic** search that involves finding an **approximate solution** when an exact solution is not feasible must be performed.

Sequences:

AAG
AAA
GGA
AGA



Parsimony Informative sites

- For a nucleotide to be useful, it needs to provide information that can differentiate between possible trees.
- **Informative**: can differentiate between possible trees.
- **Uninformative**: cannot differentiate between possible trees.

Site	1	2	3	4
Species A	A	T	C	C
Species B	A	C	C	C
Species C	A	T	C	A
Species D	A	C	A	A

Parsimony-informative sites:

- 2 and 4 are informative
- 1 is uninformative (invariable)
- 3 is uninformative (singleton)

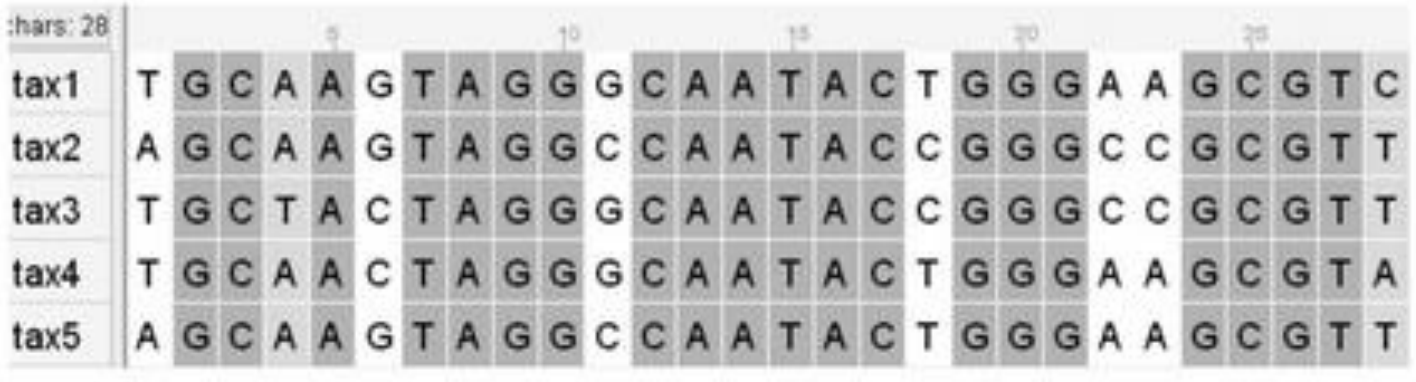
Uninformative sites

- A nucleotide which is the same in all sequences is termed **invariant**.
- If only varies in ONE sequence:
  - It does not cluster with any other sequence.
  - All possible trees require the same number of mutations.
- Three or more states, only one of which is represented more than once.

Informative sites

Any informative site must:

- Have at least two different nucleotides.
- Each nucleotide has to be present at least twice (different taxa)



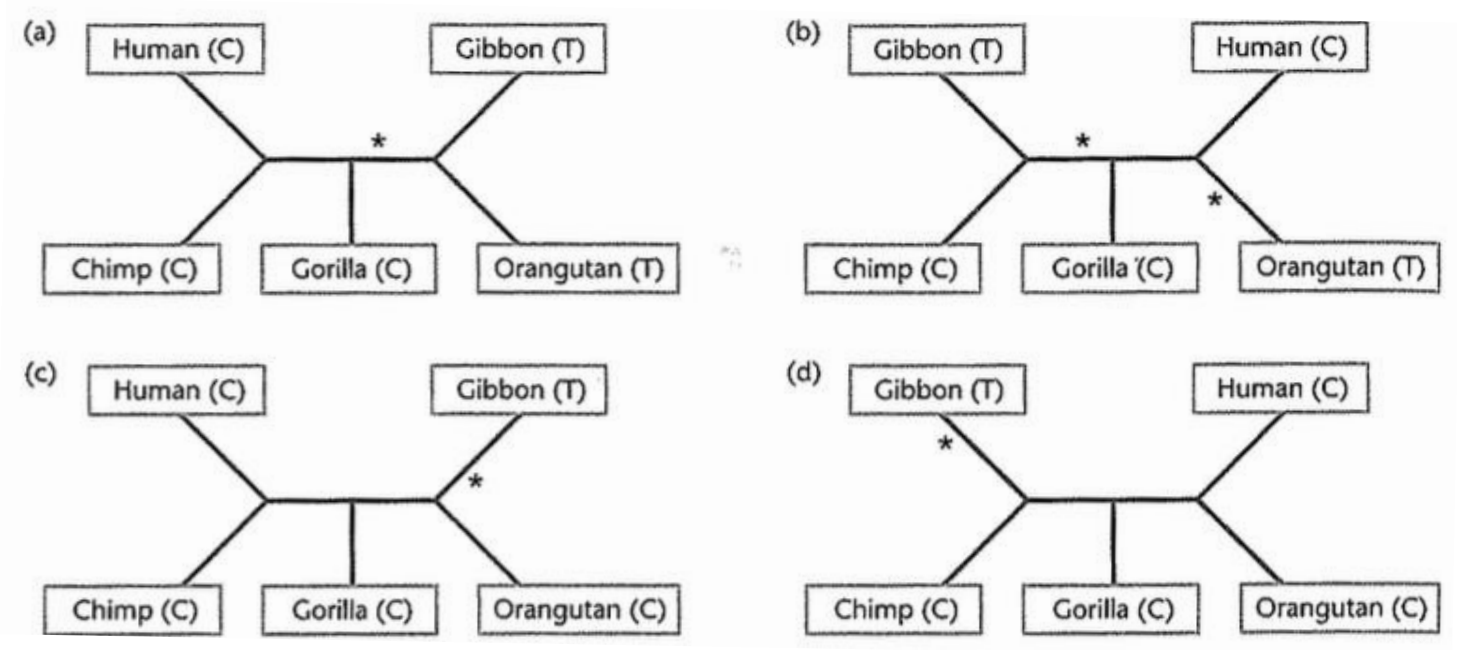
chars: 28																										
					5					10					15					20					25		
tax1	T	G	C	A	A	G	T	A	G	G	G	C	A	A	T	A	C	T	G	G	G	A	A	G	C	G	T	C
tax2	A	G	C	A	A	G	T	A	G	G	C	C	A	A	T	A	C	C	G	G	G	C	C	G	C	G	T	T
tax3	T	G	C	T	A	C	T	A	G	G	G	C	A	A	T	A	C	C	G	G	G	C	C	G	C	G	T	T
tax4	T	G	C	A	A	C	T	A	G	G	G	C	A	A	T	A	C	T	G	G	G	A	A	G	C	G	T	A
tax5	A	G	C	A	A	G	T	A	G	G	C	C	A	A	T	A	C	T	G	G	G	A	A	G	C	G	T	T

- >= 2 states, >= 2 of which occur in >=2 taxa

Informative sites

By parsimony criterion:

- (a) tree is better than (b) according to this informative site.
- Parsimony does not distinguish between trees (c) and (d) as this is not informative.

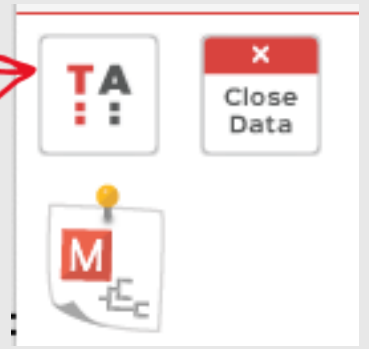


Source: *Bioinformatics and Molecular Evolution*, Higgs and Attwood

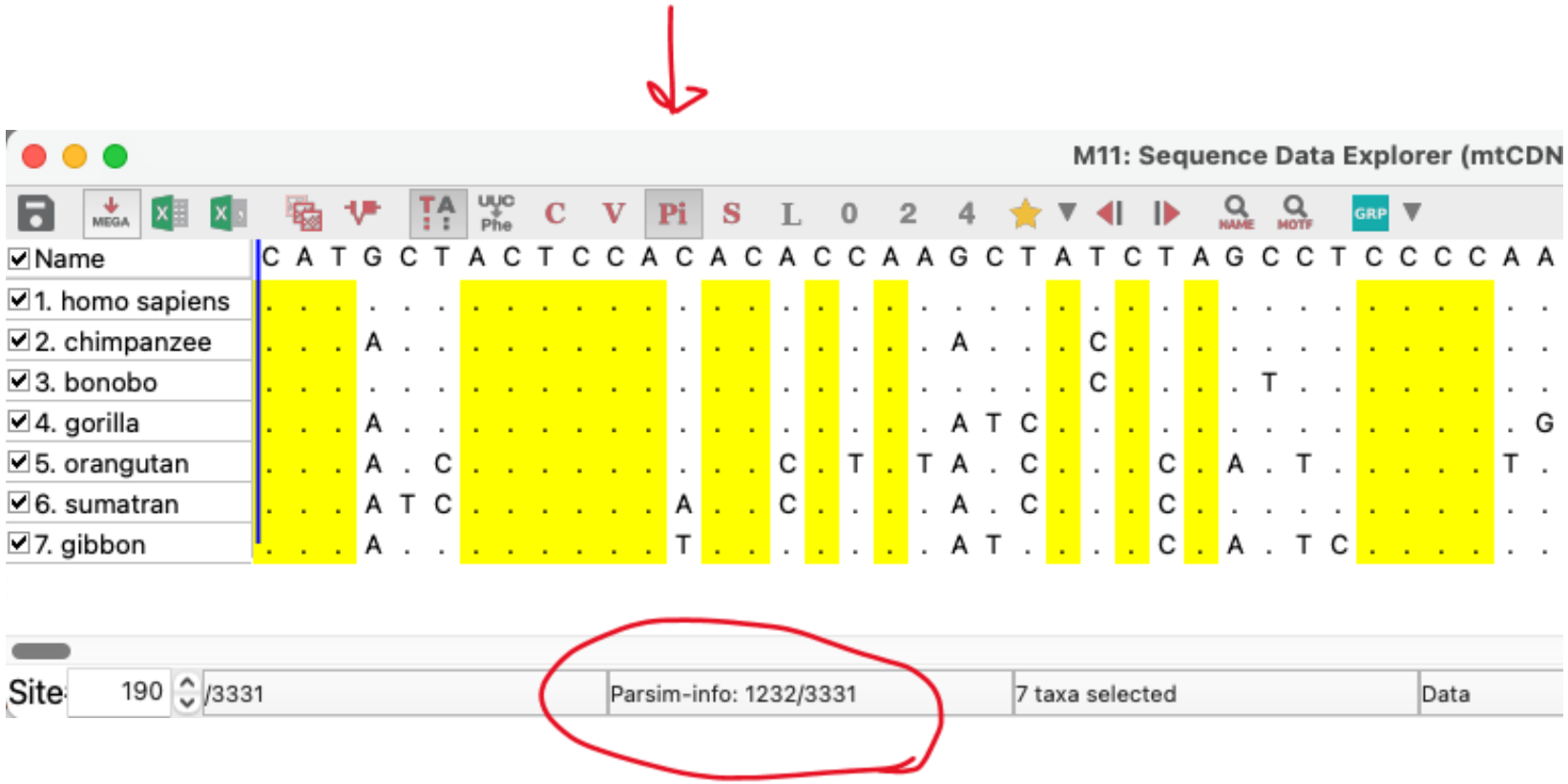
Small in-class activity!!

MEGA: Parsimony analysis / Informative sites / Parsimony tree

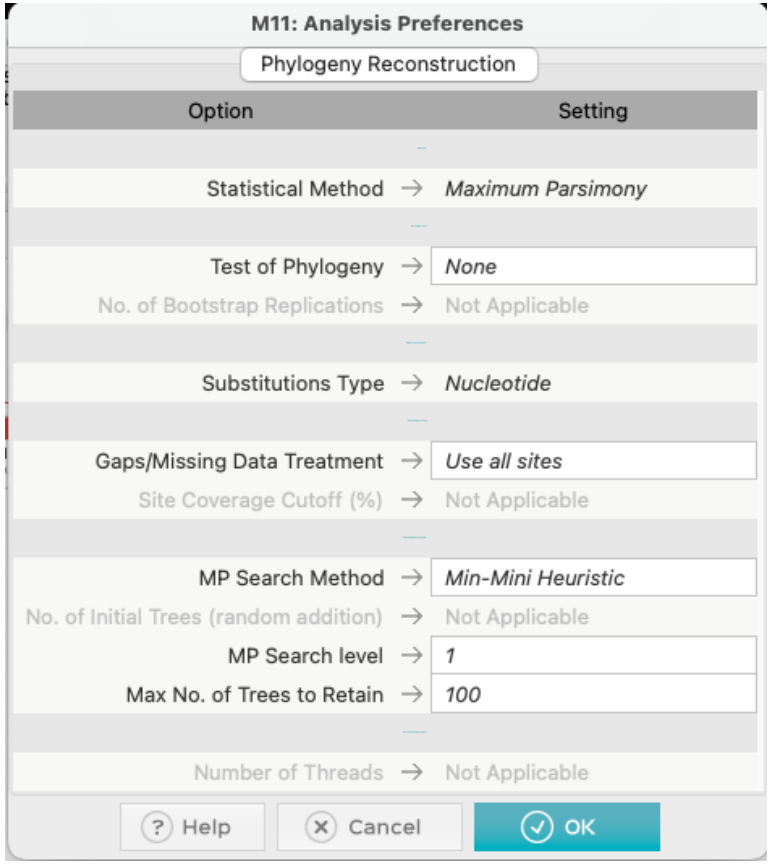
- Open MEGA, under “Analyze” Mode.
- Click on Examples, and open mtCDNA.meg
- Open Data explorer
- Explore the options:
  - C: conserved sites
  - V: variable sites
  - Pi: Parsimony informative
  - S: singleton
  - What are the proportions for each?

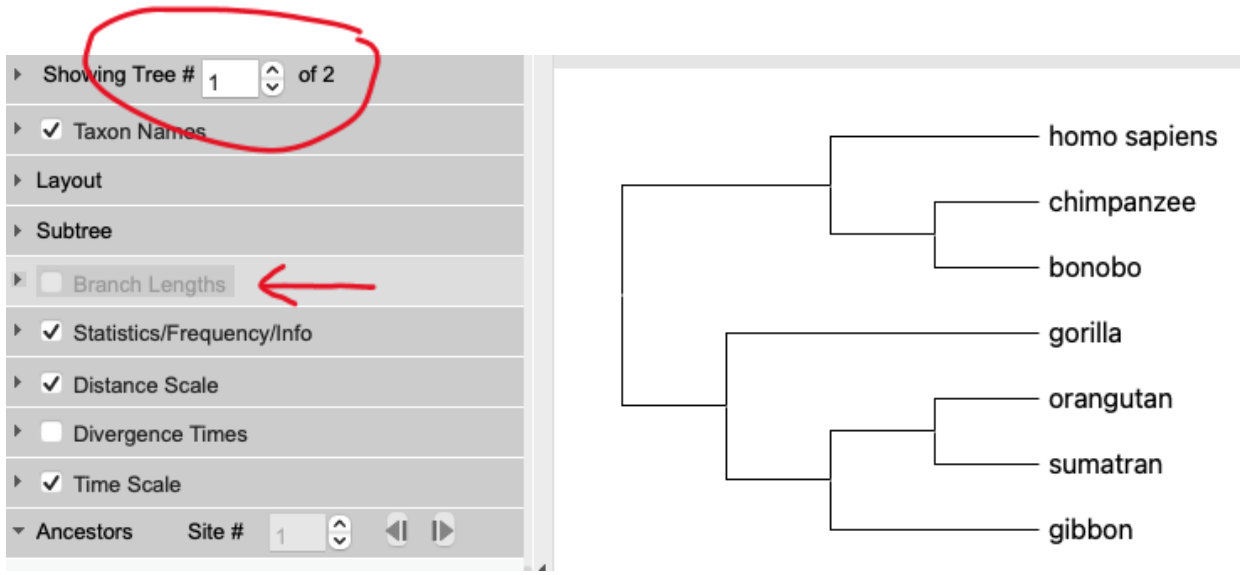


Small in-class activity!!



Small in-class activity!!





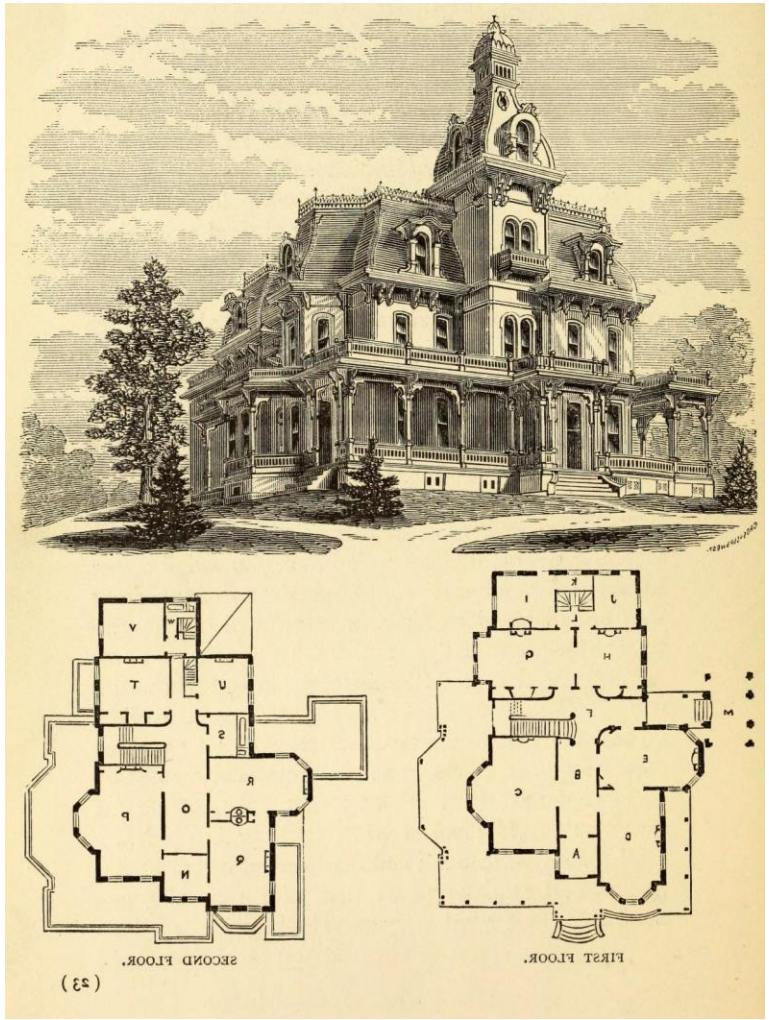
Parsimony problem

- If we are given a tree a priory, finding the most parsimonious ancestor is ***easy***.
    - But in real life, we are not given a tree.
    - There is no way to deduce a tree from the taxa.
    - Only solution: try many different trees.

Number of Leaves	Unrooted trees	Rooted trees
2	1	1
4	3	15
6	105	945
8	10,395	135,135
10	2,027,025	34,459,425

    - Not the best solution.

Parsimony problem



- We have to search for a good tree:
  - Consider one tree at a time.
  - After each tree, find a small change to give a better answer.
  - Stop when you think you have the best tree.
- This is a heuristic approach:
  - The best tree is not guaranteed.
  - But it cuts time, and we are reasonably sure we have a good (not best) tree.

Searching for Parsimonious Trees

Here are the steps to the algorithm:

1. Start with a tree.

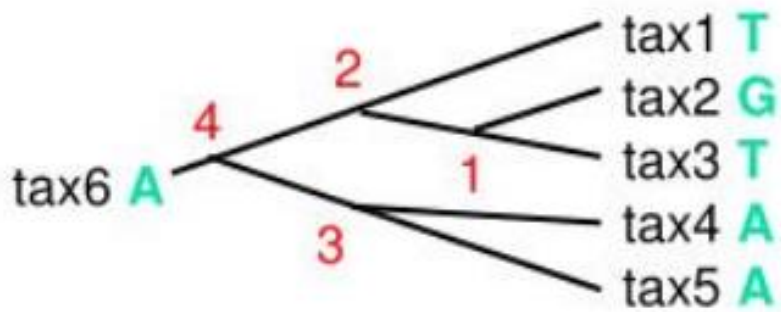
2. Do *small* parsimony on it.

3. Keep doing:
   - Do small parsimony on all neighbors of the current tree.
   - Chose neighbor with best cost, and switch to that one.
   - If the current tree had a lower cost than all its neighbors, stop the loop.

4. Current tree is the best one found so far.

But where do you start the search?
There are ways to shortcut the process: **Branch and Bound** or **Pruning**.

Calculating a tree length



1) $\{T\} \cap \{G\} = \{\} \rightarrow S_1 = \{T, G\}$

2) $\{T, G\} \cap \{T\} = \{T\} \neq \{\} \rightarrow S_2 = \{T\}$

3) $\{A\} \cap \{A\} = \{A\} \neq \{\} \rightarrow S_3 = \{A\}$

4) $\{T\} \cap \{A\} = \{\} \rightarrow S_1 = \{T, A\}$

L:
+ 1
+ 0
+ 0
+ 1

= 2

Exhaustive search – brute force

- Examine every possible alternative to find one particular solution.
- These are the easiest algorithms to design and understand.
- Sometimes they work acceptably for certain practical problems in biology.
- But brute force algorithms are too slow to be practical for anything.

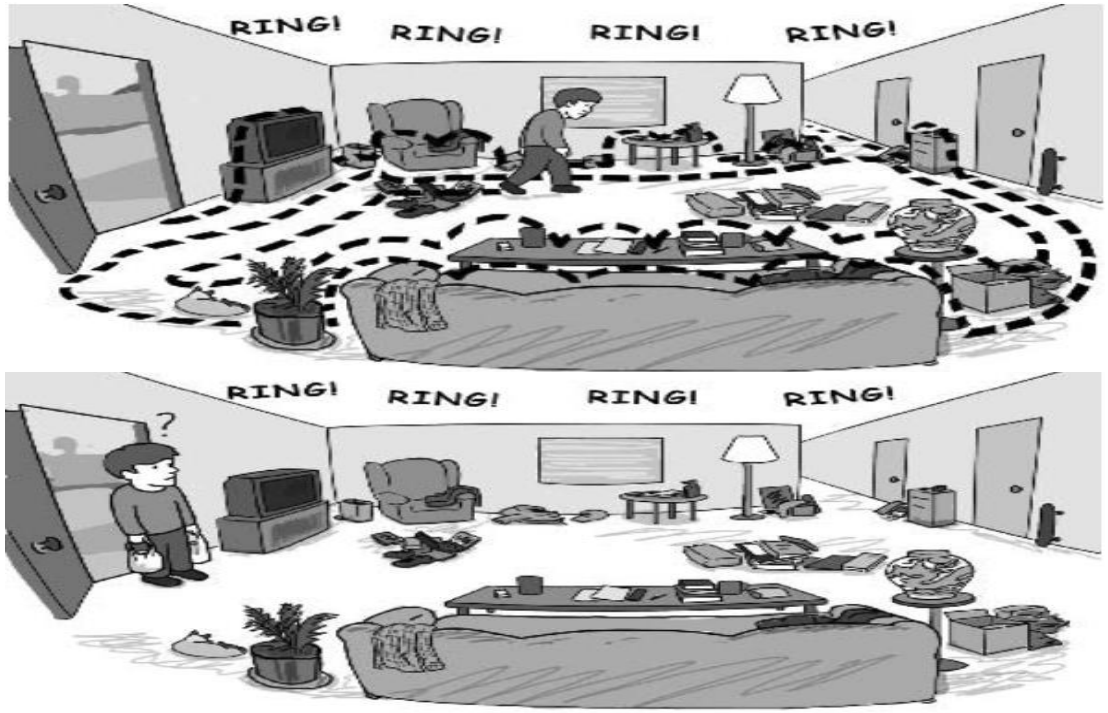
Source: *Introduction to bioinformatics algorithms*, Jones and Pevzner

Branch and Bound algorithms

- Omit a large number of alternatives when performing brute force.
- Suppose you are exhaustively searching the first floor and hear the phone ringing above your head.

second floor

first floor



Half the time!!

Branch and Bound

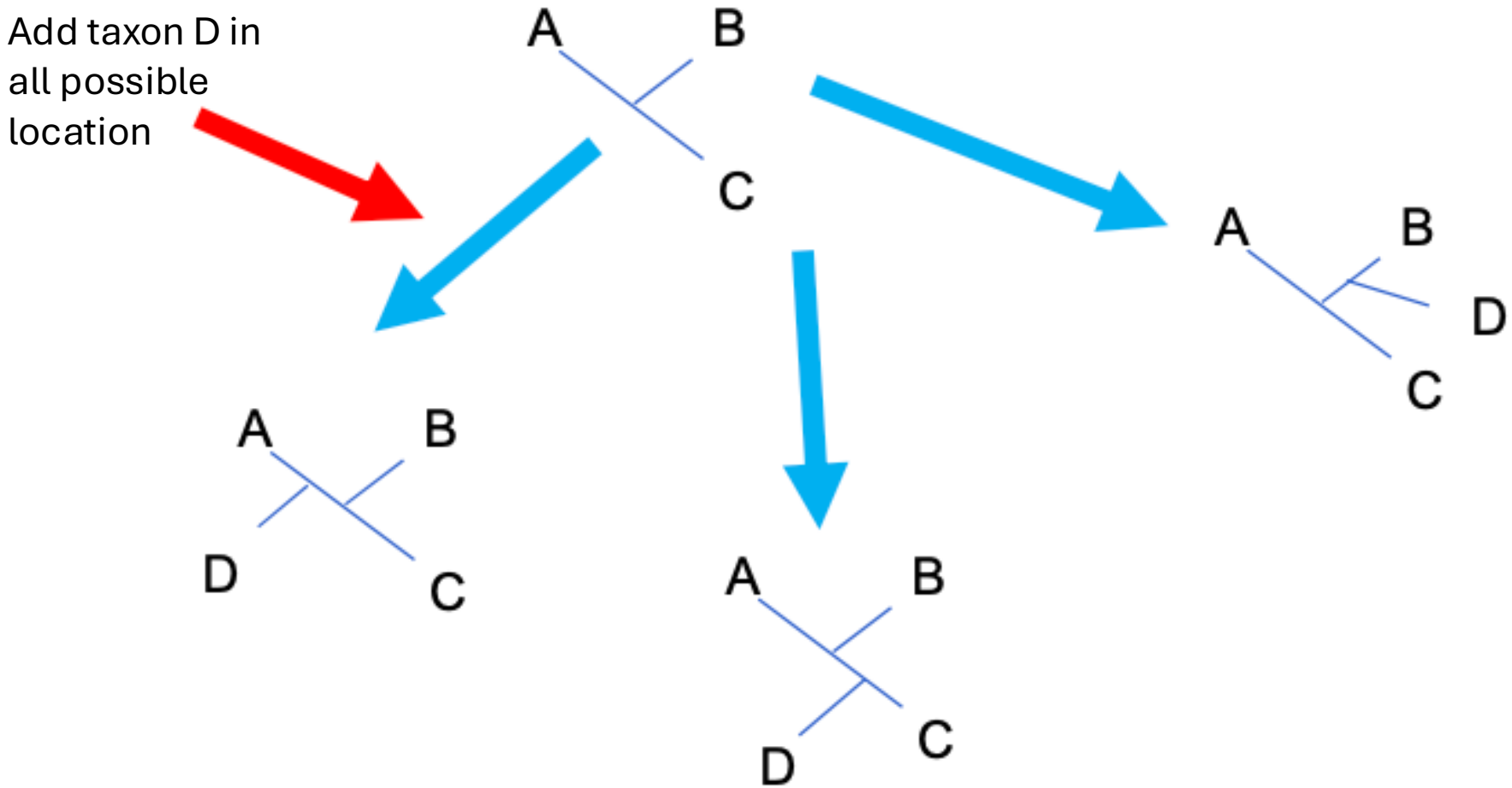
As the tree gets bigger (adding more segment between nodes), its **cost** (Parsimony score) will only increase.

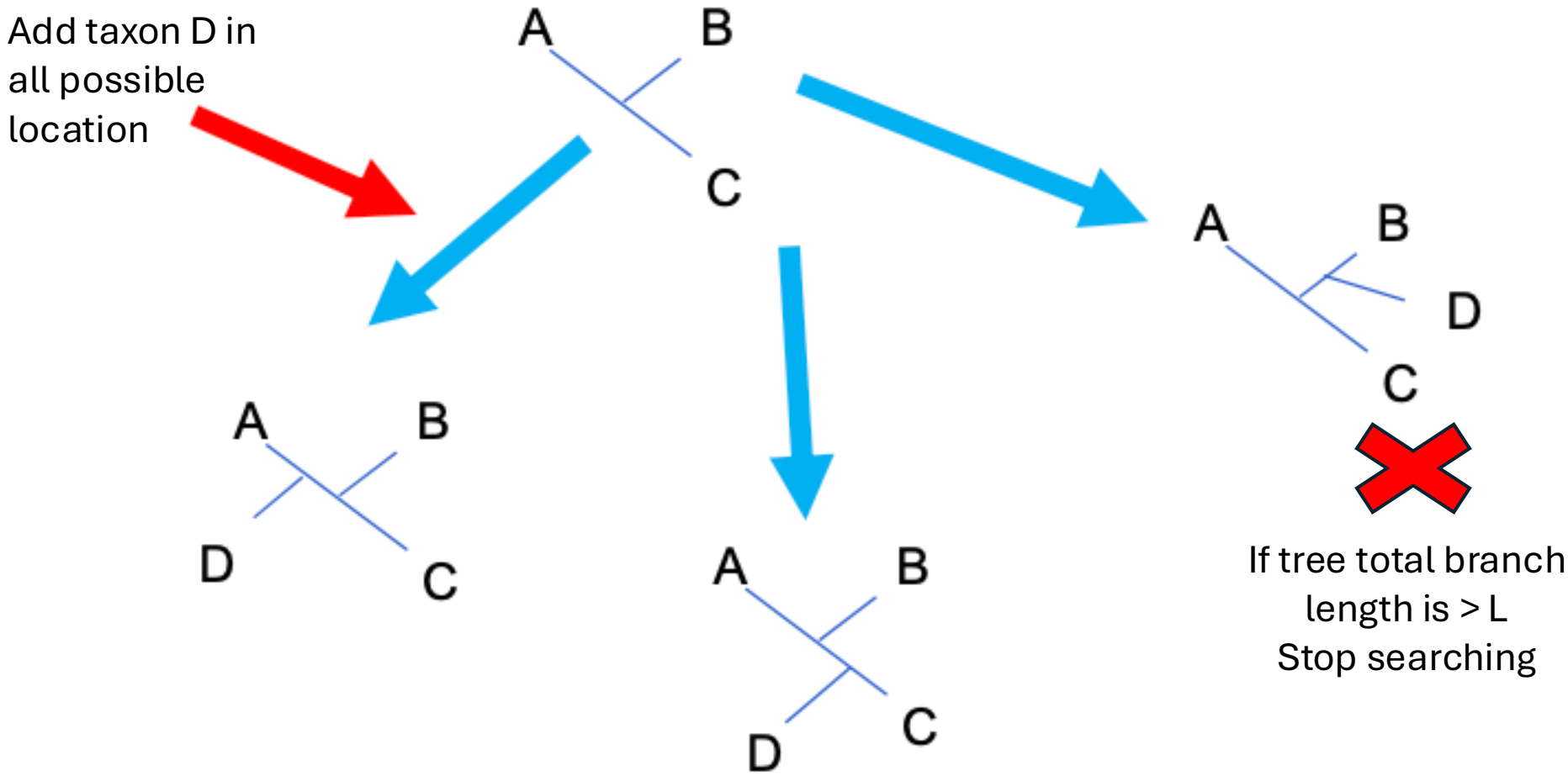
- No segment has a negative value.
- If we add a segment, it cost will either go up or stay the same.
- If a subtree already costs more than some complete tree, you don’t need to look any further.
- You can exclude many trees from the search in a single step.

Branch and Bound

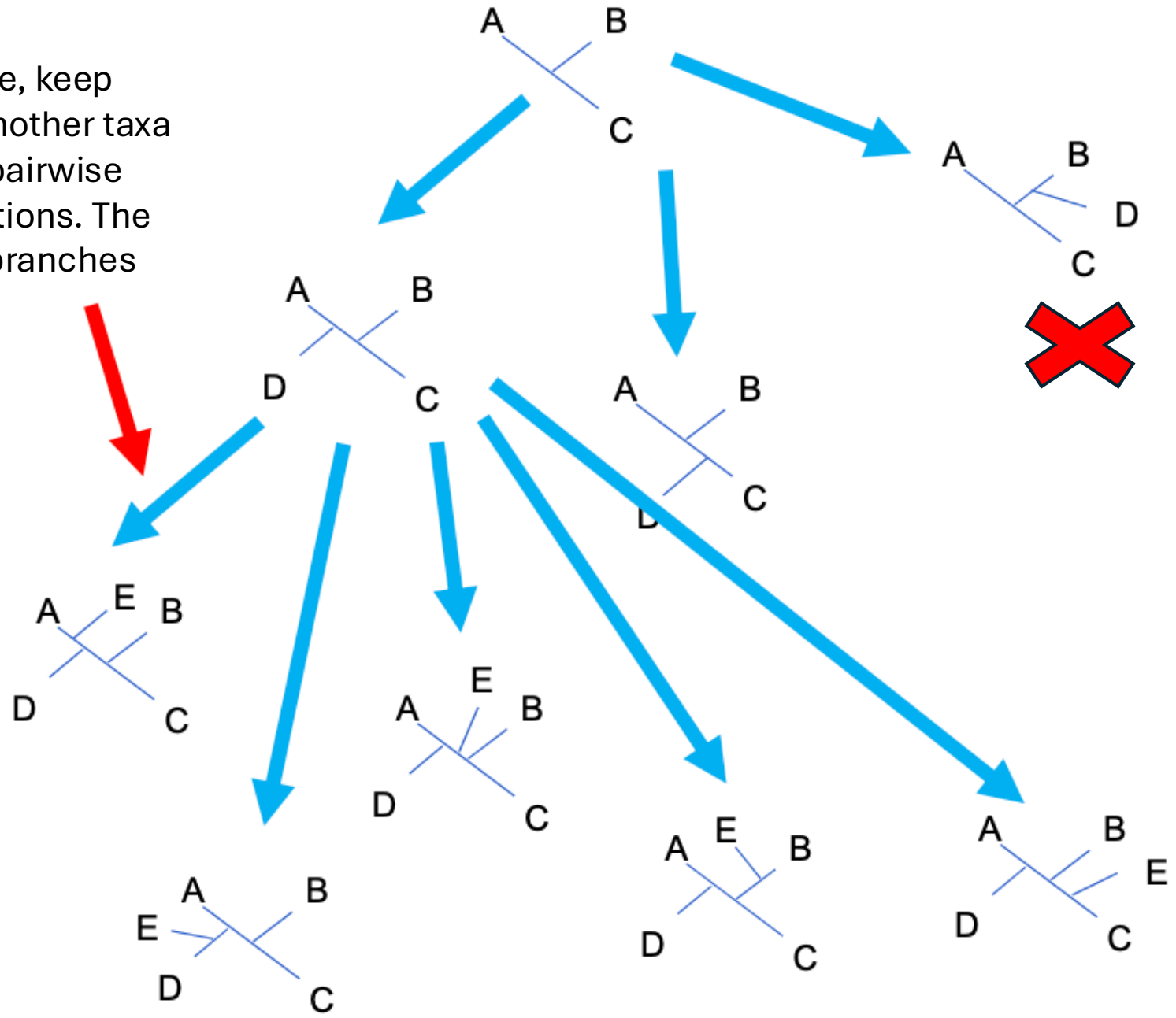
- Intelligent elimination process.
- It uses two main steps to reduce computing search.
- Start by determining the upper bound or cost (***L***) to the most parsimonious tree.
- We can initialize ***L*** by creating a quick tree (UPGMA, NJ) and taking ***L*** as the total branch length of that tree.
- The second step is to build up the tree from three terminal nodes gradually.
- Add a taxon to the three in all possible combinations.
- If the total branch length is >= ***L***, then adding more branches will only exceeds ***L***!


Start here
A
B
C

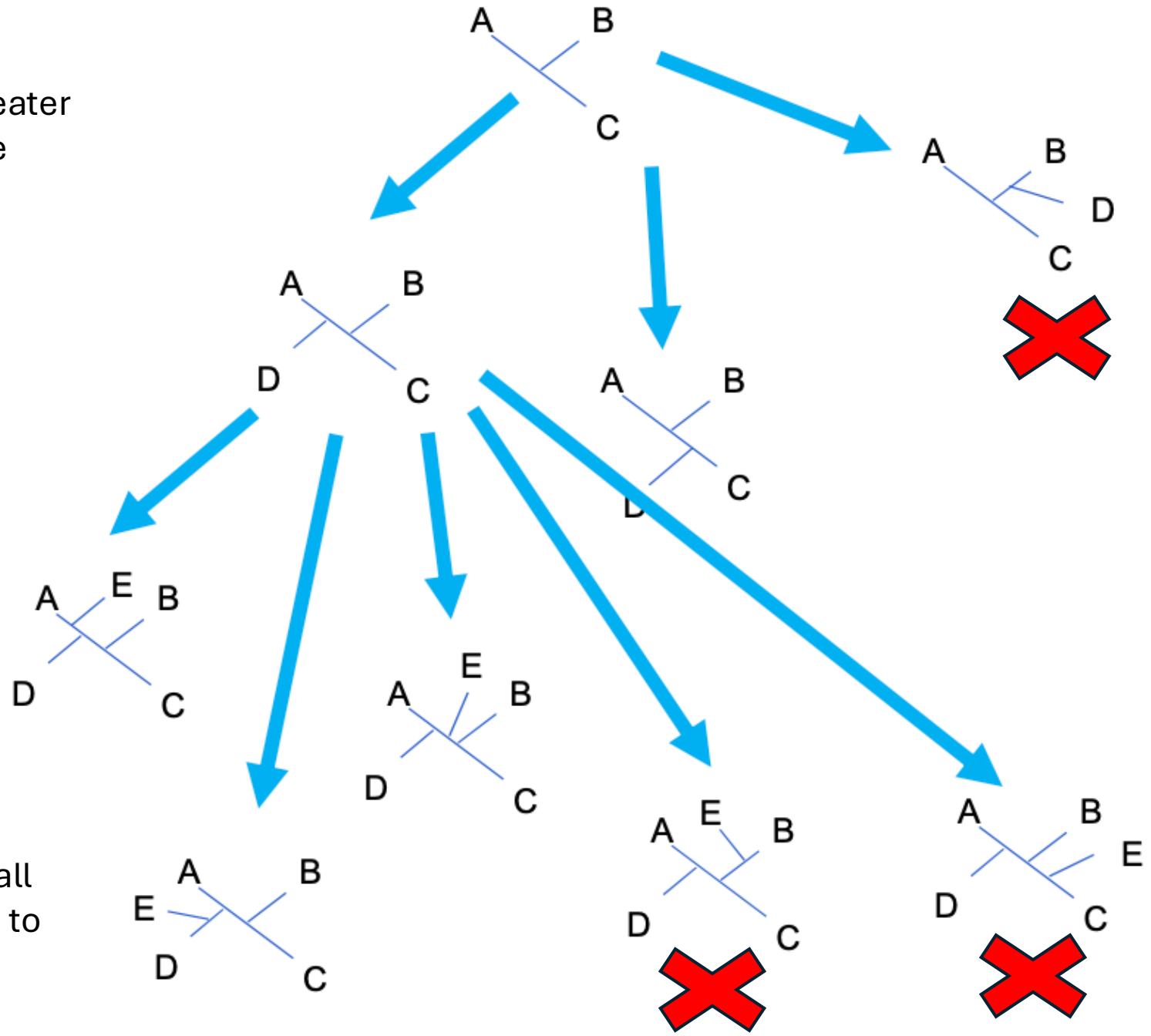

Add taxon D in all possible location
A
B
C
A
B
D
C
A
B
D
C
A
B
D
C


Add taxon D in all possible location
A
B
C
A
B
D
C
A
B
D
C
A
B
D
C
If tree total branch length is > L
Stop searching

Otherwise, keep adding another taxa (E) in all pairwise combinations. The search “branches out”.



If any tree is greater than L, stop the process.



Keep going until all nodes are added to the trees.

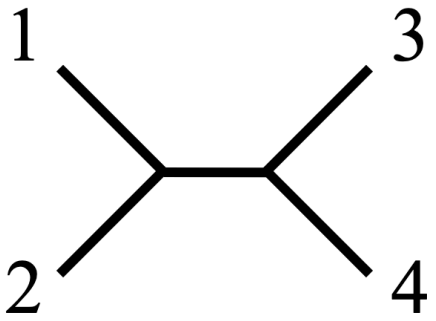
Branch and Bound

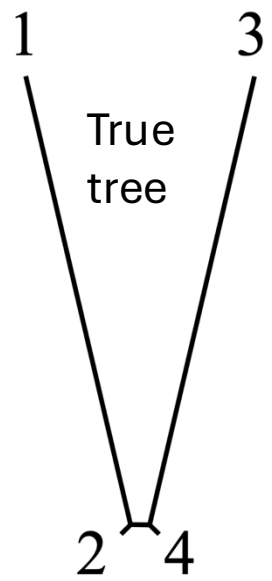
- B&B will find the best tree.
- However, it is still computationally expensive.
- In the worst case, it will take as long as a brute force search.
- In the best case, many bounds make the search very quick.

Construction Phylogenetic Trees

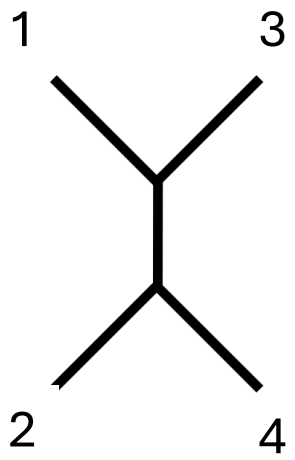
Long Branch attraction

Accurate topology





LBA
topology



- Phylogenetic trees are reconstructed based on sequence similarities between organisms.
- But if a set of similar organisms contains few organisms that are very different from the main group, these outgroup organisms cluster together even though they are very distinct to each other.
- They are clustered together due to shared dissimilarity to the main group of organisms.
- The more data -> closer to getting the correct answer.

Felsenstein 1979

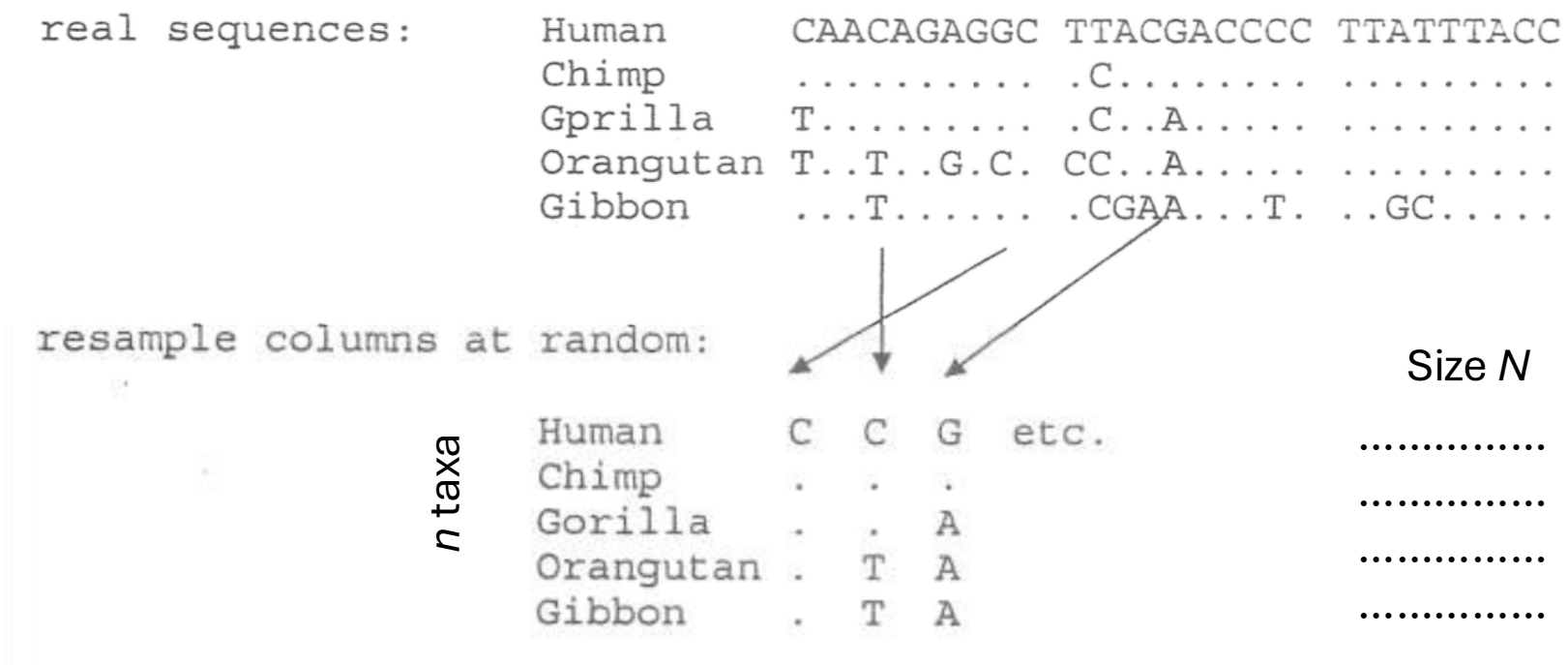
Assessing the trees: Bootstrapping

- The bootstrap is a method of assessing the significance of some phylogenetic feature, such as the segregation of a particular set of species (taxa, or leaves) on their own branch (clade).

- It is a confident measure.

- The bootstrap works as follows:
  - Given an alignment, an artificial dataset of the same size (taxa *x* seq-length) is generated by picking columns from the alignment **at random with replacement**.
  - A given column in the original dataset can appear several times in the artificial dataset.
  - The tree building algorithm is then applied to the new (created) dataset.
  - The procedure is repeated some number of times (1000).
  - The frequency with which a chosen phylogenetic feature (a particular branch) appears is taken as a measure of confidence.

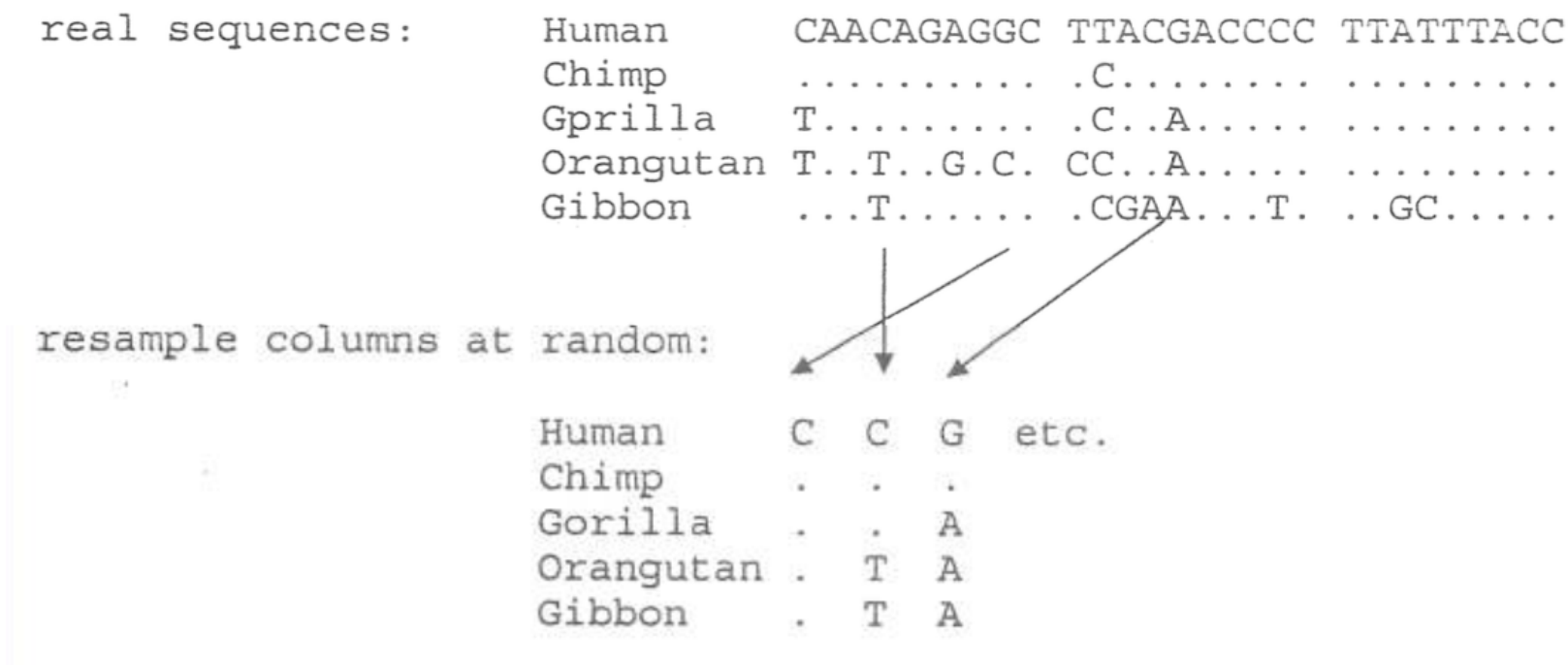
Assessing the trees: Bootstrapping



Synthetic datasets
n x N

Source:
Bioinformatics and Molecular Evolution, Higgs and Attwood

Assessing the trees: Bootstrapping



- Repeat to create many "synthetic" data sets.
- Generate trees from each.
- See the frequency of support in the original tree's branches.
- 100, 500, 1000 synthetic data sets…

Assessing the trees: Bootstrapping

- If only a small fraction of the resamples data sets support the relationships, they are probably not accurate.
- If only a large fraction of the resamples data sets support the relationships, they are more likely to be accurate.

- But it is not true inference. Not statistical probability (no p-value).
- You could get false positives if the wrong evolutionary model is being used.
- We can use the confidence to generate a consensus trees.

Assessing the trees: Consensus tree

- It is not uncommon for a given algorithm to produce more than one “best tree”.

- This happens when trees (with different topologies) have identical branch lengths (*L*). Example: we will find in Lab05 several “most” parsimonious trees with the same *L*.

- In such cases, a consensus tree is generated for better representation.
  - A consensus tree will collapse a branch if we have a poorly supported common ancestor into an *n*-furcating node (clade).

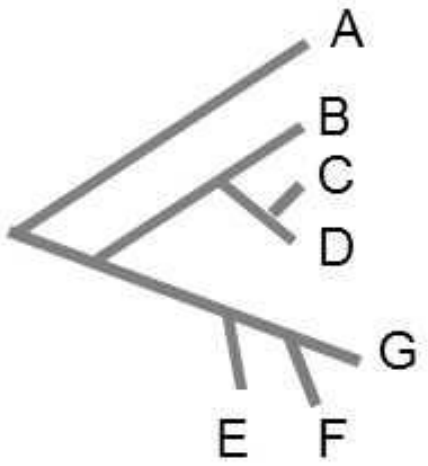
- We can use the resampling (bootstrap) procedure to assign a numerical confidence to internal branches.

- Methods:

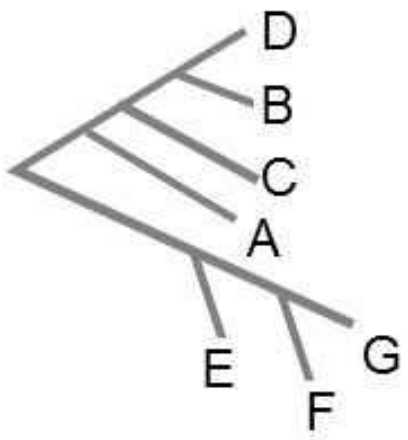
  - Strictly consensus tree
  - 50% majority rule consensus tree

Assessing the trees: Consensus tree

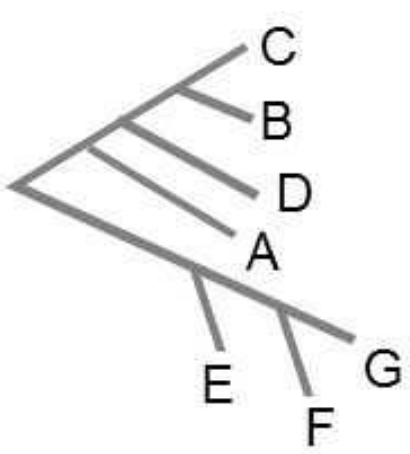
1



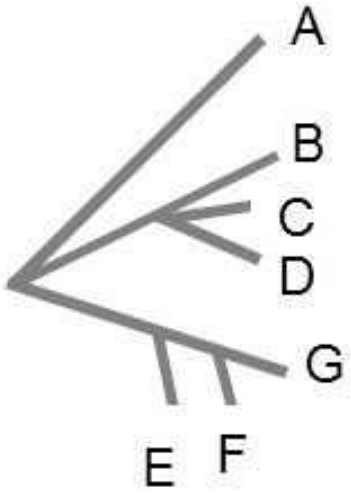
2



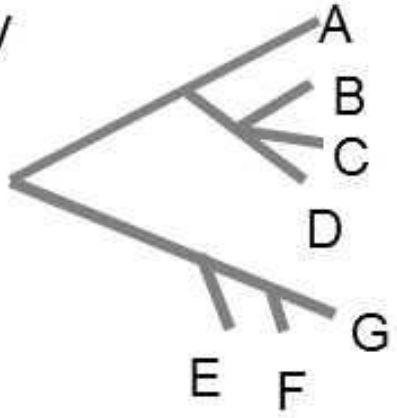
3



strictly consensual tree



50% majority consensual tree



Source:
Frozen Evolution, Flegr