Algorithms in bioinformatics (CSI 5126)[1]

Marcel Turcotte
(turcotte@site.uottawa.ca)

School of Information Technology and Engineering
University of Ottawa
Canada

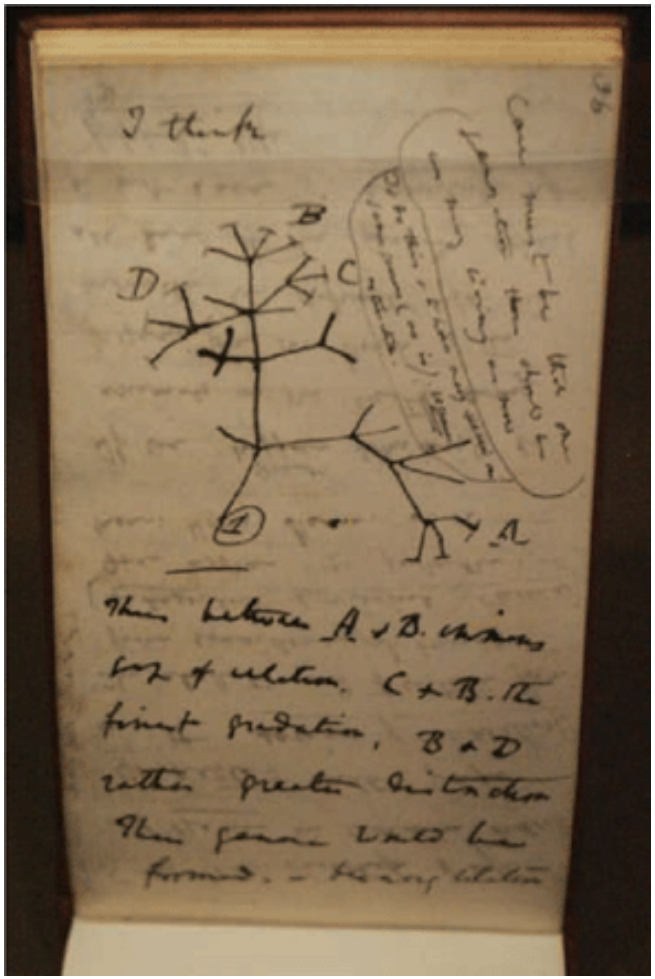
October 23, 2009

[1] Please don't print these lecture notes unless you really need to!

Plan

- Extension until Monday (Oct 26) 16:00 for Assignment 2
- Scientific paper review
- Projects
- Gotoh optimization for the affine gap function
- Inferring phylogenies
  - Motivation
  - Distance-based methods
  - **Character-based methods**
  - Maximum likelihood methods

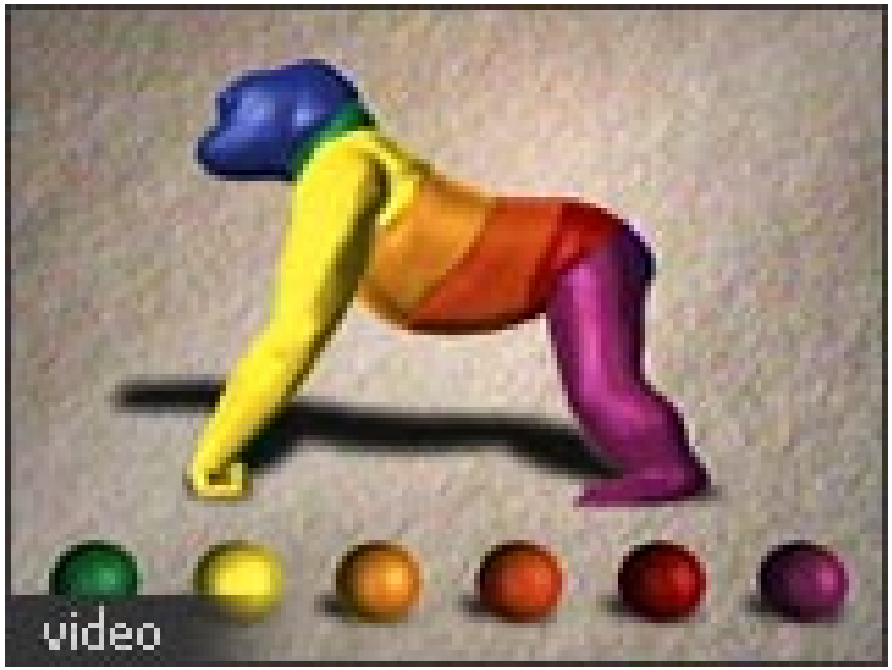
Evolution



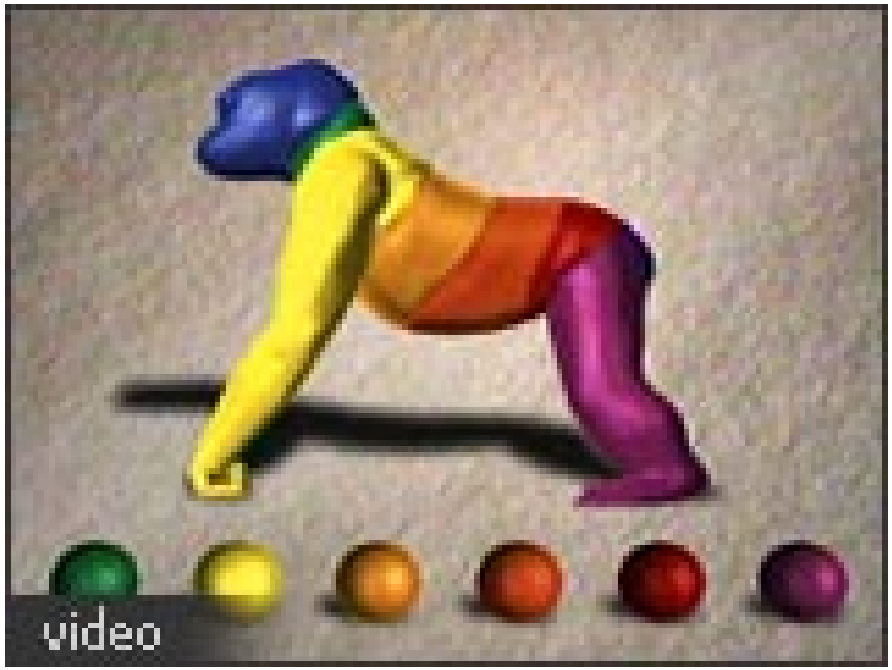
"Charles Darwin's famous notebook B containing the first known sketch of an **evolutionary tree**."

A. Rokas (2006) *Genomics and the Tree of Life. Science* **313**(5795): 1897–1899. (DOI: 10.1126/science.1134490)

Prelude: Evidences of evolution



Prelude: Evidences of evolution (cont.)



Genetic Tool Kit

(www.pbs.org/wgbh/evolution/library/03/4/l_034_04.html)

Prelude: Evidences of evolution

Prelude: Evidences of evolution (cont.)

The Common Genetic Code

(www.pbs.org/wgbh/evolution/library/04/4/l_044_02.html)

See also:

- **Evolving Ideas: How Do We Know Evolution Happens?** (www.pbs.org/wgbh/evolution/library/11/2/e_s_3.html)
- **Evolution library** (www.pbs.org/wgbh/evolution/library)
- **Understanding Evolution - Misconceptions about evolution and the mechanisms of evolution**

Introduction

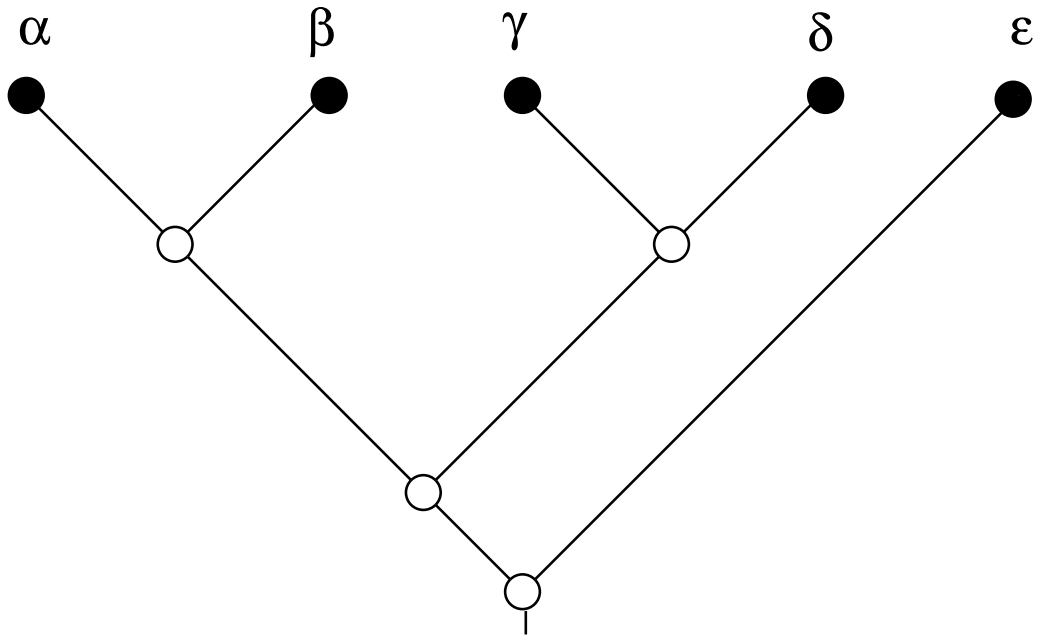
- "The objectives of phylogenetic studies are (1) to **reconstruct** the correct genealogical ties between organisms and (2) to **estimate the time of divergence** between organisms since they last shared a common ancestor."
- "A phylogenetic tree is a **graph** composed of nodes and branches, in which only one branch connects any two adjacent nodes."
- "The nodes represents the **taxonomic units**, and the branches define the **relationships** among the units in terms of **descent and ancestry**."
- "The **branch length** usually represents the number of changes that have occurred in that branch." (or some amount of time)

$\Rightarrow$ Li, W.-H. and Graur, D. (1991) Fundamentals of Molecular Evolution. Sinauer.

Terminology

- A **taxon** (plural **taxa**) or **taxonomic unit** is a species or grouping of species.
- Naming the different taxonomic levels: kingdom; phylum; class; order; family; genus; species.

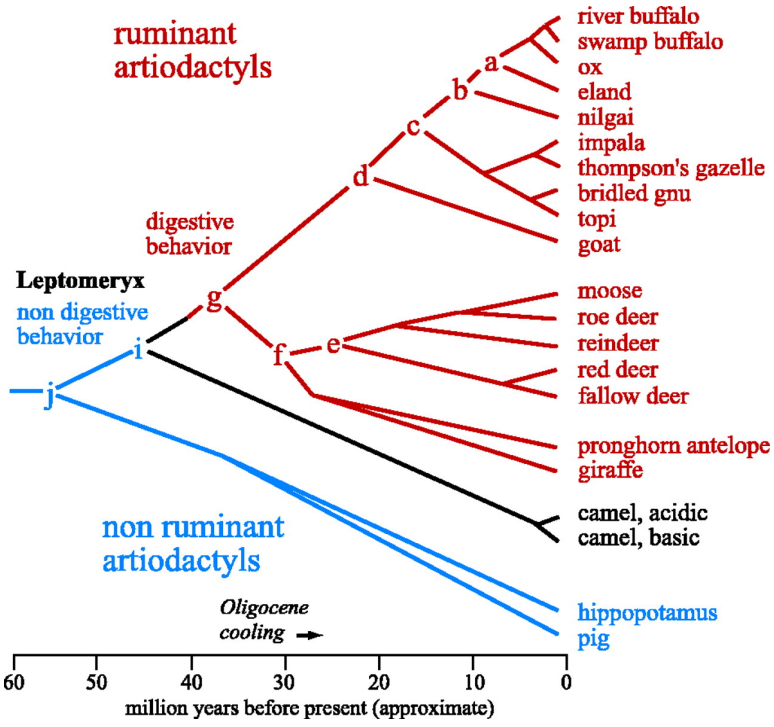
Terminology (cont.)



A **rooted tree** for 5 species. Leaves, $\alpha, \beta, \gamma, \delta$ and ϵ, correspond to **contemporary organisms**, for which data has been collected ($t = 0$). Internal nodes correspond to (inferred) **ancestors** ($t < 0$).
Newick format of that tree: $(((\alpha, \beta), (\gamma, \delta)), \epsilon)$

Why?

- **Comparative studies**: understanding gene function, adaptation, correlating the appearance of a trait to environmental factors, etc.;
- **Drug design**: designing compounds that are specific to a group of organisms;
- **Bioinformatics**: multiple sequence alignment, protein (secondary) structure prediction, etc.;



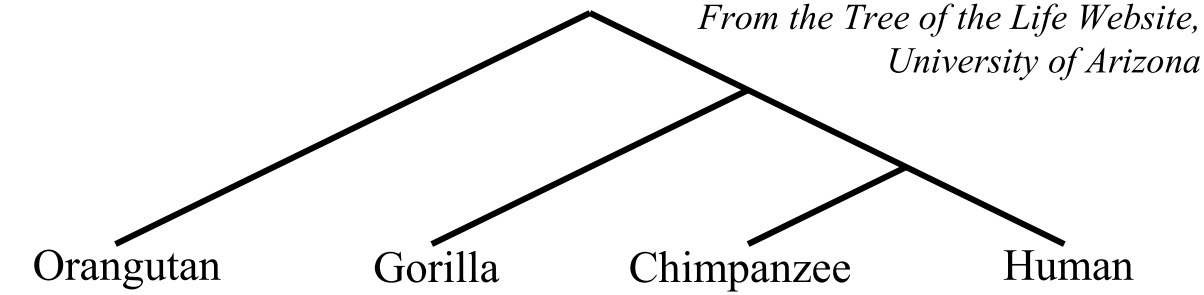
S.A. Benner (2002) *Science* **296**(5569): 864–868.

"One way of testing such hypotheses is to **resurrect the ancestral proteins** and study their behavior in the laboratory. To do this, a DNA molecule encoding **the ancestral protein is synthesized** and expressed in an appropriate host. The ancient protein is then recovered and studied to determine whether its properties are consistent with its inferred ancestral role. (...) **digestive ribonuclease emerged near the time when ruminant digestion emerged, in animals in which ruminant digestion developed, at a time where difficult-to-digest grasses emerged, permitting their descendants to exploit a newly available resource emerging at a time of global climatic upheaval.**"

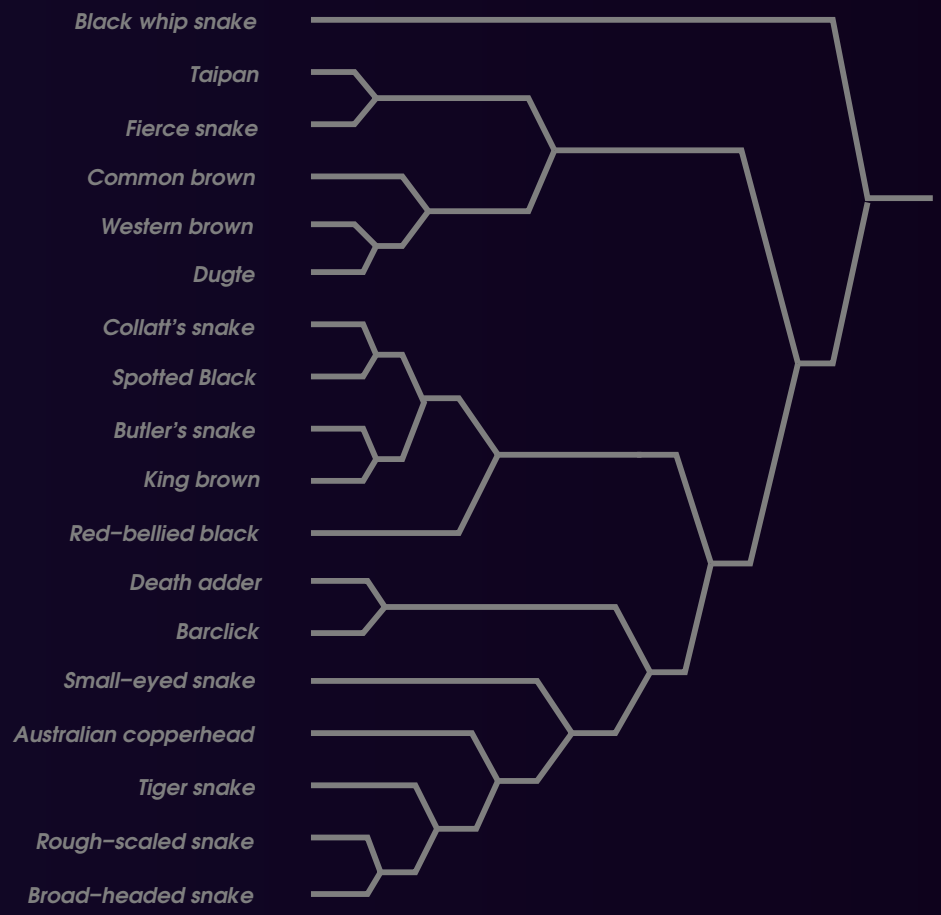
The Great Apes

Phylogeny

From the Tree of the Life Website, University of Arizona

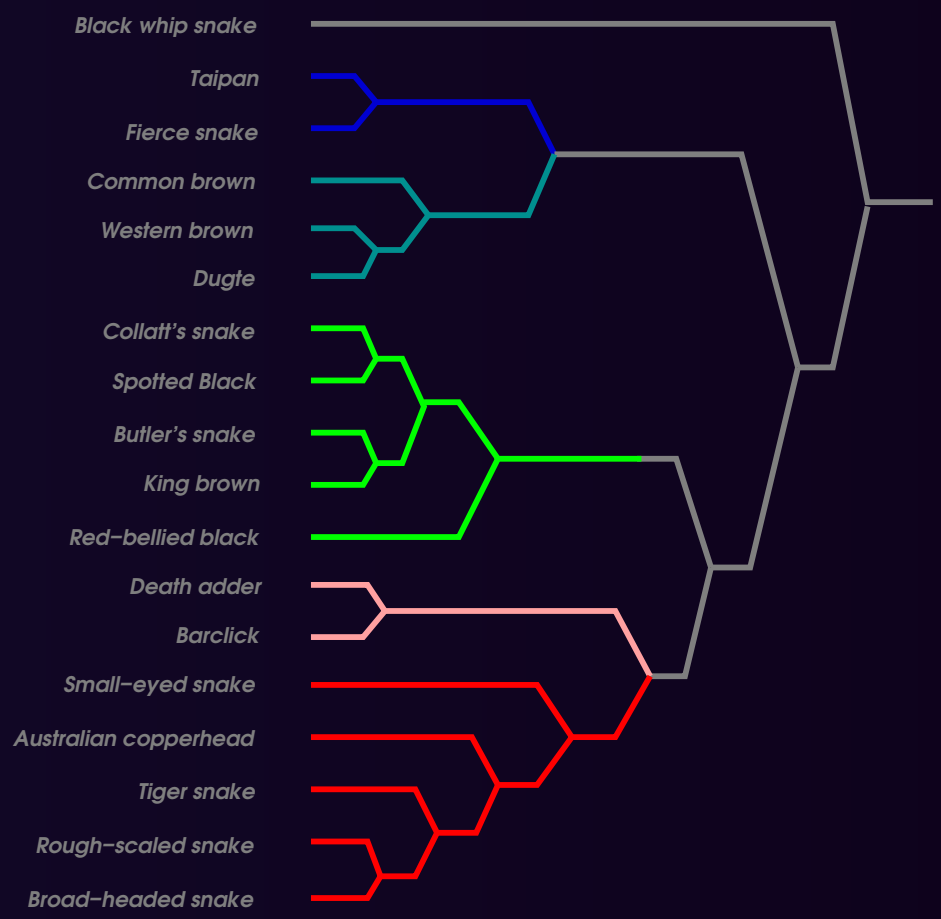


Example: Antivenins



Venomous Australian Snakes

Example: Antivenins



Scale of The Tree of Life

- 1,5 million described species.
- 10 million to 200 million existing species.

- Reconstruction tools can handle around 500 organisms.
- Reconstruction tools scale exponentially with the amount of data.

Credits for the last four slides

Bernard Moret

1980–2006: University of New Mexico
(www.cs.unm.edu/∼moret)

2006–: École Polytechnique Fédérale de Lausanne
(people.epfl.ch/bernard.moret)

See also: Tree of Life Web Project (www.tolweb.org).

Summary of the applications

- Study ancestor-descendant relationships
  (Evolutionary biology, adaption, genetic drift, selection, speciation, etc.)
- Paleogenomics: inferring ancestral genomic information from extinct species
  (Comparing Chimpanzee, Neanderthal and Human DNA)
- Origins of epidemics
  (Comparing, at the molecular level, various virus strains)
- Drug design: specifically targeting groups of organisms
  (Efficient enumeration of phylogenetically informative substrings)
- Forensic
  (Relationships among HIV strains)
- Linguistics
  (Languages tree divergence times)

Caveats

$\Rightarrow$ Felsenstein, J. (2004) Inferring phylogenies. Sinauer.

- "Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyze those differences statistically."
- "I estimated that there are about **3,000 papers** on methods for inferring phylogenies."
- "The field of inferring phylogenies has been wracked by **outrageously excessive controversy** (...) there have been many biologists who strove to bring back the field to normality (...)."

Joe Felsenstein is the author of a software package called PHYLIP, which is one the most widely used software system for phylogenetic studies.

What's the data? (1/2)

	Characters					
Species	**1**	**2**	**3**	**4**	**5**	**6**
α	1	0	0	1	1	0
β	0	0	1	0	0	0
γ	1	1	0	0	0	0
δ	1	1	0	1	1	1
ϵ	0	0	1	1	1	0

Here, the 0s and 1s are indicating the presence or absence of a character (has feathers?, lays eggs?, curved beak?, flies?, ...).
A character is a measurable feature having well-defined mutually exclusive states.

What's the data? (1/2) (cont.)

- Based on anatomical and behavioural characters, the panda was classified as a raccoon (1870). Recently, 1985, the panda was re-classified as a bear when an analysis based on molecular data was done.

$\Rightarrow$ Images from www.wikipedia.org

What's the data? (1/2)

Platypus (*Ornithorhynchus anatinus*) "the only **mammals that lay eggs** instead of giving birth to live young", "[i]t is the sole representative of its family (Ornithorhynchidae) and genus (Ornithorhynchus)", "[t]he platypus is considered to be one of the strangest specimens of the animal kingdom: a **venomous**, **egg-laying**, **duck-billed mammal**".

(commons.wikimedia.org/wiki/Image:Ornithorhynchidae-00.jpg)

Genome analysis of the platypus reveals unique signatures of evolution. Nature (2008) vol. 453 (7192) pp. 175-183

What's the data? (1/2)

- The **thylacine** (*Thylacinus cynocephalus*) is a now extinct (wolf-like) carnivorous marsupial.

(commons.wikimedia.org/wiki/Image:Thylacinus.jpg)

Hard to resolve relationships using morphology and behaviour alone

1. Similar characteristics can evolve independently in distantly related organisms — **convergent-evolution**;
2. It is often difficult to find characteristics that are common to all the organisms under study.

What's the data? (2/2)

	Characters					
Species	**1**	**2**	**3**	**4**	**5**	**6**
α	A	G	A	C	G	G
β	C	G	T	G	A	G
γ	A	C	A	G	A	G
δ	A	C	A	C	G	A
ϵ	C	G	T	C	G	G

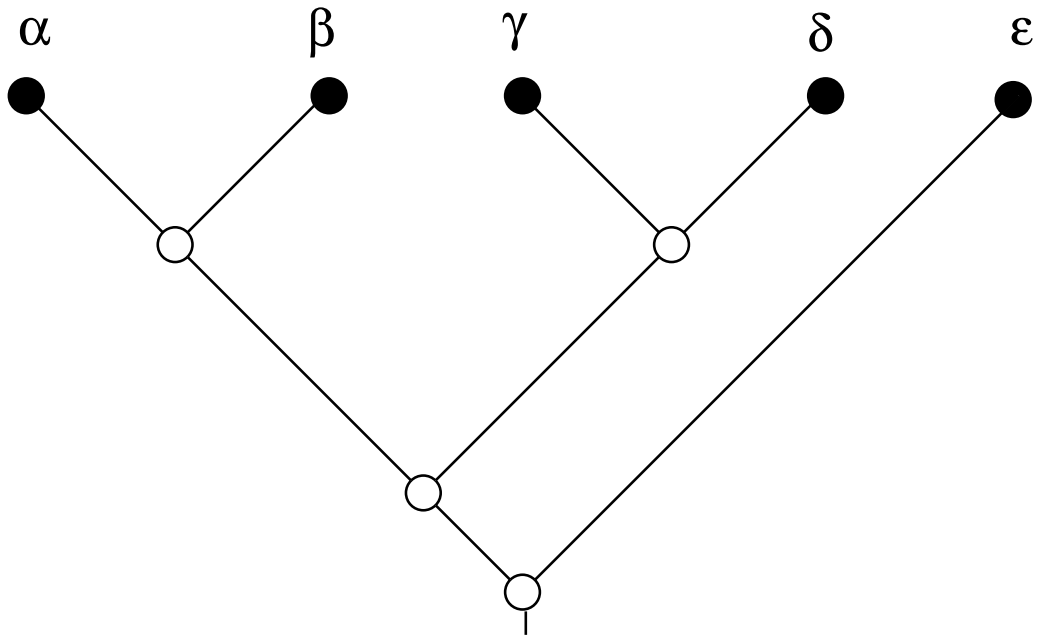
Nowadays, biologists rely on molecular sequence data, in particular DNA or RNA sequences, which allows the comparison of a broader range of species. **What characters, other than molecular sequence data, would allow to compare E. coli, yeast, clam shell and human?**

Genes or species trees

- Herein, a molecular sequence alignment (DNA, RNA or proteins) is used as input. Each column (site) of this alignment represents a character.
- The taxonomic units (nodes of the tree) can represent genes, species or populations; not all at once obviously.
- A gene tree represents the evolutionary history of a single gene; e.g. the evolution of the globin family (with its numerous gene duplication events).
- A species tree will generally be built using multiple genes (say 100).
- Since our main interest is to study the methods, we will limit our discussion to species trees.

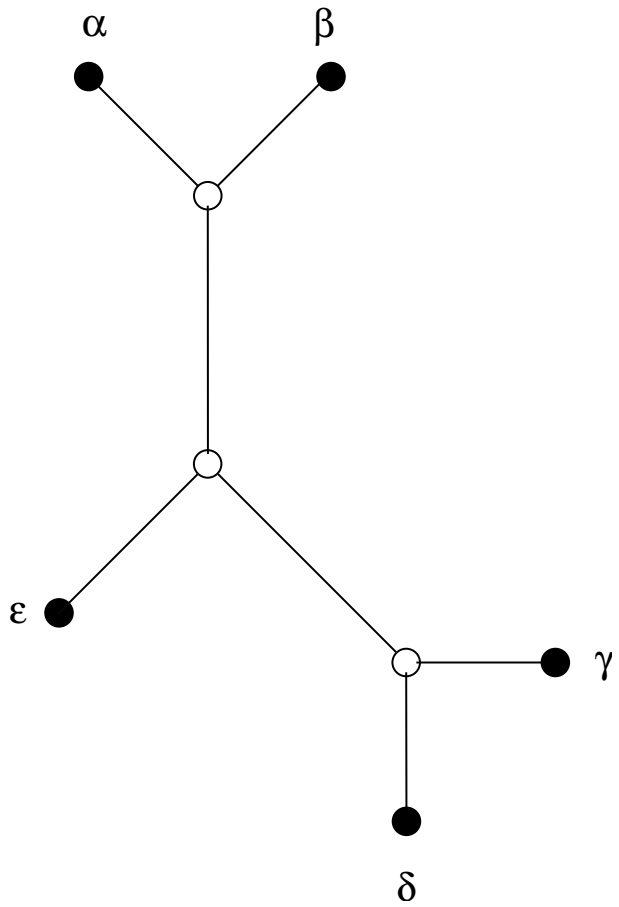
Rooted tree

A **rooted tree** not only gives the relationships between the taxonomic units, it also indicates the direction of evolution (time). Such trees can be scaled or unscaled.



Unrooted tree

An **unrooted tree** specifies the relationships between species.



Rooting the tree

Biologists generally prefer rooted trees.

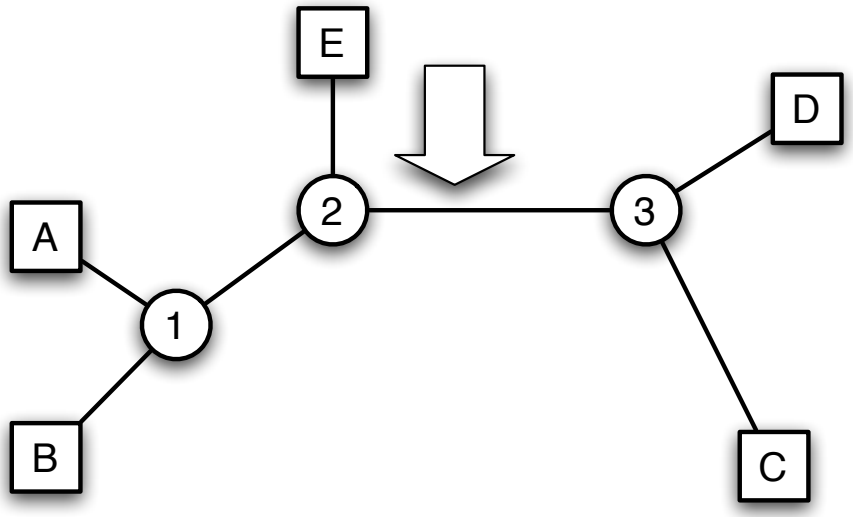
- Under the **molecular clock assumption**, the root of the tree would be located at equal distance from all the leaves (contemporary organisms);
- The **outgroup method** consists of including into the analysis an organism that is known to have branched off earlier than the taxa under study (for which paleontological evidences exist, for instance), the root will be placed along the edge connecting the outgroup to the ancestor of the ingroup (taxa under study).

Molecular clock theory

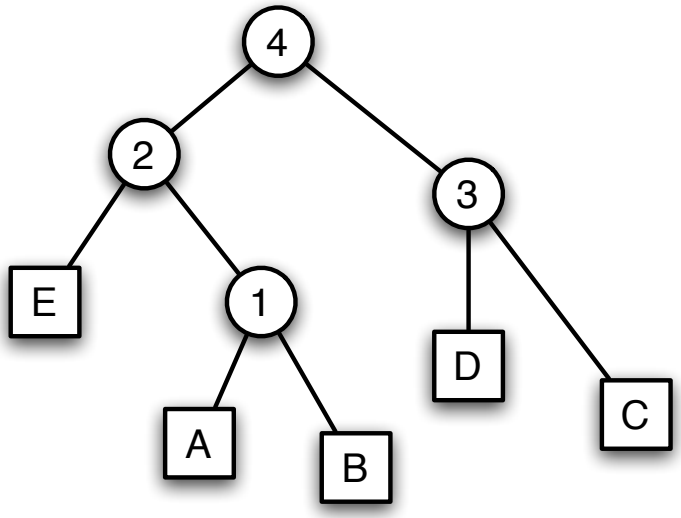
- Proposed by Emile Zuckerkandl and Linus Pauling, 1962.
- **Accepted mutations occur at a constant rate.**
- The number of accepted mutations is proportional to the length of the time interval.
- Once the "**clock**" has been calibrated (using fossil evidences, for instance) the unknown length of some time interval can be deduced from the number of accepted mutations.
- **Note**: different proteins have different clocks (hemoglobin ticks faster than cytochrome c).
- "A great deal of ink (and blood) has been spilt over the molecular clock (...)"

$\Rightarrow$ [3, pages 453–455]

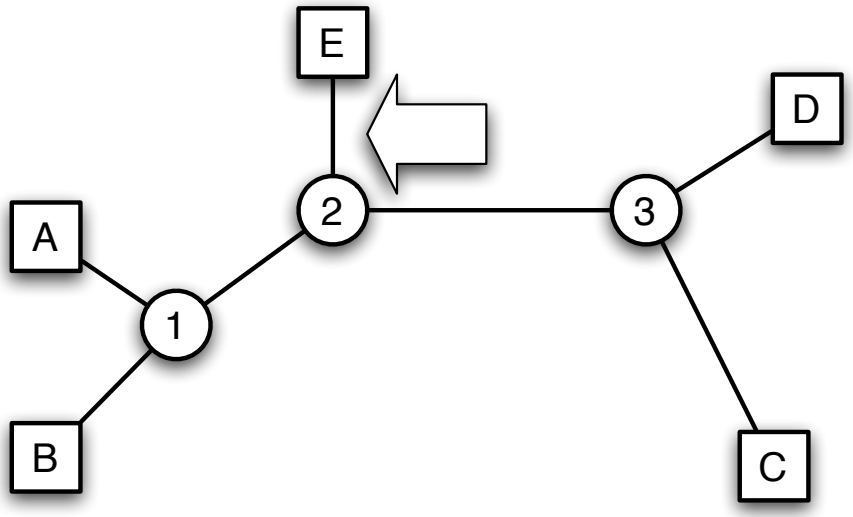
Rooting a tree: molecular clock



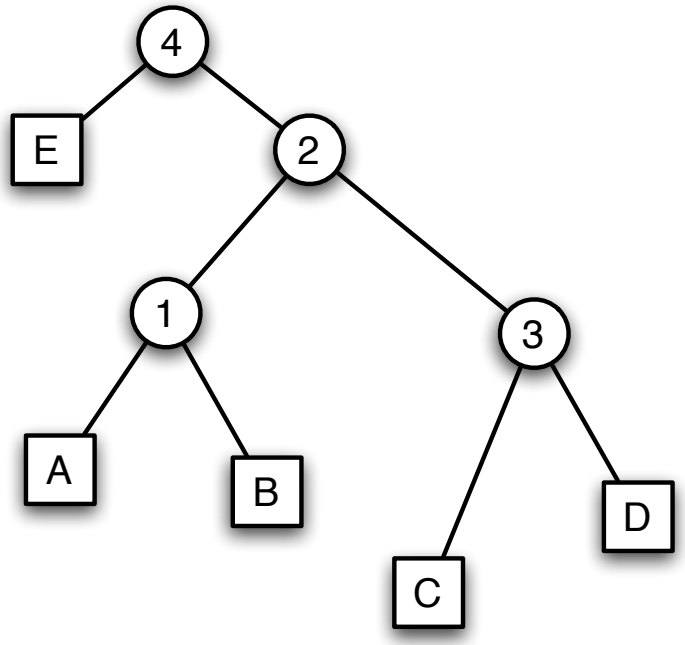
Rooting a tree: molecular clock (cont.)



Rooting a tree: outgroup

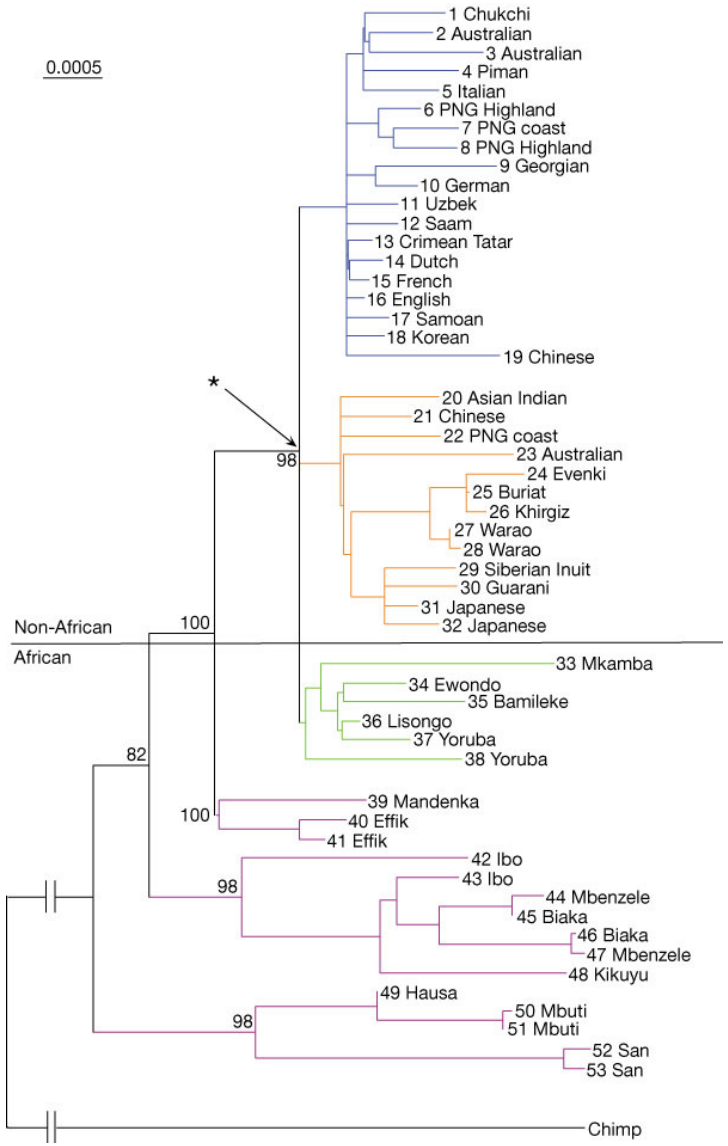


Rooting a tree: outgroup (cont.)



Hypothetically, E is the outgroup — e.g. chimpanzee while the ingroup consists of human populations.

Rooting a tree: outgroup (cont.)



Neighbour-joining phylogram based on complete mtDNA genome sequences.
Source: Max Ingman, Henrik Kaessmann, Svante Pääbo and Ulf Gyllensten (2000) Nature 408, 708-713

Number of trees

# Species	# rooted trees	# unrooted trees
5	105	15
10	34,459,425	2,027,025
15	213,458,046,676,875	7,905,853,580,625
20	8,200,794,532,637,891,559,375	221,643,095,476,699,771,875

It can be shown that the number of rooted and unrooted trees for a given n (number of species) are as follows.

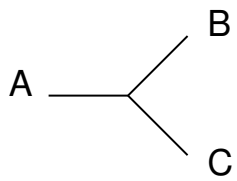
$$N_{rooted}(n) = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

$$N_{unrooted}(n) = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

General paradigm

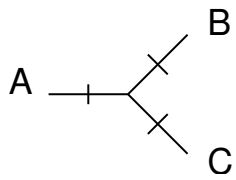
1. Enumerate trees;
2. Select the "best" tree.

Sequential addition strategy



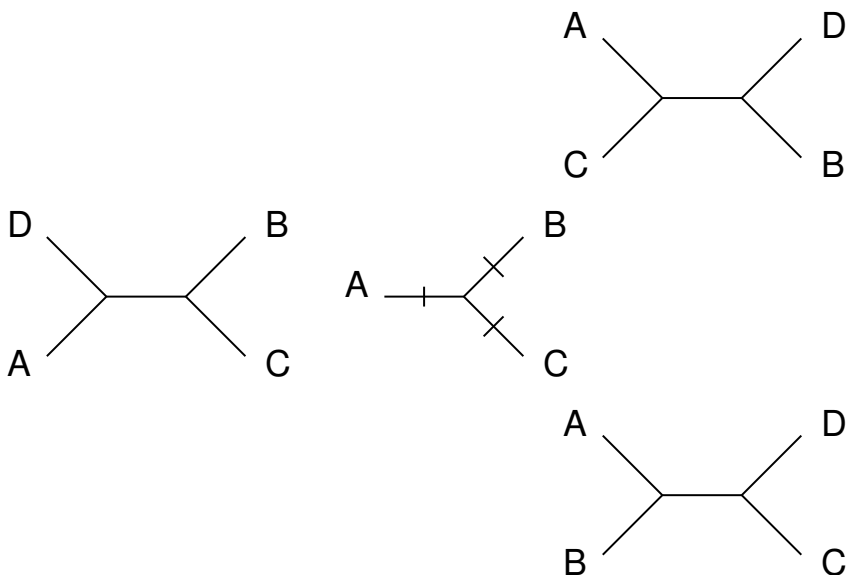
Given three species, there is a single unrooted tree.

Sequential addition strategy



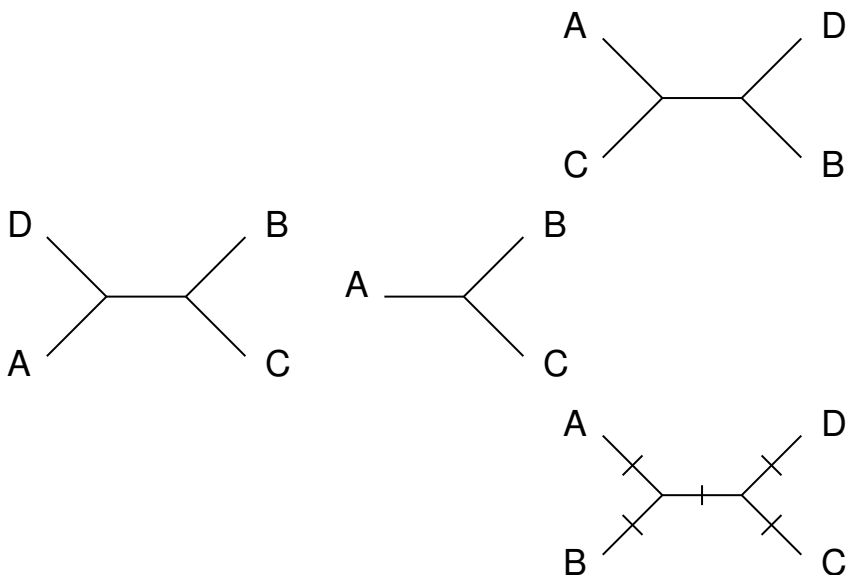
Each branch can serve as an insertion point, adding a new branch off the middle of any existing branch.

Sequential addition strategy



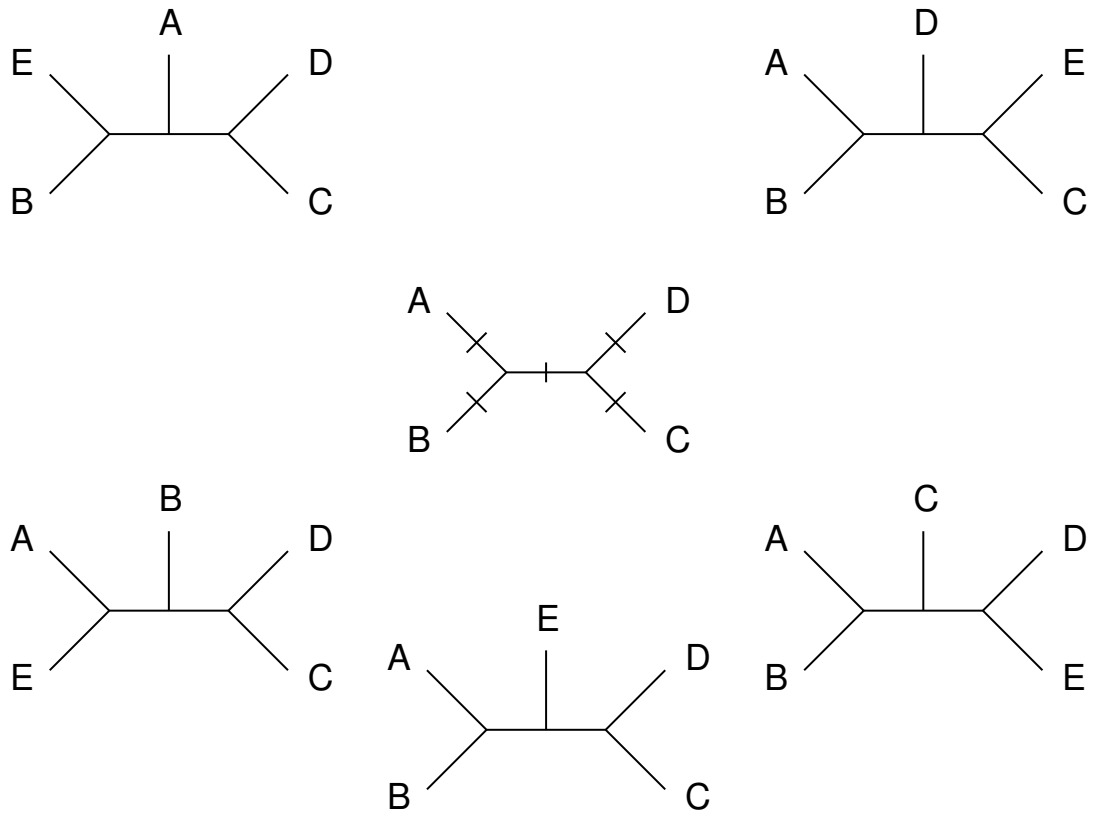
Therefore producing 3 four species unrooted trees.

Sequential addition strategy



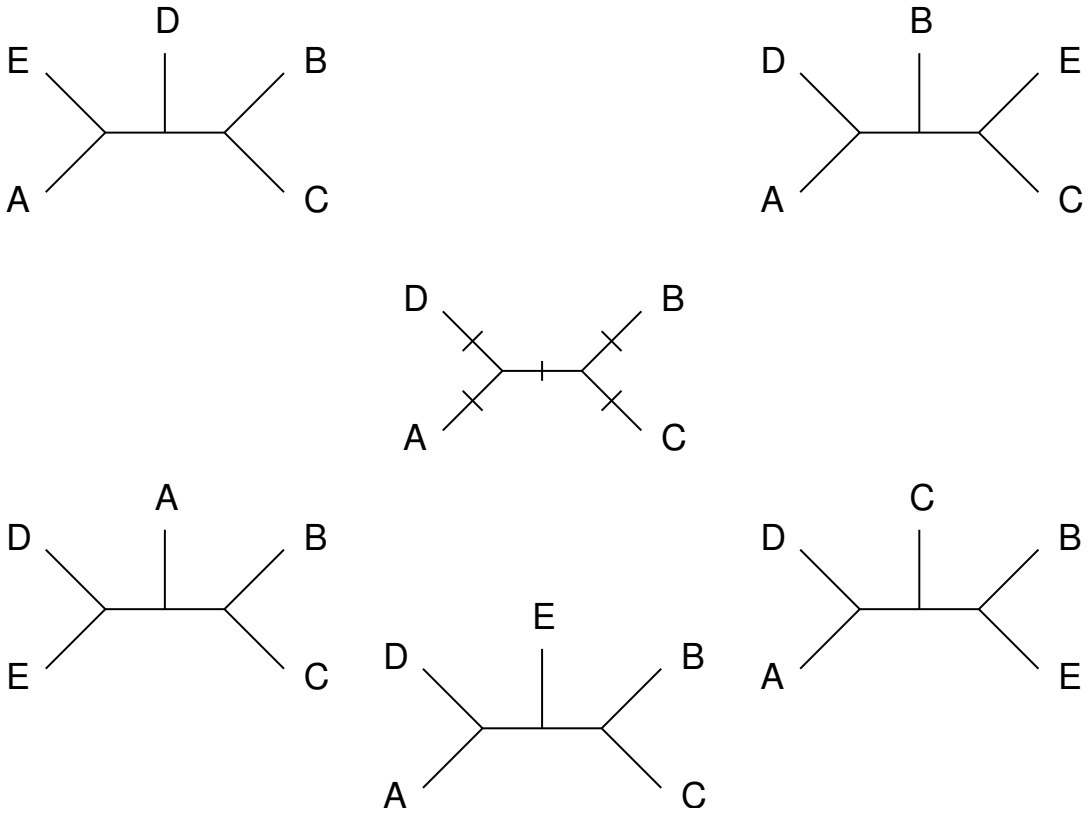
The same process is applied to all 3 four species trees.

Sequential addition strategy



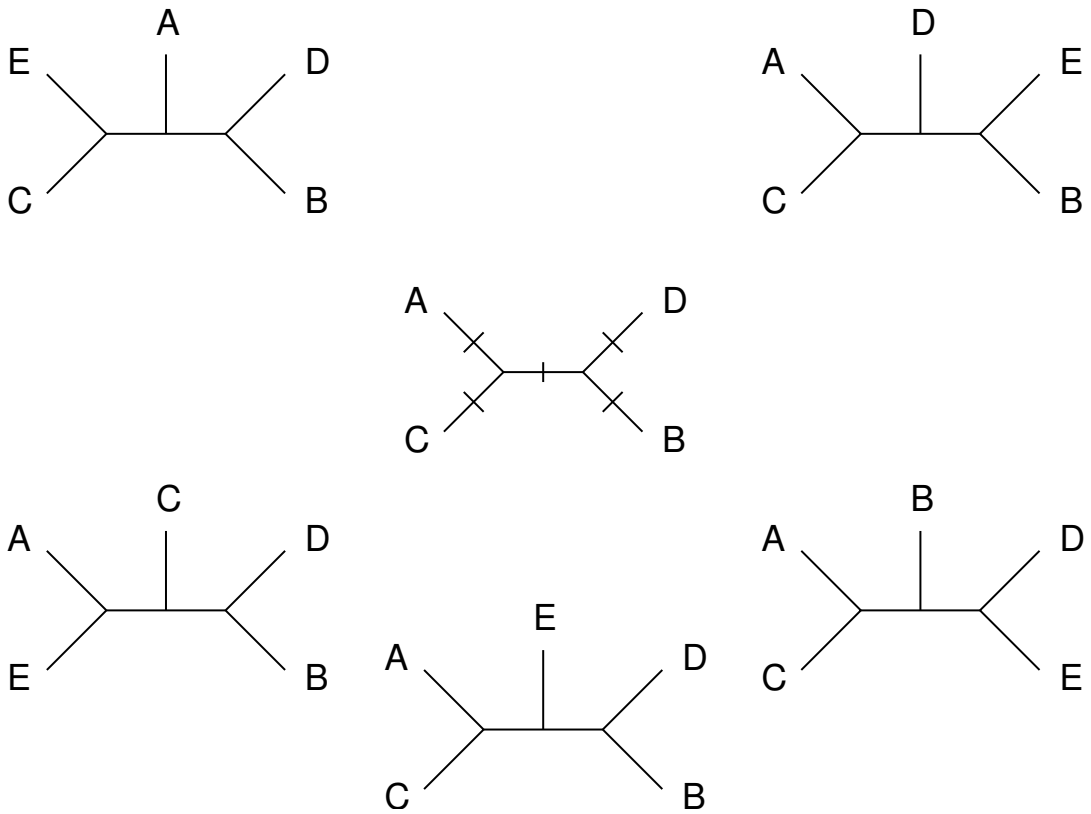
A four species unrooted tree has 5 edges, thus leading to 5 new unrooted trees.

Sequential addition strategy



There will be 15 five species unrooted trees.

Sequential addition strategy



Methods

- **Distance-based**
  a distance is a measure of the overall differences/similarities between two objects
- **Character-based**
  a character is a characteristic that has well-defined, limited number of states
- **Maximum likelihood**
  Finds a tree such that the likelihood of the data given the tree structure is maximum

"While disputes between the champions of the two approaches [character-based and distance-based] have often been surprisingly intense, it is fair to say that both approaches are widely used and work well with most data sets." [5, page 85]

I. Distance-based reconstruction

- Let D_{ij} be the pairwise distance between two species; for instance, measured by comparing, in some ways, sequence data from the two species.
- Let d_{ij} be the distance between i and j in some tree; the sum of the length of all the branches on the unique path between i and j.
- Distance-based methods seek to find a tree (topology + branch length) such the D_{ij} and the d_{ij}, for all i and j, are in **good agreement**.
- For instance, find a tree minimizing

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(D_{ij} - d_{ij})^2 \tag{1}$$

Minimum least squares approach.

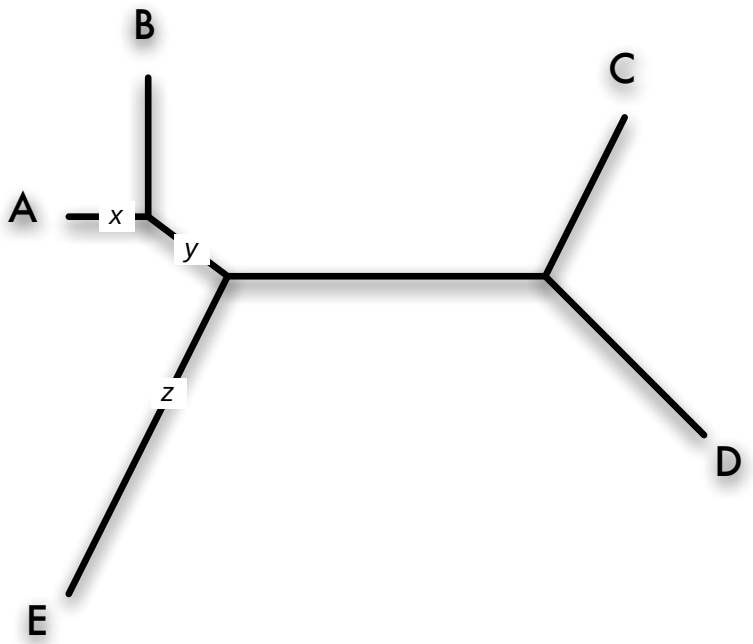
If the tree topology was given, the problem would simply be to find the length of the branches minimizing Equation 1, which is a quadratic function.
The d_{ij} need to be expressed as the sum of the length of the branches, x, on the unique path between i and j, $\mathcal{P}_{ij}$.

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij}(D_{ij} - \sum_{x \in \mathcal{P}_{ij}} x)^2 \tag{2}$$

and

$$d_{ij} = \sum_{x \in \mathcal{P}_{ij}} x$$

I. Distance-based reconstruction



$$d_{AE} = x + y + z$$

This involves solving the set of linear equations obtained by taking the derivative of Q with respect to the branch lengths and equating those to 0.

$$\frac{dQ}{dx} = -2\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(D_{ij} - \sum_{x \in \mathcal{P}_{ij}} x) = 0$$

Methods for finding exact solutions for the weighted and unweighted least squares branch length equations have been proposed that run in polynomial time (order 3 or less), and iterative/heuristic methods have also been proposed, which in practice converge rapidly towards the the correct lengths.

I. Distance-based reconstruction

- What's next?
- Well, we assumed that the topology was known!
- In principle, it would be possible to enumerate all tree topologies, solve the branch length equations for each one, and select the topology that minimizes the least squares.
- This is not practical since the number of topologies grows rapidly w.r.t. the number of species.
  [The same approaches as for parsimony methods for enumerating tree topologies can be applied. However, under certain assumptions, a simple algorithm, UPGMA, performs adequately.]

I. Distance-based reconstruction

Given n species and m characters.

1. Compute a distance matrix between all ($\Theta(n^2)$) pairs, this matrix is symmetrical;
2. Generate the topology of the tree;
3. Estimate the length of the branches.

Distance-based reconstruction

- For now, let's assume a simple distance measure, e.g. **Hamming distance** or **fractional alignment difference** ($\frac{D}{L}$) where D is the number of sites that differ (excluding indel containing sites), L is the number of sites
- More realistic models will be presented together with the Maximum Likelihood approach

Distance-based reconstruction

```
alpha   GTGCTGCACGG CTCAGTATA GCATTTACCC TTCCATCTTC AGATCCTGAA
beta    ACGCTGCACGG CTCAGTGCG GTGCTTACCC TCCCATCTTC AGATCCTGAA
gamma   GTGCTGCACGG CTCGGCGCA GCATTTACCC TCCCATCTTC AGATCCTATC
delta   GTATCACACGA CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCCTAAA
epsilon GTATCACATAG CTCAGCGCA GCATTTGCCC TCCCGTCTTC AGATCTAAAA
```

Build a multiple sequence alignment and **select columns**.

$\Rightarrow$ [5, page 87]

Distance-based reconstruction

Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0

The above matrix has been filled by computing the Hamming distance for all pairs of sequences.
$\Rightarrow$ [5, page 87]

UPGMA

UPGMA = Unweighted Pair Group Method using Arithmetic averages[2].

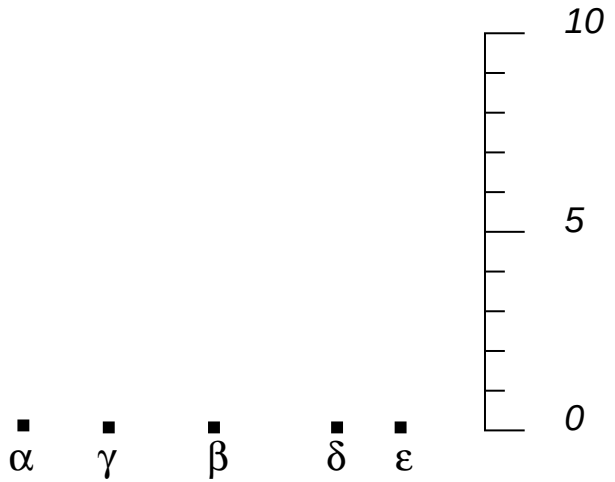
{ Initialization }
Assign each species i to its own cluster C_i
Define one leaf of T for each species, place it at height zero
{ Iterations }
Find the pair of clusters i and j which minimises d_{ij}.
Define a new cluster $C_k = C_i \bigcup C_j$.
Calculate d_{kl} for all l.
Create the parent node k of i and j
at height $d_{ij}/2$ in T.
Add k to the current list of clusters
and remove i and j.
{ Termination }
Stop when the list of clusters contains only one entry.

$\Rightarrow$ Early 1960s, simple and intuitive.

[2] See [1, page 166]

Distance-based reconstruction

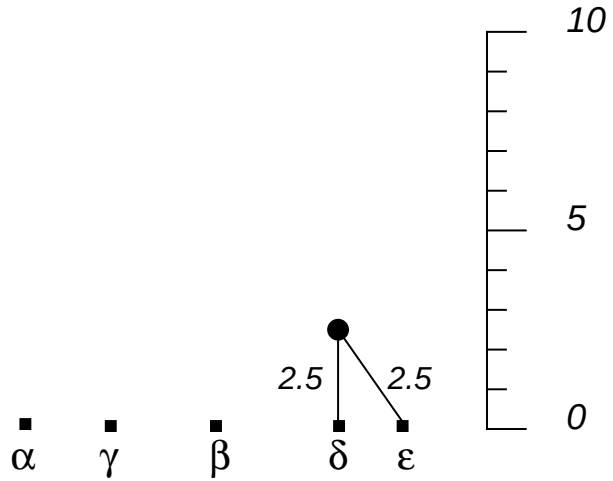
Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0



$\Rightarrow$ Assign each species to its own cluster, place it at height 0.

Distance-based reconstruction

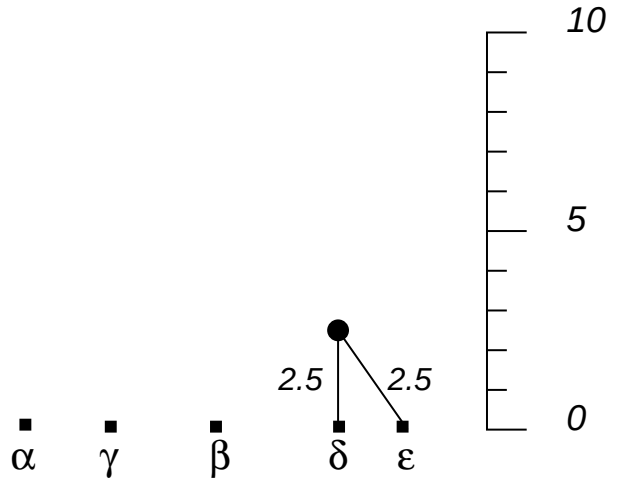
Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0



$\Rightarrow$ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

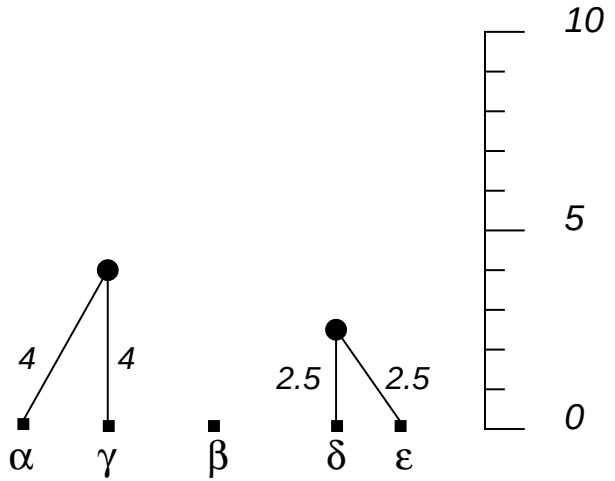
Species	α	β	γ	(δ, ϵ)
α	0	9	8	13.5
β	9	0	11	16.5
γ	8	11	0	11.5
(δ, ϵ)	13.5	16.5	11.5	0



$\Rightarrow$ Calculate $d_{k,l}$ for all l, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$.

Distance-based reconstruction

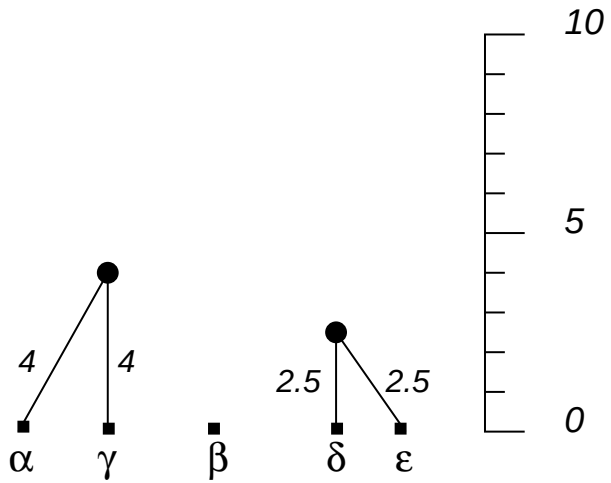
Species	α	β	γ	(δ, ϵ)
α	0	9	8	13.5
β	9	0	11	16.5
γ	8	11	0	11.5
(δ, ϵ)	13.5	16.5	11.5	0



$\Rightarrow$ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

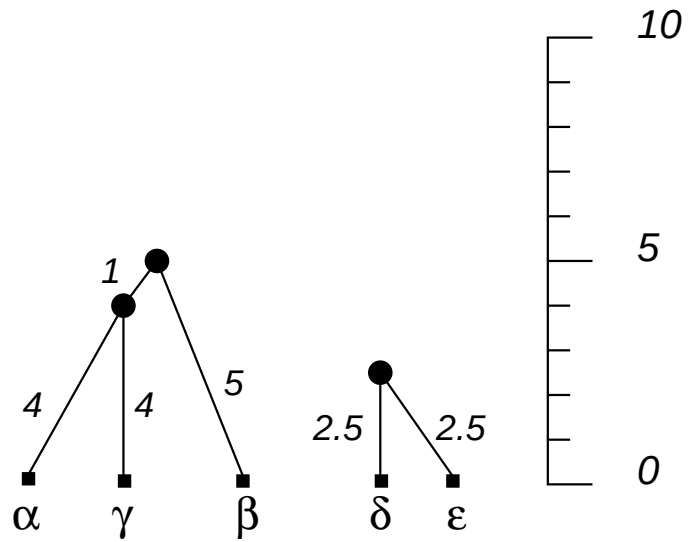
Species	(α, γ)	β	(δ, ϵ)
(α, γ)	0	10	12.5
β	10	0	16.5
(δ, ϵ)	12.5	16.5	0



$\Rightarrow$ Calculate $d_{k,l}$ for all l, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$.

Distance-based reconstruction

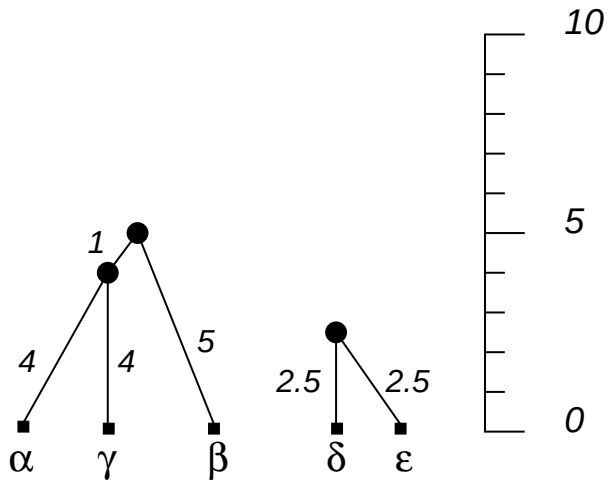
Species	(α, γ)	β	(δ, ϵ)
(α, γ)	0	10	12.5
β	10	0	16.5
(δ, ϵ)	12.5	16.5	0



$\Rightarrow$ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

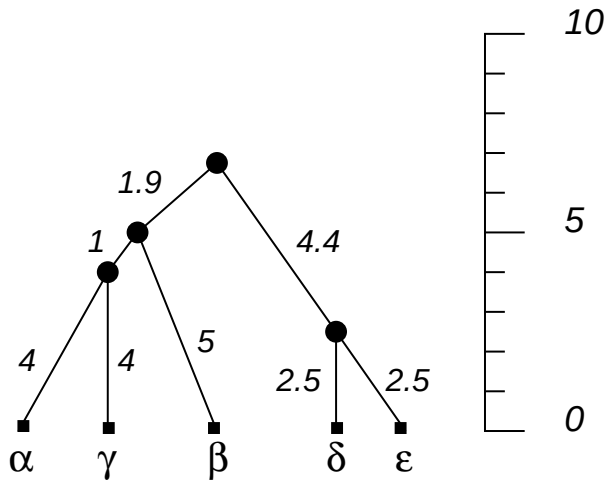
Species	$((\alpha,\gamma),\beta)$	(δ,ϵ)
$((\alpha,\gamma),\beta)$	0	$13.8\overline{3}$
(δ,ϵ)	$13.8\overline{3}$	0



$\Rightarrow$ Calculate $d_{k,l}$ for all l, where $d_{i,j} = \frac{1}{|C_i||C_j|} \sum_{p \in C_i, q \in C_j} d_{pq}$.

Distance-based reconstruction

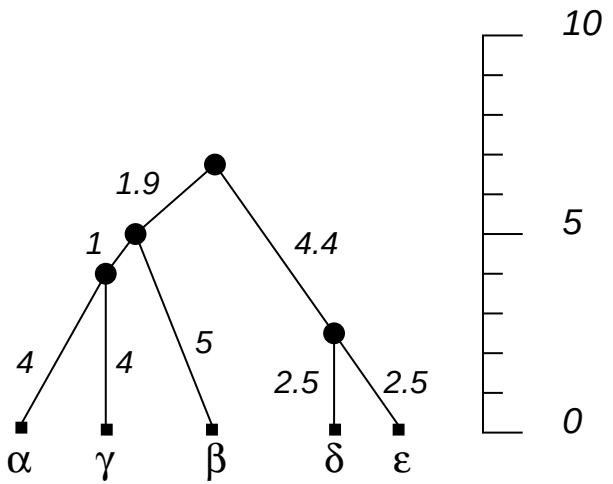
Species	$((\alpha,\gamma),\beta)$	(δ,ϵ)
$((\alpha,\gamma),\beta)$	0	$13.8\overline{3}$
(δ,ϵ)	$13.8\overline{3}$	0



$\Rightarrow$ Find the pair of clusters which minimises $d_{i,j}$, define a new cluster $C_k = C_i \bigcup C_j$, create the parent node k at height $d_{i,j}/2$.

Distance-based reconstruction

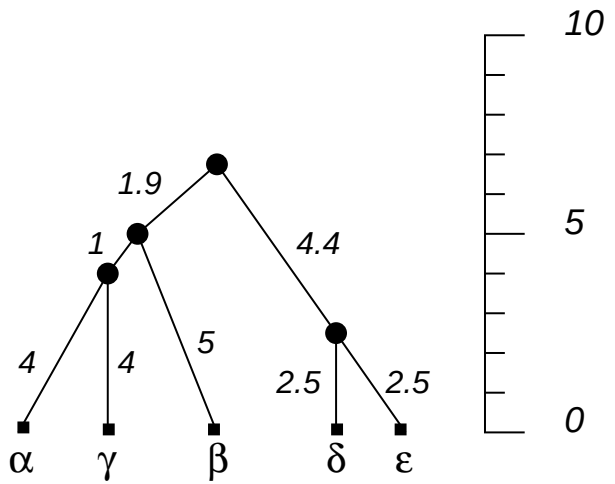
Species	$(((\alpha,\gamma),\beta),(\delta,\epsilon))$
$(((\alpha,\gamma),\beta),(\delta,\epsilon))$	0



$\Rightarrow$ UPGMA produces **ultrametric** trees, a tree such that the distance from any internal node (including the root) to its descendant leaves is the same. Thus, UPGMA assumes that evolution proceeds at the same rate in all the lineages, this is called the **molecular clock hypothesis**. (An assumption that is often violated)

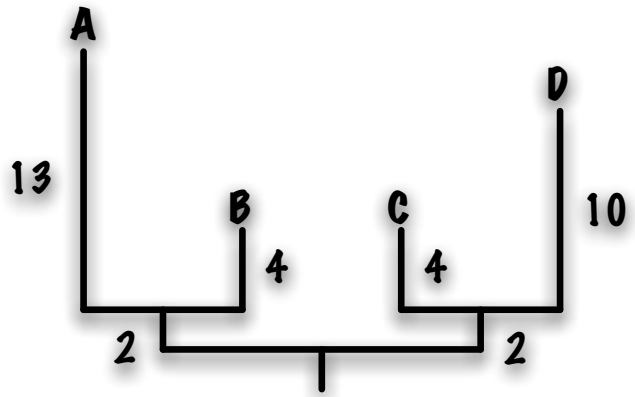
Distance-based reconstruction

Species	α	β	γ	δ	ϵ
α	0	9	8	12	15
β	9	0	11	15	18
γ	8	11	0	10	13
δ	12	15	10	0	5
ϵ	15	18	13	5	0



$\Rightarrow$ Consider $d_T(\alpha, \beta)$ and $d_{\alpha,\beta}$. Is $d_T(i,j) = d_{i,j}$ for all i, j?

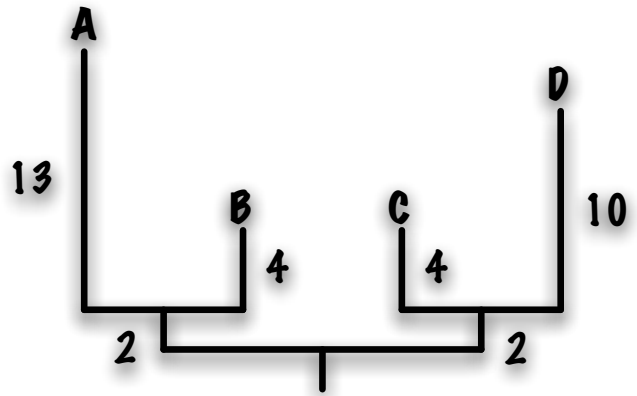
UPGMA — failure of the molecular clock hypothesis



"True tree"

Felsenstein 2004, page 167

UPGMA — failure of the molecular clock hypothesis



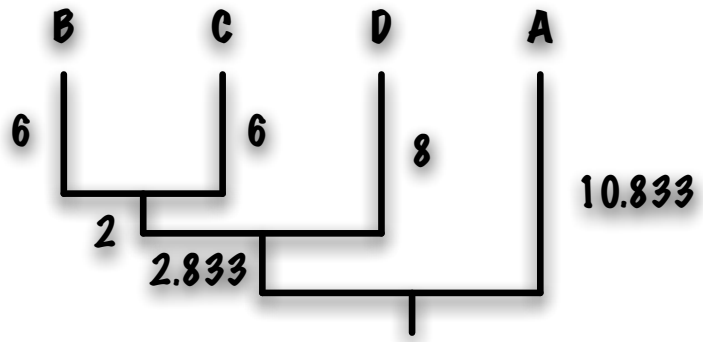
$$\Rightarrow \begin{array}{rrrr} 0 & 17 & 21 & 27 \\ 17 & 0 & 12 & 18 \\ 21 & 12 & 0 & 14 \\ 27 & 18 & 14 & 0 \end{array}$$

"True tree"

Felsenstein 2004, page 167

UPGMA — failure of the molecular clock hypothesis

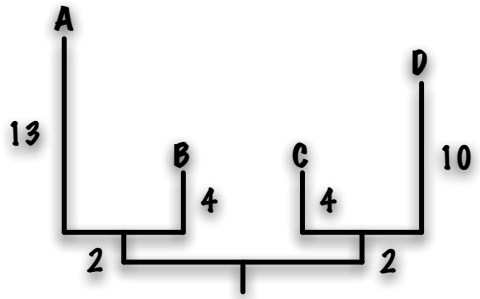
$$\begin{matrix} 0 & 17 & 21 & 27 \\ 17 & 0 & 12 & 18 \\ 21 & 12 & 0 & 14 \\ 27 & 18 & 14 & 0 \end{matrix} \Rightarrow$$



Reconstructed tree

Felsenstein 2004, page 167

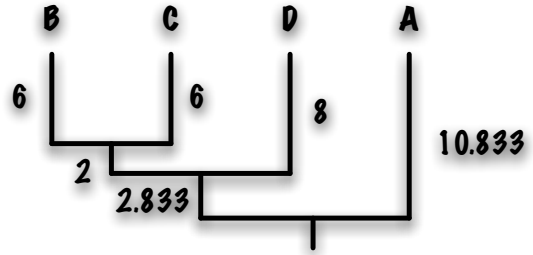
UPGMA — failure of the molecular clock hypothesis



$\Rightarrow$

$$\begin{matrix} 0 & 17 & 21 & 27 \\ 17 & 0 & 12 & 18 \\ 21 & 12 & 0 & 14 \\ 27 & 18 & 14 & 0 \end{matrix}$$

$\Rightarrow$



UPGMA joins B and C together since both are evolving slowly (short branch attraction).

Distance-based tree reconstruction

- Given a tree T, a distance matrix $d_{i,j}$ is *additive* if $d_T(i,j) = d_{i,j}$, and *nonadditive* otherwise;
- UPGMA has no guarantee to produce "additive" trees;
- UPGMA produces *ultrametric* trees, i.e. a tree such that the distance from the root to any leaf is the same;
- UPGMA assumes that evolution proceeds at a fixed constant rate, the so called molecular clock hypothesis;
- Other distance-based methods, such as the *neighbour-joining* algorithm, do not assume the existence of a molecular clock; good approximation of the least squares methods, fast (works well with hundreds of species).

Summary

- Distance-based methods are taking as input a multiple sequence alignment (n sequences by m columns). A distance is calculated for each pairwise alignment in order to fill an $n \times n$ (distance) matrix. The distance matrix is used to infer the topology of a tree, as well as the length of its branches.
- **Neighbour-joining** is the most widely used distance-based approach. It produces un-rooted, additive (but not necessarily ultrametric) trees. It is an iterative algorithm that requires $\mathcal{O}(n^3)$ time.
- Distance-based methods are fast!

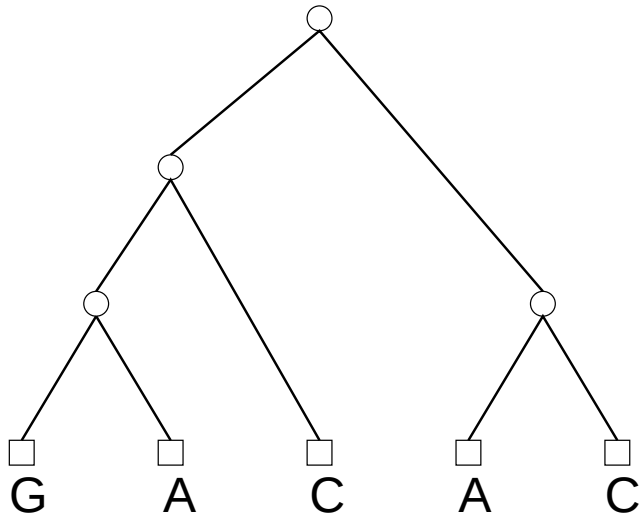
Remarks

- "However, the real problems arise when the data are not additive. Then, one has to find a tree whose distances *best approximate* the given data."
- "(...) the problem [distance-based tree reconstruction] remains open as there is no approach that both leads to a provably efficient algorithm and that follows a completely accepted definition of a *good approximation*." [3, page 448]
- Information is lost when reducing a pairwise alignment to a single number! In particular, the reconstructed tree says nothing about the "characters" of the internal nodes (ancestors) and the evolutionary events that occurred along its branches.
- Hence the need for alternatives.

II. Character-based tree reconstruction

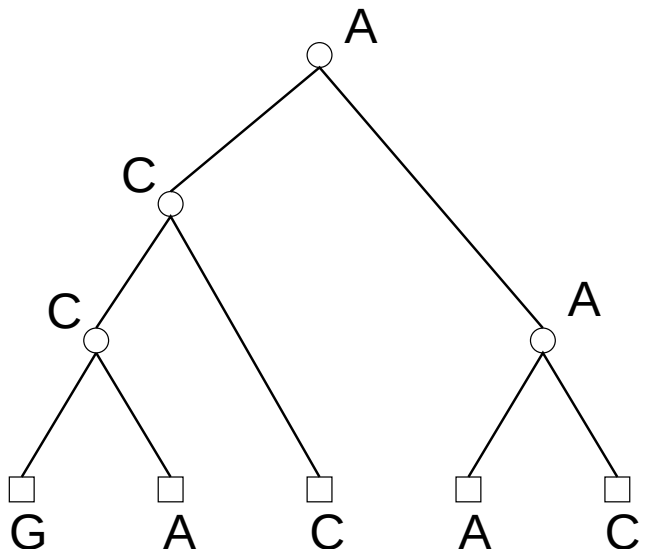
Character-based reconstruction algorithms are labelling all the nodes of the tree with characters. **Leaves** are labelled with **observed data**. While the **internal nodes** are labelled with **hypothetical characters** (ancestral states).

Character-based tree reconstruction



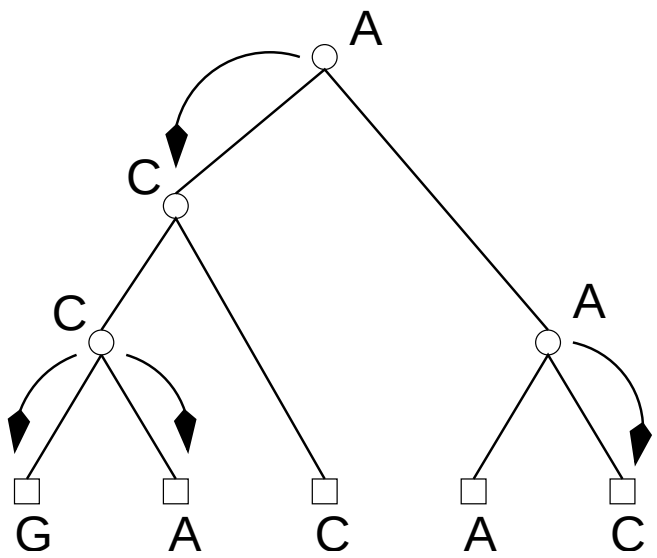
First, let's consider a single character (the ith nucleotide of a given gene in 5 species). The only observable characters are those at the leaves. Those correspond to the characters in todays organisms.

Character-based tree reconstruction



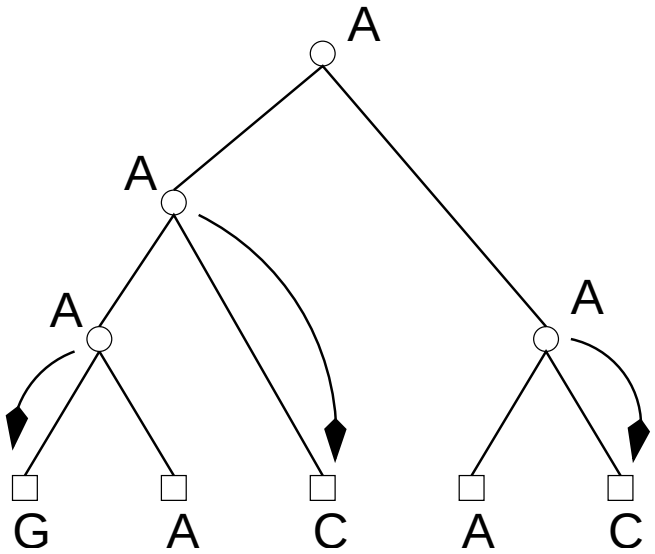
- Several reconstructions of the ancestral states are possible.
- How many events are represented on this tree?

Character-based tree reconstruction



The tree represents 4 events.
Can you find a reconstruction that requires fewer events?

Character-based tree reconstruction



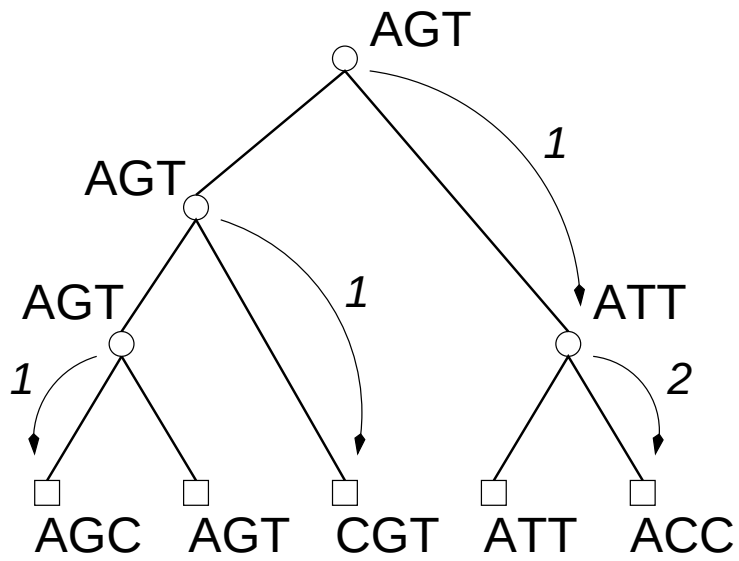
A 3 events tree.

Character-based tree reconstruction

Now considering 3 characters (sites).

	Sites		
Species	1	2	3
A	A	G	C
B	A	G	T
C	C	G	T
D	A	T	T
E	A	C	C

Character-based tree reconstruction



A tree for five species and three characters. The reconstruction involves 5 mutations (evolutionary events).

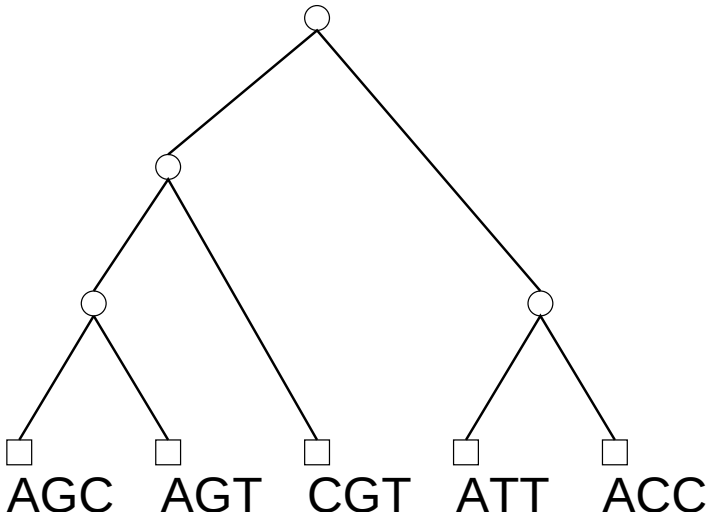
Parsimony

- "Adoption of the simplest assumption in the formulation of a theory or in the interpretation of data, especially in accordance with the rule of Ockham's razor." **The American Heritage [Online] Dictionary**
- **Ockham's Razor**: "Plurality should not be posited without necessity."
- "(1) Mutations are exceedingly rare events and (2) the more unlikely events a model invokes, the less likely the model is to be correct. As a result, the relationship that requires the fewest number of mutations to explain the current state of the sequences being considered is the relationship that is most likely to be correct." [5, page 98]

Issues

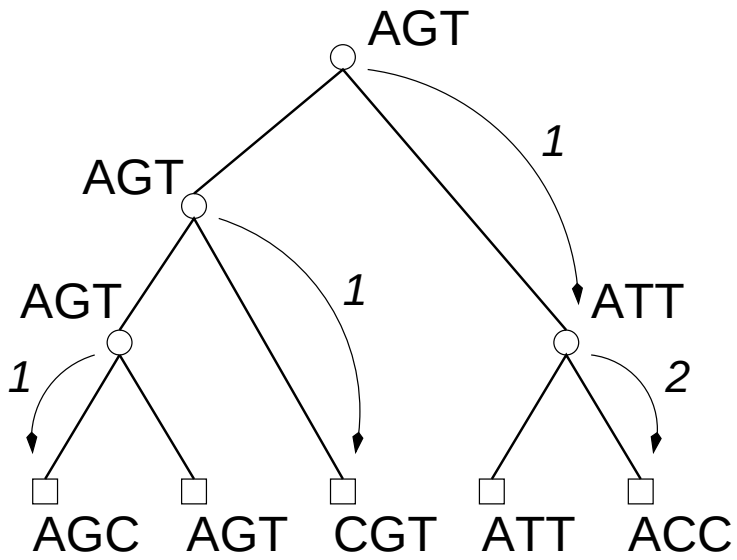
- Reconstructing the ancestral states;
- Counting the number of changes;
- Find all most parsimonious trees;
- Infer branch lengths;
- Is the most parsimonious tree the "real one"?
- Given several most parsimonious trees, is there a better one?

Small parsimony problem



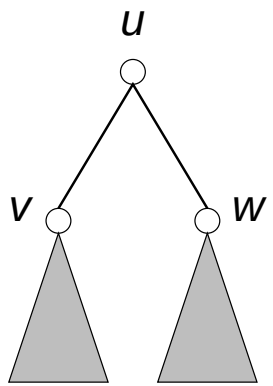
- **Problem:** *Find the most parsimonious labelling of the internal vertices in a given evolutionary tree.*
- **Input:** A tree T with each leaf labelled by an m-character array.
- **Output:** Labels (m-character arrays) for all the internal nodes such that $\Sigma d_H(u, v)$ for all the edges (u, v) is minimum; d_H is the Hamming distance.

Observation



Notice that the characters are independent. The total number of changes is the total number of changes for the first character plus the total number of changes for the second character plus the total number of changes for the third character. Thus, it suffices to develop a method that works for a single character and to apply it to all the characters. Proposals?

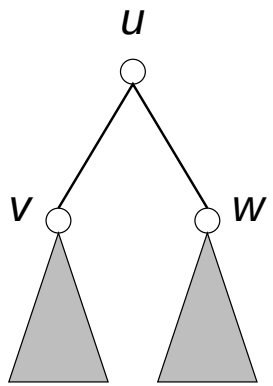
Small parsimony problem



Let's define $s_c(u)$ as the minimum parsimony score obtained when u is labelled with c. How to computer $s_c(u)$? What do you need to know? What are the dependencies?

$$s_c(u) = \ldots$$

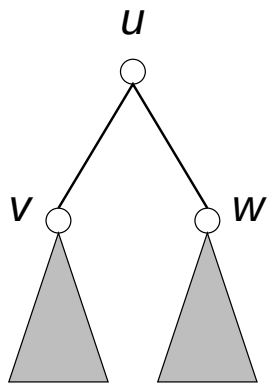
Small parsimony problem



For instance, what would be the most parsimonious score if u was labelled with A.

$$\begin{matrix} s_A(v)+? \\ s_C(v)+? \\ s_G(v)+? \\ s_T(v)+? \end{matrix} \quad ? \quad \begin{matrix} s_A(w)+? \\ s_C(w)+? \\ s_G(w)+? \\ s_T(w)+? \end{matrix}$$

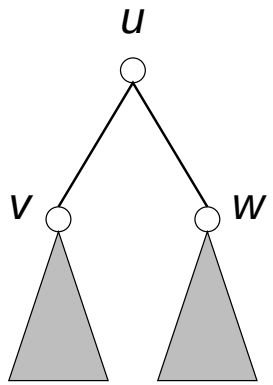
Small parsimony problem



For instance, what would be the most parsimonious score if u was labelled with A.

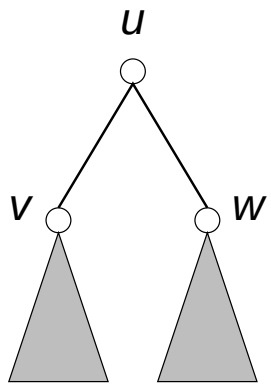
$$\begin{matrix} s_A(v)+0 \\ s_C(v)+1 \\ s_G(v)+1 \\ s_T(v)+1 \end{matrix} \quad ? \quad \begin{matrix} s_A(w)+0 \\ s_C(w)+1 \\ s_G(w)+1 \\ s_T(w)+1 \end{matrix}$$

Small parsimony problem



$$s_A(u) = \min \begin{cases} s_A(v) + 0 \\ s_C(v) + 1 \\ s_G(v) + 1 \\ s_T(v) + 1 \end{cases} + \min \begin{cases} s_A(w) + 0 \\ s_C(w) + 1 \\ s_G(w) + 1 \\ s_T(w) + 1 \end{cases}$$

Weighted small parsimony problem (Sankoff 1975)



$$s_c(u) = \min_i\{s_i(v) + \delta_{i,c}\} + \min_j\{s_j(w) + \delta_{j,c}\}$$

where $\delta_{j,c}$ is a $k \times k$ scoring matrix.

Examples of scoring matrices

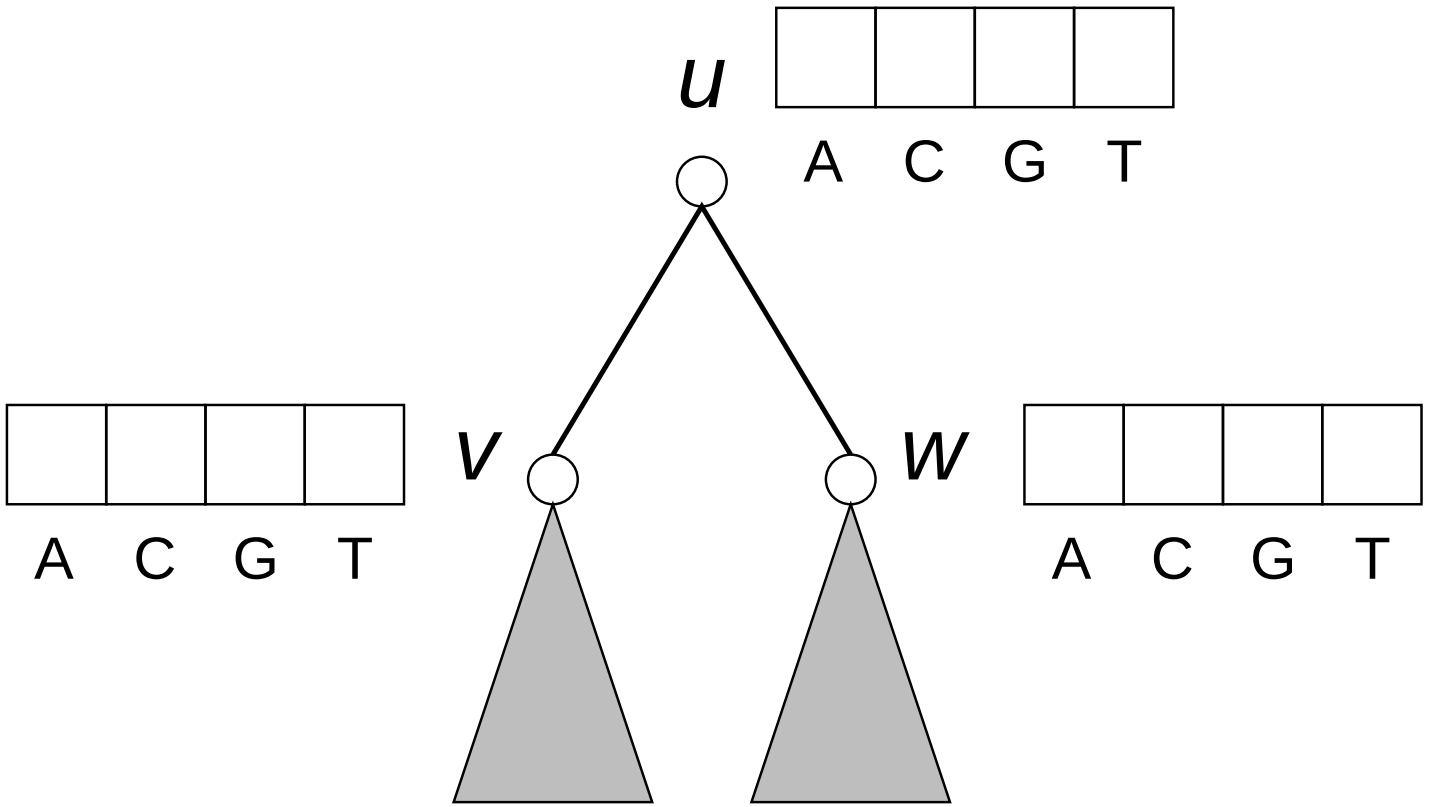
	A	C	G	T
A	0	1	1	1
C	1	0	1	1
G	1	1	0	1
T	1	1	1	0

	A	C	G	T
A	0	1	0.33	1
C	1	0	1	0.33
G	0.33	1	0	1
T	1	0.33	1	0

Solving the small parsimony problem

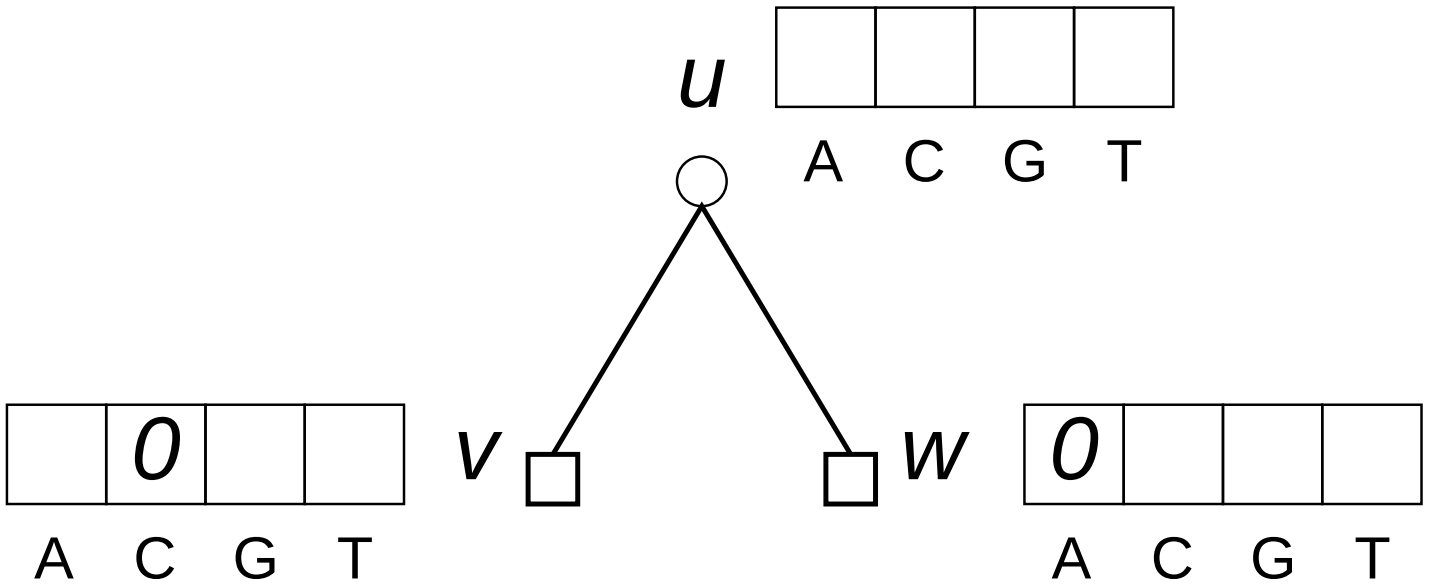
General case.

$$s_c(u) = \min_i\{s_i(v) + \delta_{i,c}\} + \min_j\{s_j(w) + \delta_{j,c}\}$$

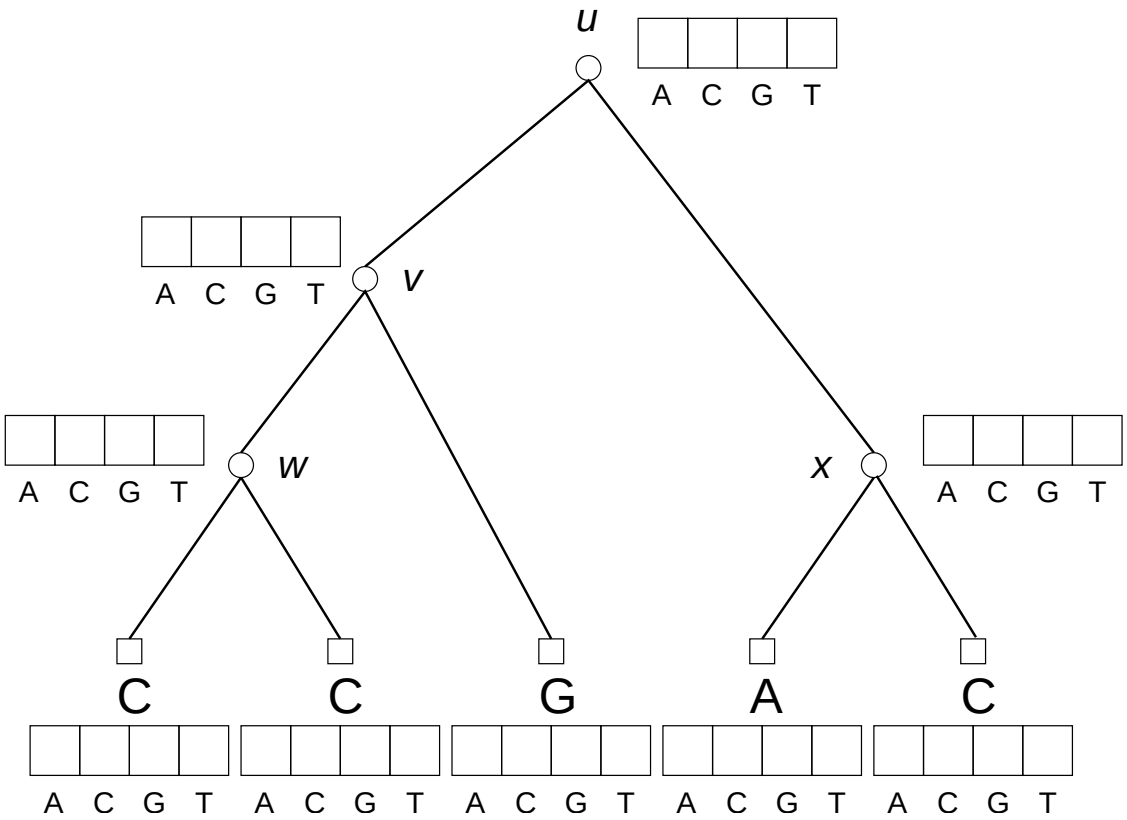


Solving the small parsimony problem

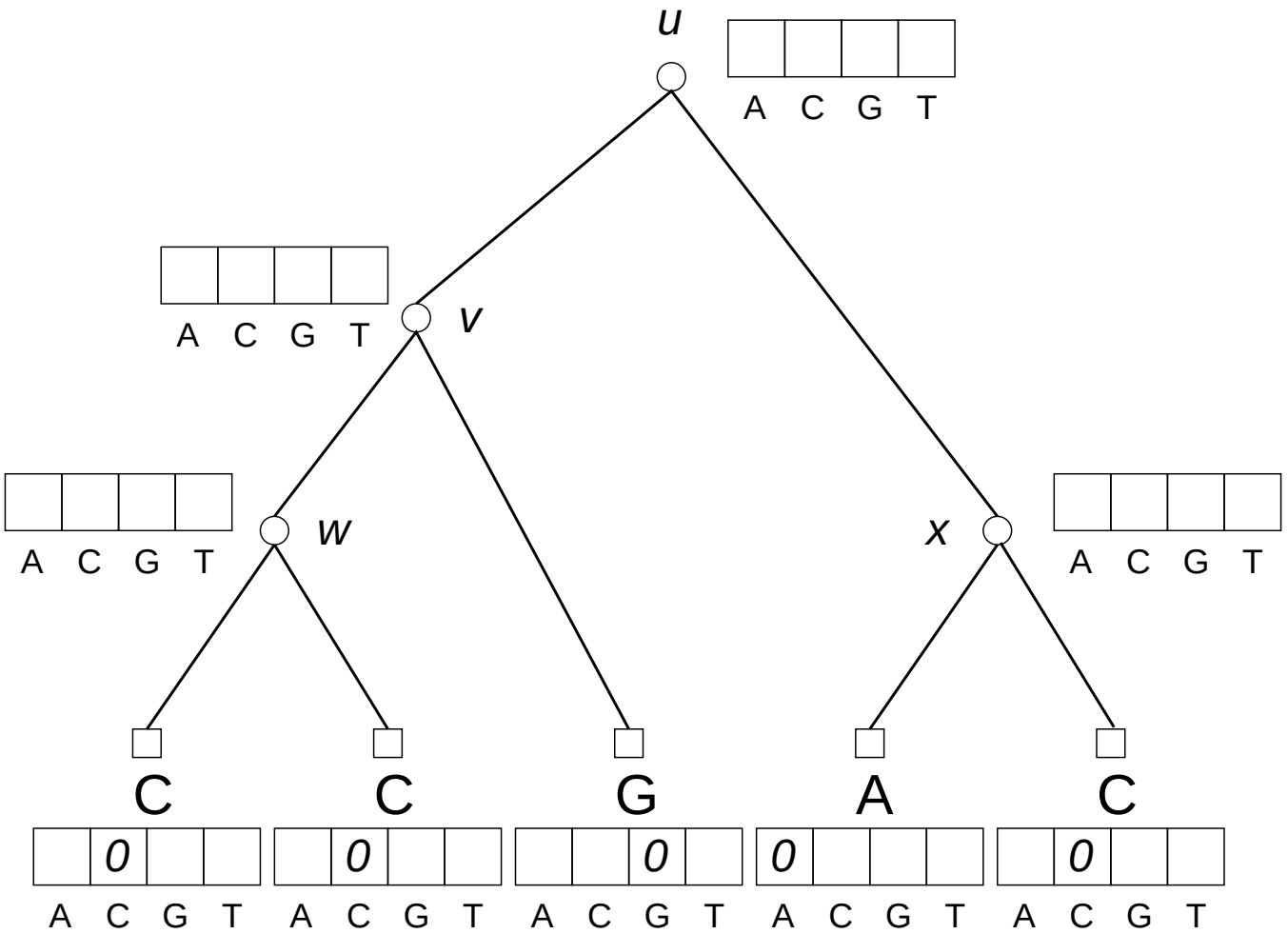
Initialisation.
For each leaf, $s_c(v) = 0$ if character c is found at that node and infinity otherwise.



Small parsimony problem



Small parsimony problem



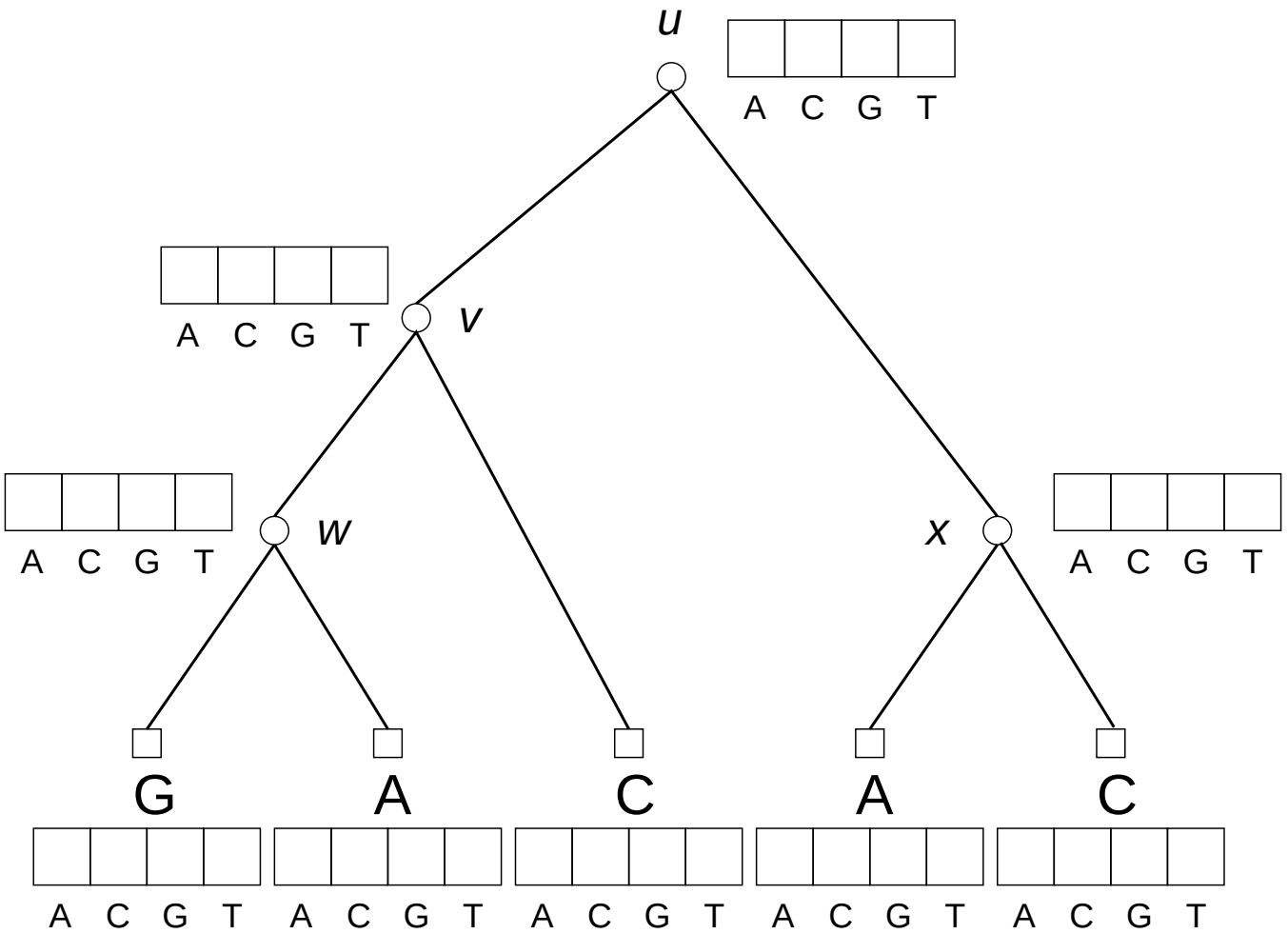
$$S_A(w) = 2 = \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\}$$
$$S_C(w) = 0 = \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\}$$
$$S_G(w) = 2 = \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\}$$
$$S_T(w) = 2 = \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}$$

$$S_A(x) = 1 = \min\{0 + 0, \infty + 1, \infty + 1, \infty + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\}$$
$$S_C(x) = 1 = \min\{0 + 1, \infty + 0, \infty + 1, \infty + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\}$$
$$S_G(x) = 2 = \min\{0 + 1, \infty + 1, \infty + 0, \infty + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\}$$
$$S_T(x) = 2 = \min\{0 + 1, \infty + 1, \infty + 1, \infty + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}$$

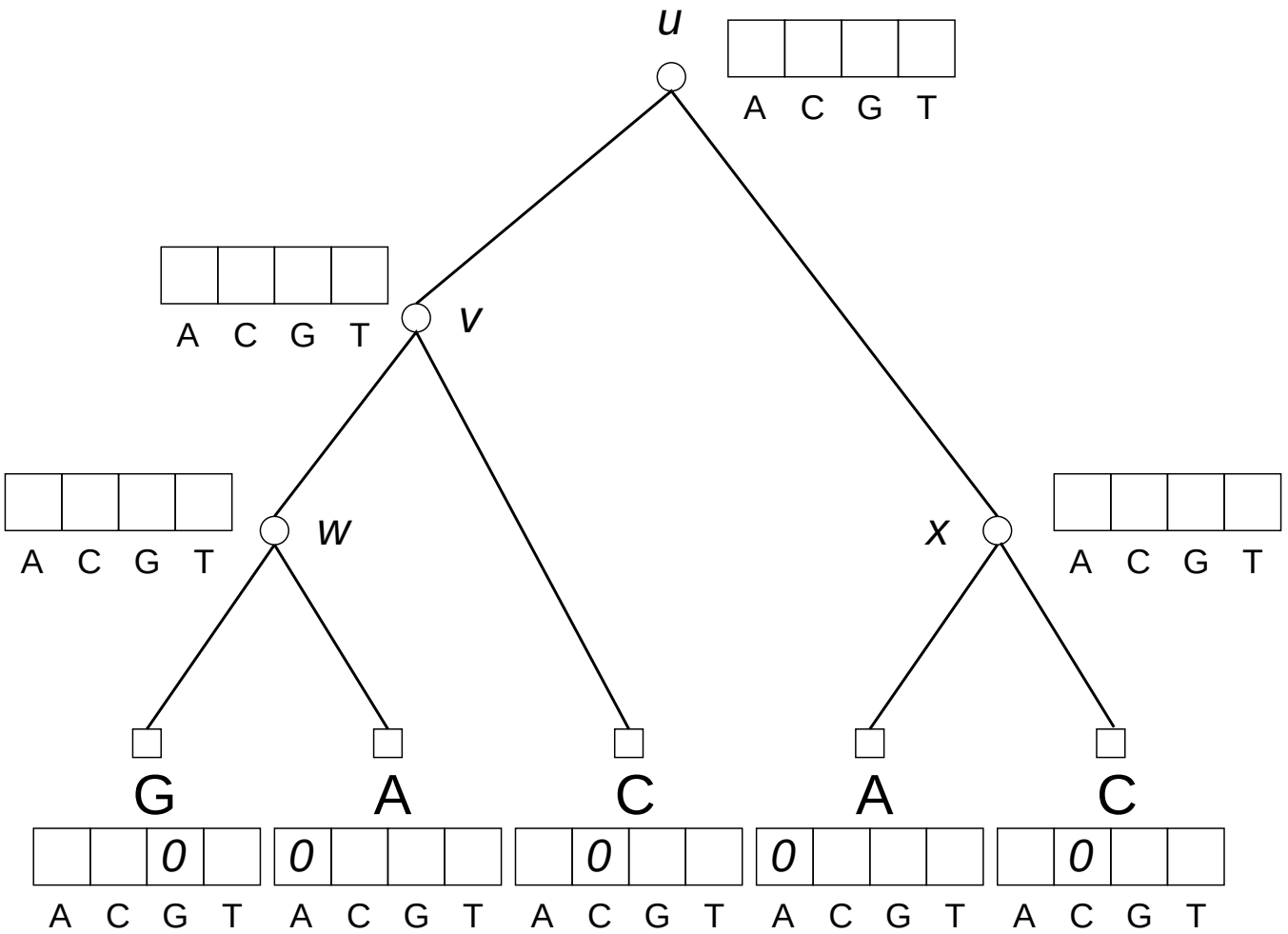
$$S_A(v) = 2 = \min\{2 + 0, 0 + 1, 2 + 1, 2 + 1\} + \min\{\infty + 0, \infty + 1, 0 + 1, \infty + 1\}$$
$$S_C(v) = 1 = \min\{2 + 1, 0 + 0, 2 + 1, 2 + 1\} + \min\{\infty + 1, \infty + 0, 0 + 1, \infty + 1\}$$
$$S_G(v) = 1 = \min\{2 + 1, 0 + 1, 2 + 0, 2 + 1\} + \min\{\infty + 1, \infty + 1, 0 + 0, \infty + 1\}$$
$$S_T(v) = 2 = \min\{2 + 1, 0 + 1, 2 + 1, 2 + 0\} + \min\{\infty + 1, \infty + 1, 0 + 1, \infty + 0\}$$

$$S_A(u) = 3 = \min\{2 + 0, 1 + 1, 1 + 1, 2 + 1\} + \min\{1 + 0, 1 + 1, 2 + 1, 2 + 1\}$$
$$S_C(u) = 2 = \min\{2 + 1, 1 + 0, 1 + 1, 2 + 1\} + \min\{1 + 1, 1 + 0, 2 + 1, 2 + 1\}$$
$$S_G(u) = 3 = \min\{2 + 1, 1 + 1, 1 + 0, 2 + 1\} + \min\{1 + 1, 1 + 1, 2 + 0, 2 + 1\}$$
$$S_T(u) = 4 = \min\{2 + 1, 1 + 1, 1 + 1, 2 + 0\} + \min\{1 + 1, 1 + 1, 2 + 1, 2 + 0\}$$

Small parsimony problem



Small parsimony problem



$$S_A(w) = 1 = \min\{\infty + 0, \infty + 1, 0 + 1, \infty + 1\} + \min\{0 + 0, \infty + 1, \infty + 1, \infty + 1\}$$
$$S_C(w) = 2 = \min\{\infty + 1, \infty + 0, 0 + 1, \infty + 1\} + \min\{0 + 1, \infty + 0, \infty + 1, \infty + 1\}$$
$$S_G(w) = 1 = \min\{\infty + 1, \infty + 1, 0 + 0, \infty + 1\} + \min\{0 + 1, \infty + 1, \infty + 0, \infty + 1\}$$
$$S_T(w) = 2 = \min\{\infty + 1, \infty + 1, 0 + 1, \infty + 0\} + \min\{0 + 1, \infty + 1, \infty + 1, \infty + 0\}$$

$$S_A(x) = 1 = \min\{0 + 0, \infty + 1, \infty + 1, \infty + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\}$$
$$S_C(x) = 1 = \min\{0 + 1, \infty + 0, \infty + 1, \infty + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\}$$
$$S_G(x) = 2 = \min\{0 + 1, \infty + 1, \infty + 0, \infty + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\}$$
$$S_T(x) = 2 = \min\{0 + 1, \infty + 1, \infty + 1, \infty + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}$$

$$S_A(v) = 2 = \min\{1 + 0, 2 + 1, 1 + 1, 2 + 1\} + \min\{\infty + 0, 0 + 1, \infty + 1, \infty + 1\}$$
$$S_C(v) = 2 = \min\{1 + 1, 2 + 0, 1 + 1, 2 + 1\} + \min\{\infty + 1, 0 + 0, \infty + 1, \infty + 1\}$$
$$S_G(v) = 2 = \min\{1 + 1, 2 + 1, 1 + 0, 2 + 1\} + \min\{\infty + 1, 0 + 1, \infty + 0, \infty + 1\}$$
$$S_T(v) = 3 = \min\{1 + 1, 2 + 1, 1 + 1, 2 + 0\} + \min\{\infty + 1, 0 + 1, \infty + 1, \infty + 0\}$$

$$S_A(u) = 3 = \min\{2 + 0, 2 + 1, 2 + 1, 3 + 1\} + \min\{1 + 0, 1 + 1, 2 + 1, 2 + 1\}$$
$$S_C(u) = 3 = \min\{2 + 1, 2 + 0, 2 + 1, 3 + 1\} + \min\{1 + 1, 1 + 0, 2 + 1, 2 + 1\}$$
$$S_G(u) = 4 = \min\{2 + 1, 2 + 1, 2 + 0, 3 + 1\} + \min\{1 + 1, 1 + 1, 2 + 0, 2 + 1\}$$
$$S_T(u) = 5 = \min\{2 + 1, 2 + 1, 2 + 1, 3 + 0\} + \min\{1 + 1, 1 + 1, 2 + 1, 2 + 0\}$$

Large parsimony problem

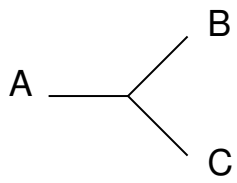
- **Problem:** *Find a tree having the minimum parsimony score.*
- **Input:** An $n \times m$ matrix (alignment).
- **Output:** A tree T with n leaves labeled by the n rows (m characters) of the input matrix. The internal nodes are labelled with arrays of m characters such that the overall parsimony score is minimum.
- The problem is known to be $\mathcal{NP}$-complete.

Exhaustive approach: 4 to 15 sequences

# Species	# unrooted trees
4	3
5	15
6	105
7	945
8	10,395
9	1,35,135
10	2,027,025
11	34,459,425
12	654,729,075
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625

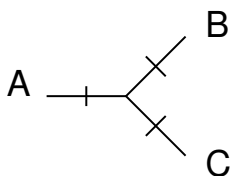
For a small number of species, say less than 15, it is be possible to exhaustively enumerate all the trees, and for each tree calculate the minimum parsimony score. The tree that has the overall minimum parsimony score is reported.

Sequential addition strategy



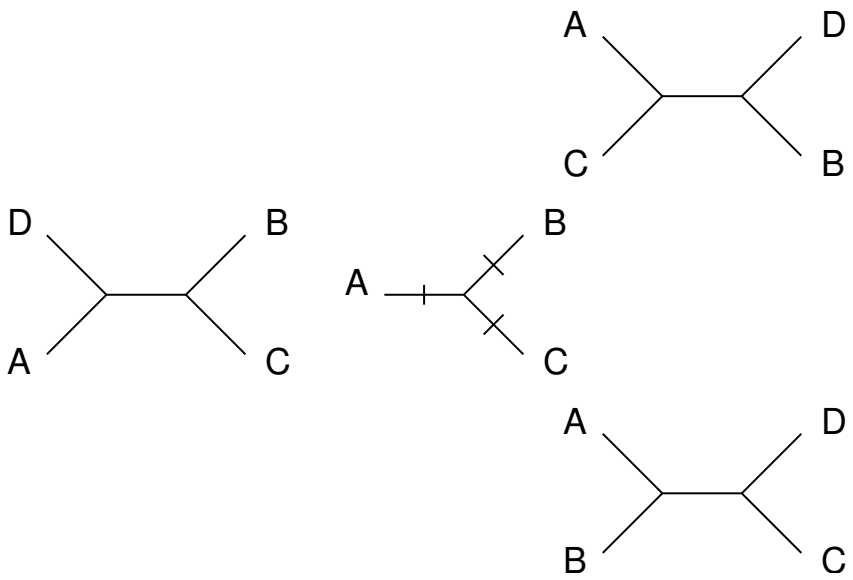
Given three species, there is a single unrooted tree.

Sequential addition strategy



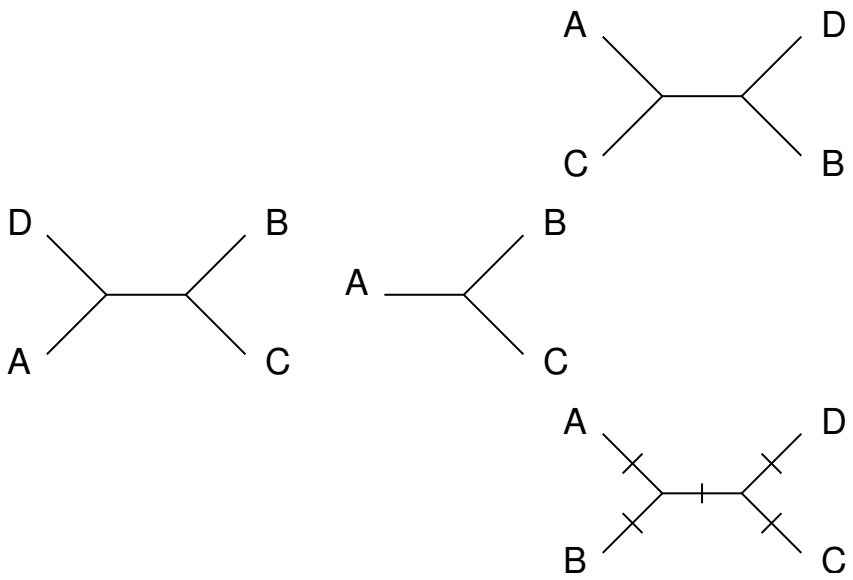
Each branch can serve as an insertion point, adding a new branch off the middle of any existing branch.

Sequential addition strategy



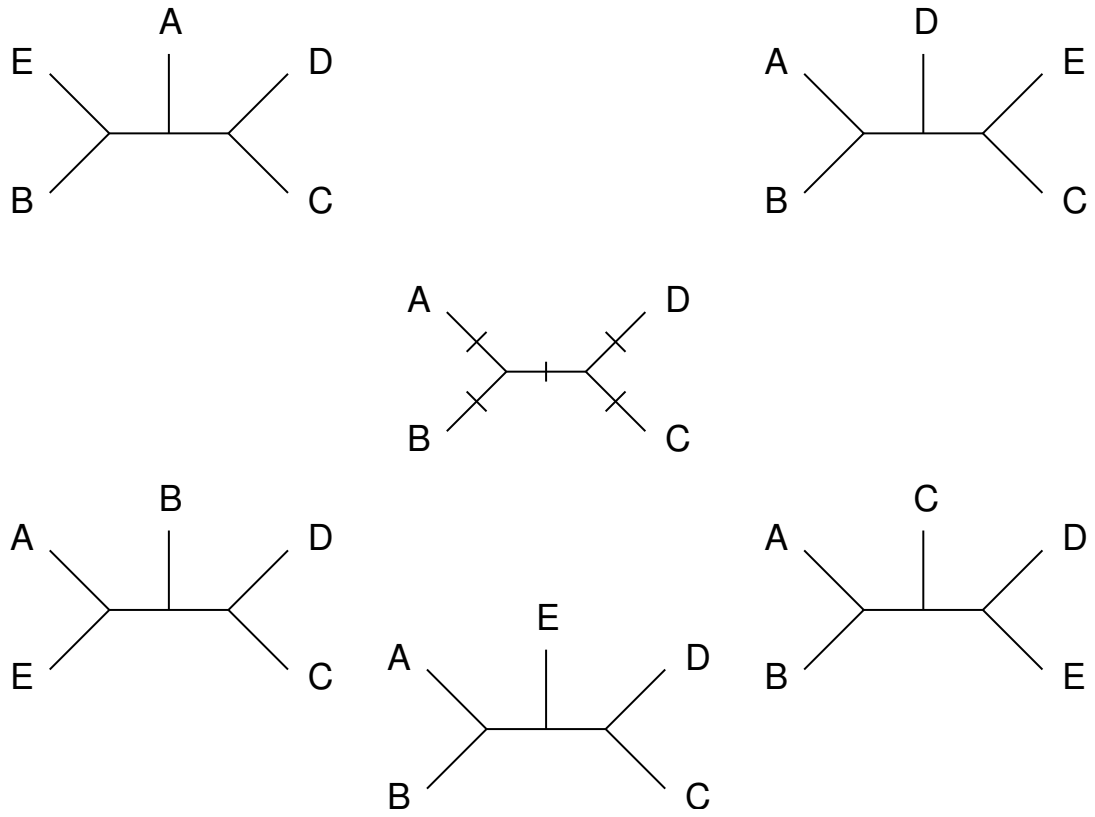
Therefore producing 3 four species unrooted trees.

Sequential addition strategy



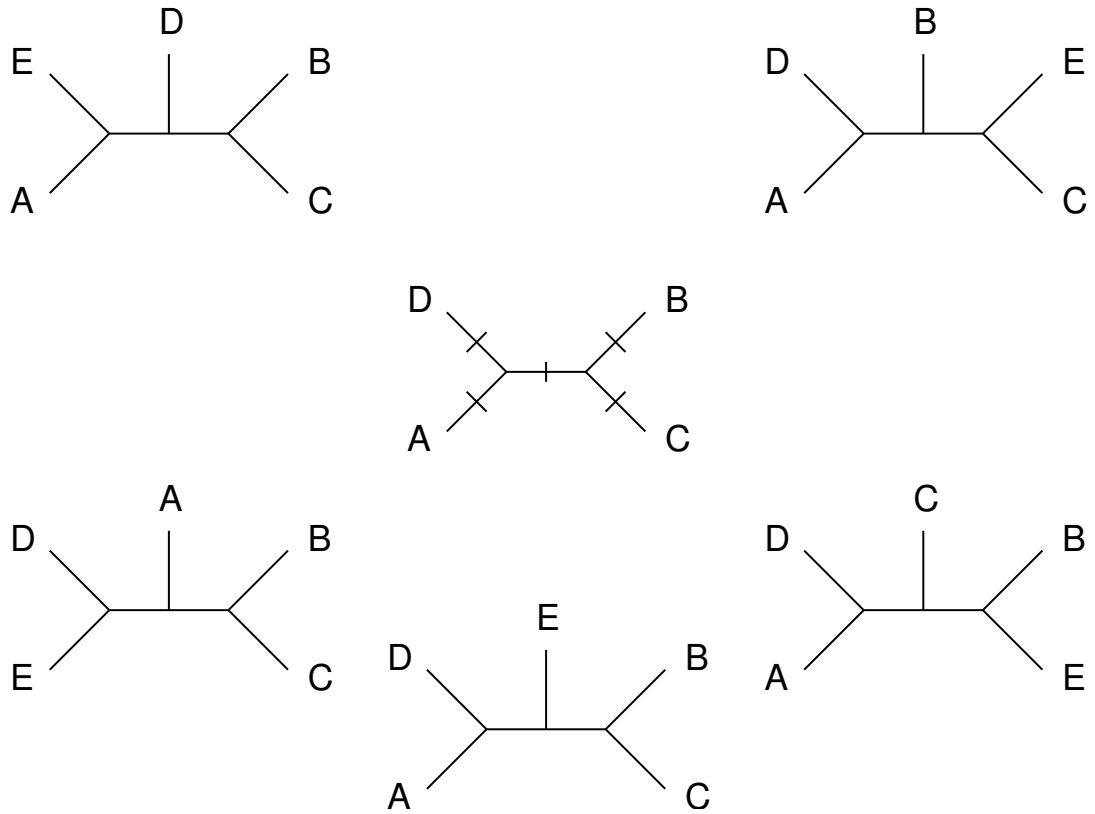
The same process is applied to all 3 four species trees.

Sequential addition strategy



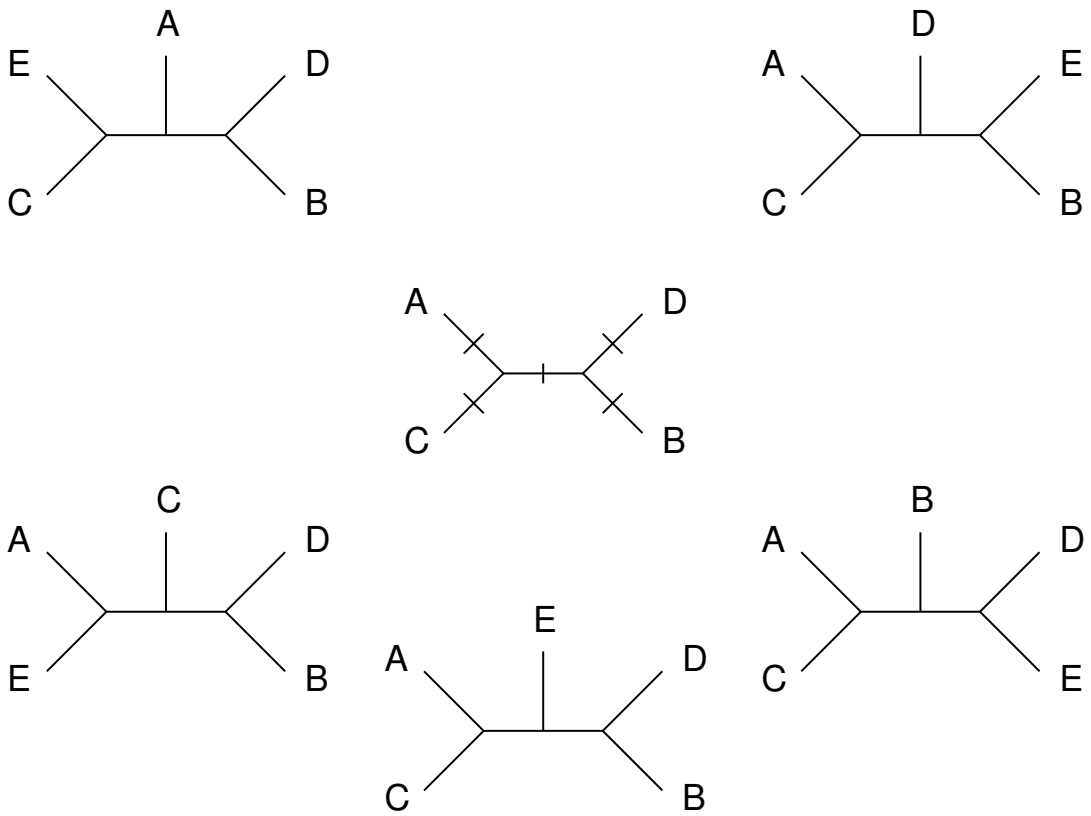
A four species unrooted tree has 5 edges, thus leading to 5 new unrooted trees.

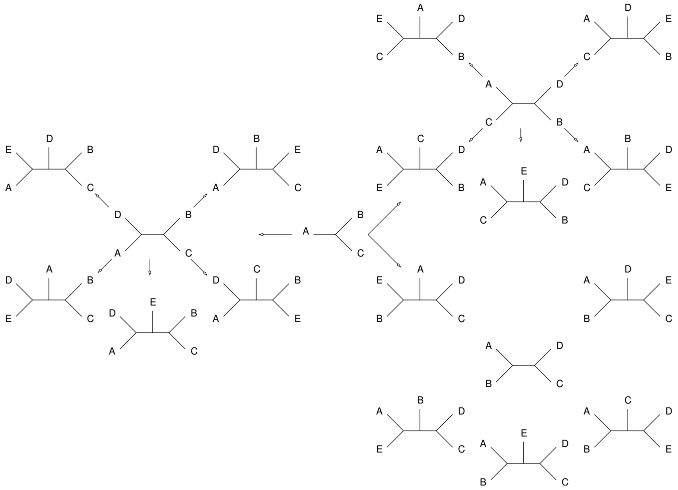
Sequential addition strategy



There will be 15 five species unrooted trees.

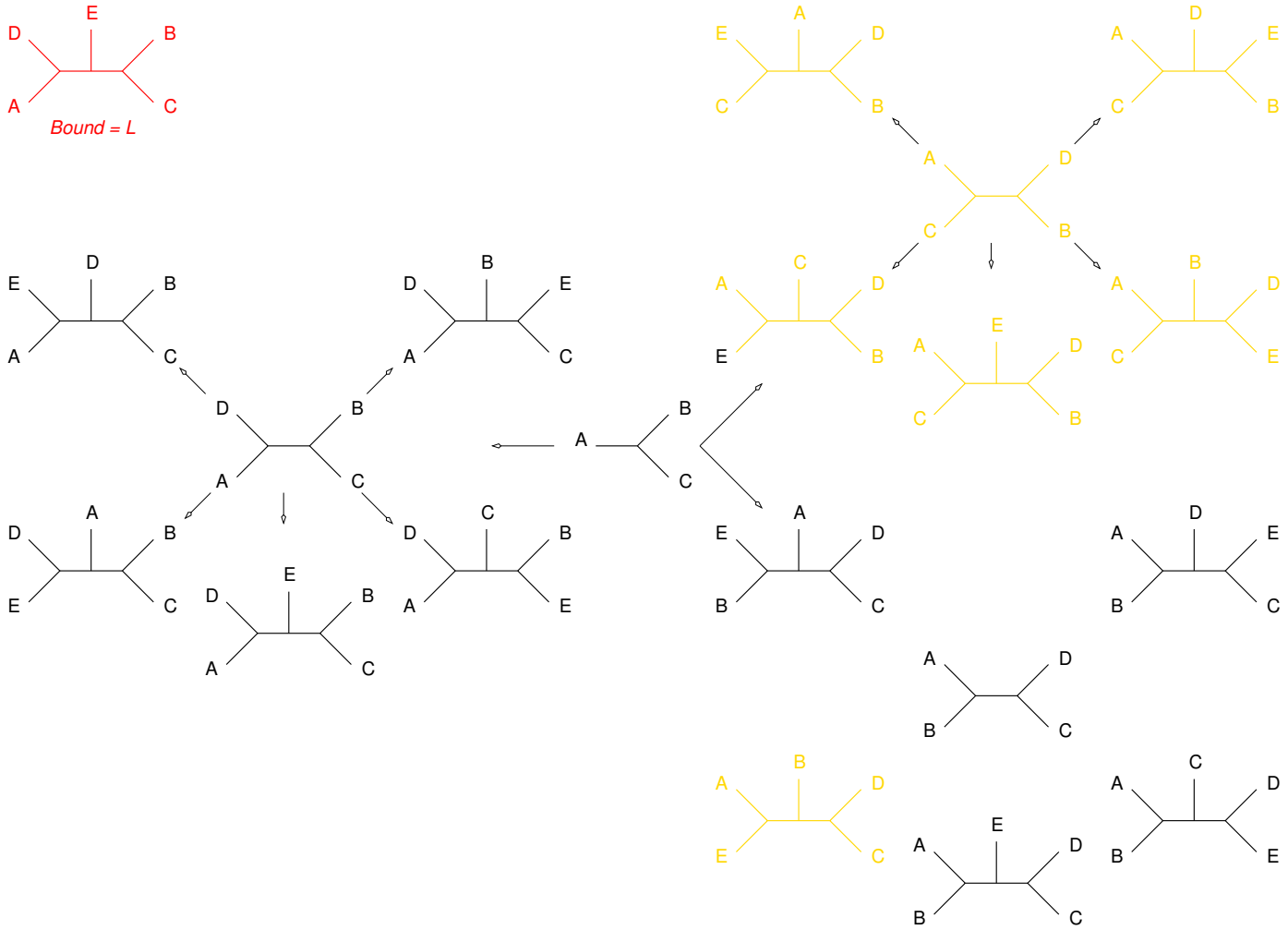
Sequential addition strategy

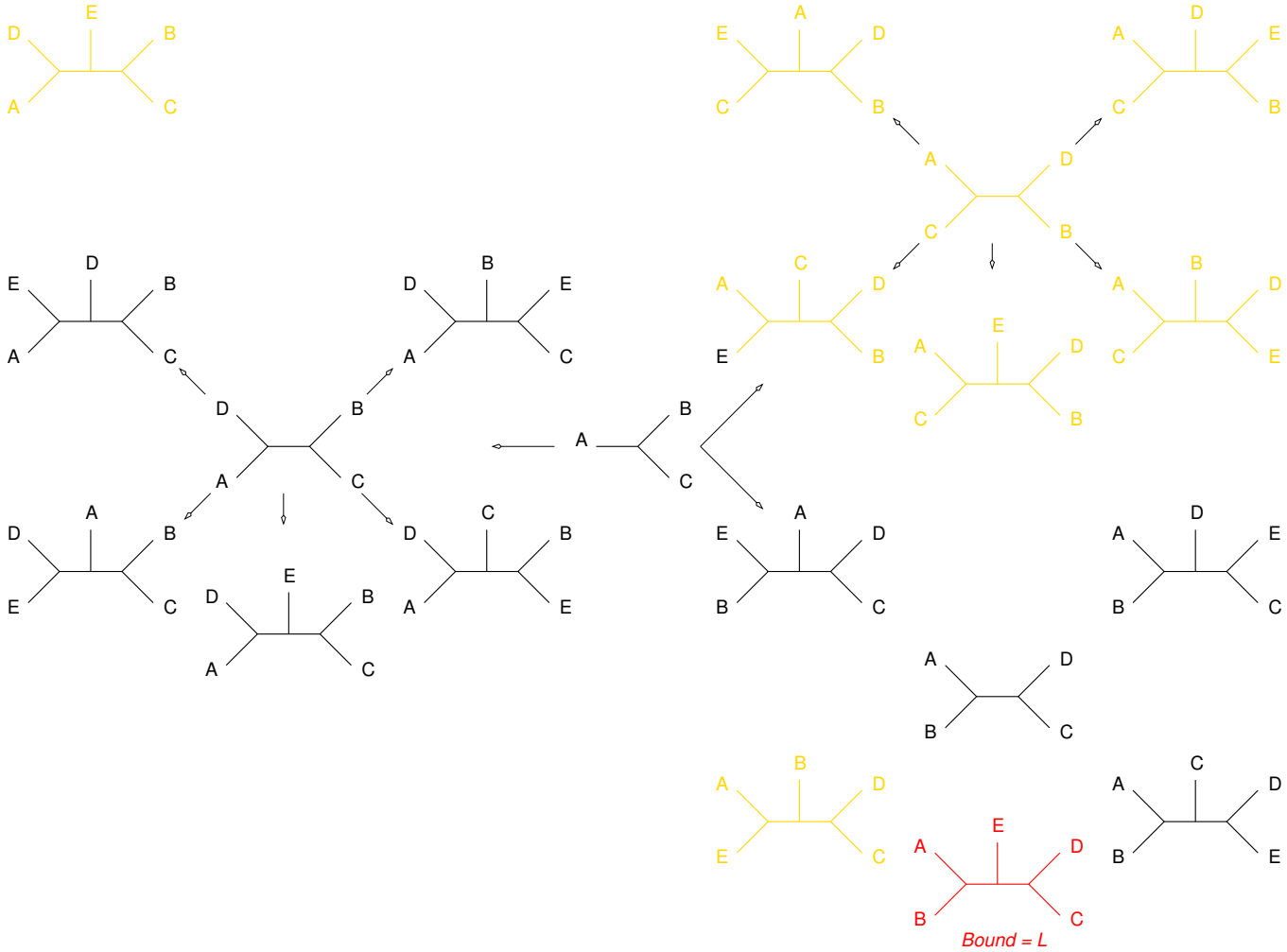


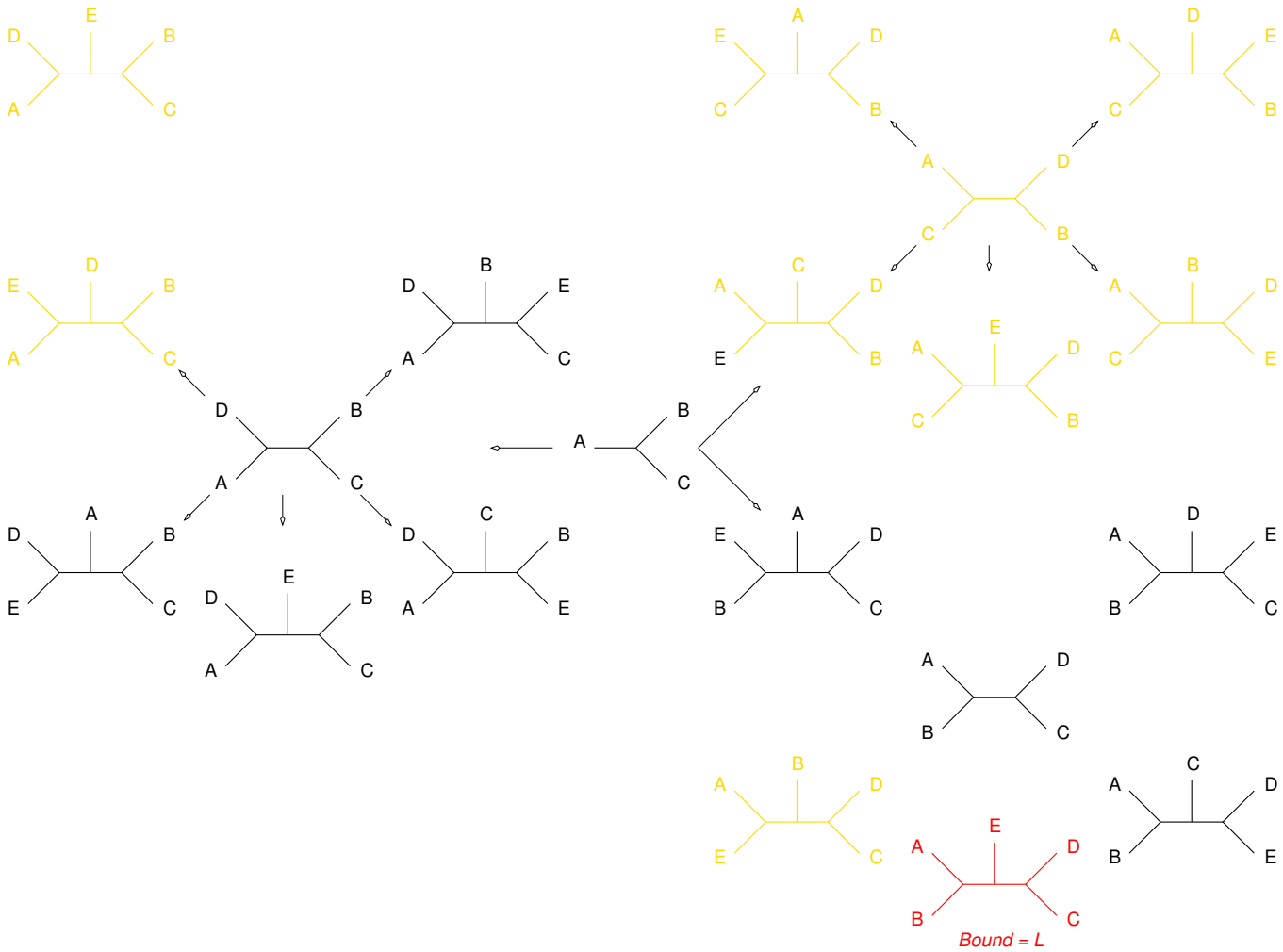


Branch and bound: 4 to 20 sequences

```
Generate an initial tree T (using UPGMA method for instance)
Compute L the minimum parsimony score of T ( lowest score so far )
Create two empty lists, open and solutions
Create an unrooted tree for three species and add it open
While open is not empty
  Remove the front element of the list and call it current
  Foreach tree t created by a sequential addition to current do
      If the minimum parsimony score of t is larger than L than discard
      If the minimum parsimony score of t is is lower than  L
         If t has n leaves:
            clear solutions
            add t to solutions
            set L to the minimum parsimony score of t
          Else add t to the rear of open
      Else (equals case)
          If t has n leaves: add t to solutions
          Else add t to the rear of open
solutions is the list of all the solutions, their score is L.
```


Bound = L

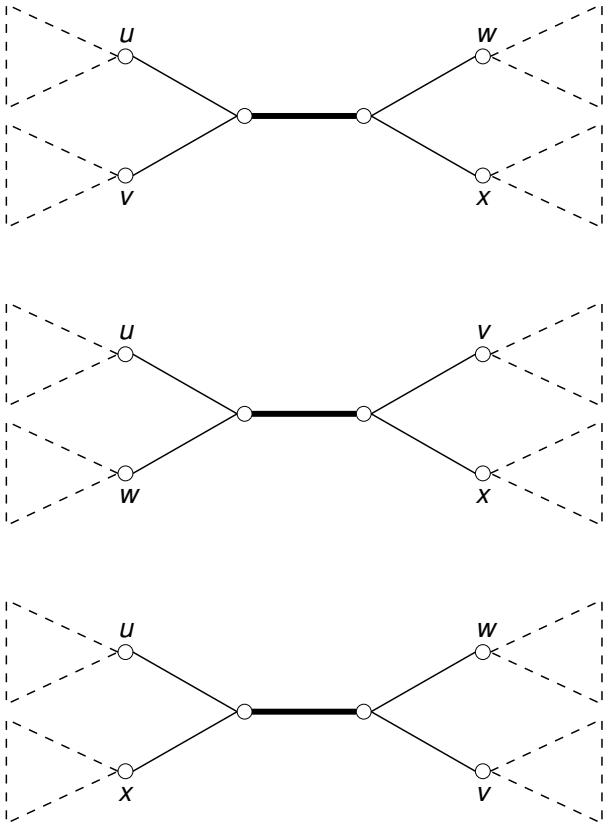

Bound = L


Bound = L

Greedy algorithm

1. Generate an initial topology (using UPGMA for instance);
2. Apply nearest neighbour interchange (NNI) transformations to all the internal edges;
3. Select the minimum parsimony tree;
4. Goto step 2.

$n > 20$: Nearest-neighbour interchange (NNI) local search heuristic



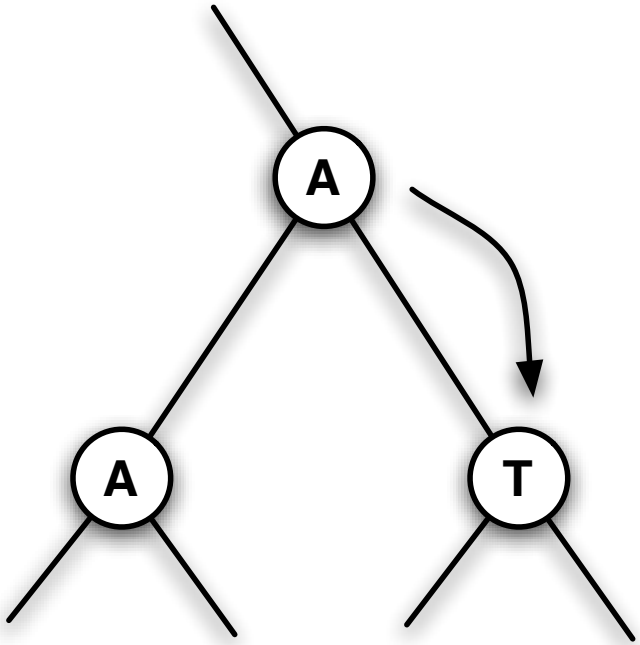
Other heuristics include: subtree pruning and regrafting (SPR) or tree bisection and reconnection (TBR).

Remarks: distance-based vs character-based

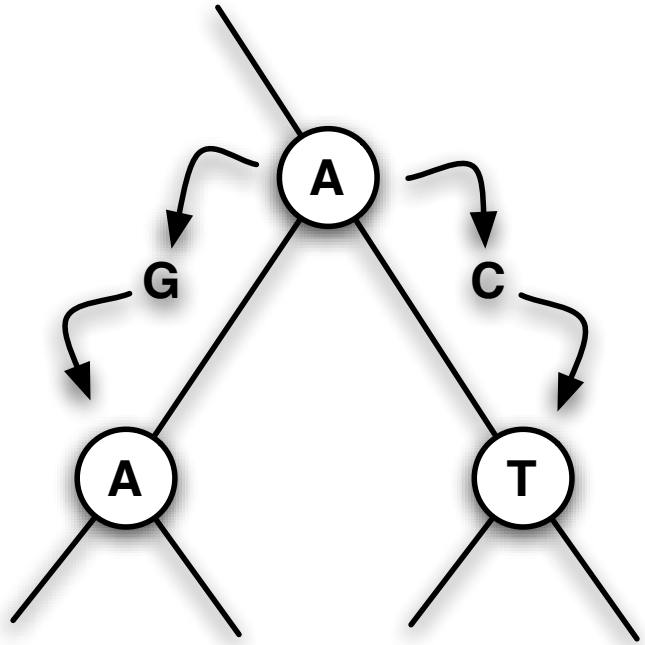
- Distance-based methods compute the pairwise sequence distances **1)** directly, **2)** in isolation, **3)** before inferring the tree topology
- Instead, for character-based methods, **1)** extant sequences are never compared directly **2)** the pairwise distances depend on the reconstructed ancestral sequences, and **3)** this process (solving the small phylogeny problem) takes all the sequences into account

Remarks

- The particular methods that were presented are not modeling the base substitutions accurately.
- Specifically, these methods are ignoring the fact that multiple substitutions (for a given site) are likely to occur in any given branch of the tree (time interval).



VS



Informal discussion!

III. Maximum likelihood methods

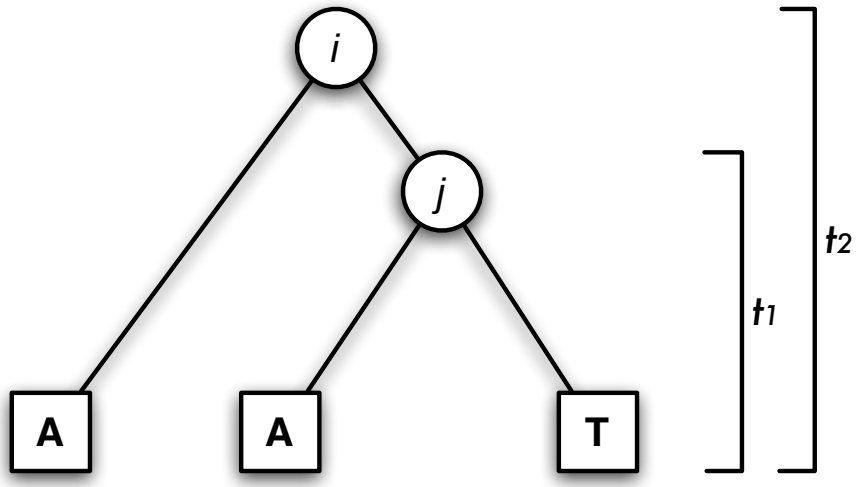
$P(D|\Theta)$ denotes the probability of the data given some model Θ (set of parameters, such as tree topology, branch length, evolutionary model...).
Let $L(\Theta) = P(D|\Theta)$ be the **likelihood function**.
The **maximum likelihood estimate** is the value of Θ that maximizes $L(\Theta)$.

III. Maximum likelihood methods

Let $L(\Theta)$ be the **likelihood** of a phylogenetic tree. $L(\Theta)$ is defined as the probability of the data (generally sequences) for a given tree (topology, branch length, evolutionary model),
$P(\text{observed sequences}|\text{tree})$.
A **maximum likelihood** approach finds a tree, amongst all possible trees, with the largest value of $L(\Theta)$. Such tree explains best the data.
[See Felsenstein 2004, pages 251–253.]
Assumptions that are generally made:

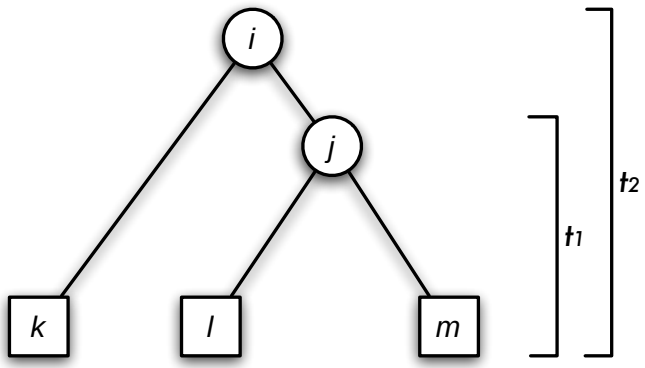
1. Sites are independent
2. Lineages are independent

Probability of a tree



$$\sum_i \sum_j p_i q_{iA}(t_2) q_{ij}(t_2 - t_1) q_{jA}(t_1) q_{jT}(t_1)$$

Probability of a tree (cont.)



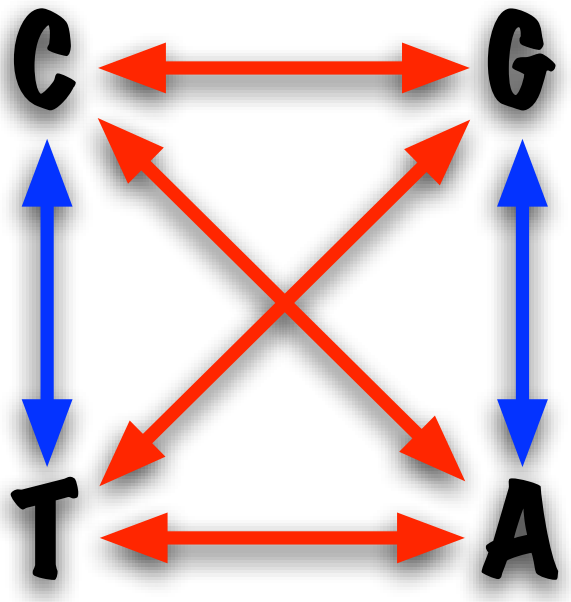
$$\sum_i \sum_j p_i q_{ik}(t_2) q_{ij}(t_2 - t_1) q_{jl}(t_1) q_{jm}(t_1)$$

where k, l, m are nucleotide types found at the given sequence position in the 3 organisms under study. Assuming that the positions (sites) are independent one from another (are evolving independently), the probability of the tree would be the product over all site probabilities.

Probability of a tree (cont.)

The $q_{ij}(t)$ terms give the probability of finding the nucleotide type j at a given site knowing that its ancestor had the nucleotide type i at the same position at time t (earlier).
Examples of substitution schemes modeling multiple substitutions for a given time interval include Jukes-Cantor one-parameter model and Kimura's two-parameter model.

Probability of a tree: model of evolution



Transition rate: blue and **transversion** rate: red

Probability of a tree: model of evolution

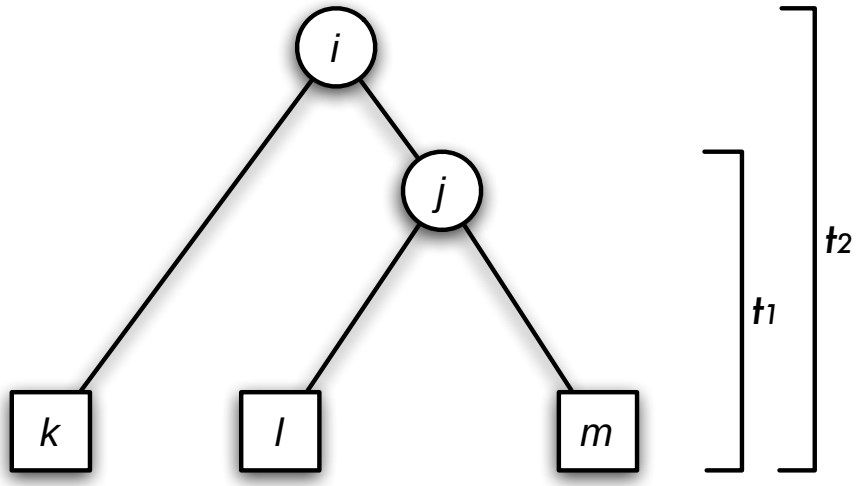
JC69: Jukes and Cantor 1969; bases are equiprobable; transition rate = transversion rate

F81: Felsenstein 1981; variable base composition; transition rate = transversion rate

K80: Kimura 1980; bases are equiprobable; transition rate $\neq$ transversion rate

HKY85: Hasegawa *et al.* 1985; variable base composition; transition rate $\neq$ transversion rate; variable transition and transversion rates

Jukes and Cantor 1969



$$p(j|i,t) = \begin{cases} \frac{1}{4}(1+3e^{-4\alpha t}) & \text{if } j = i \\ \frac{1}{4}(1-e^{-4\alpha t}) & \text{if } j \neq i \end{cases}$$

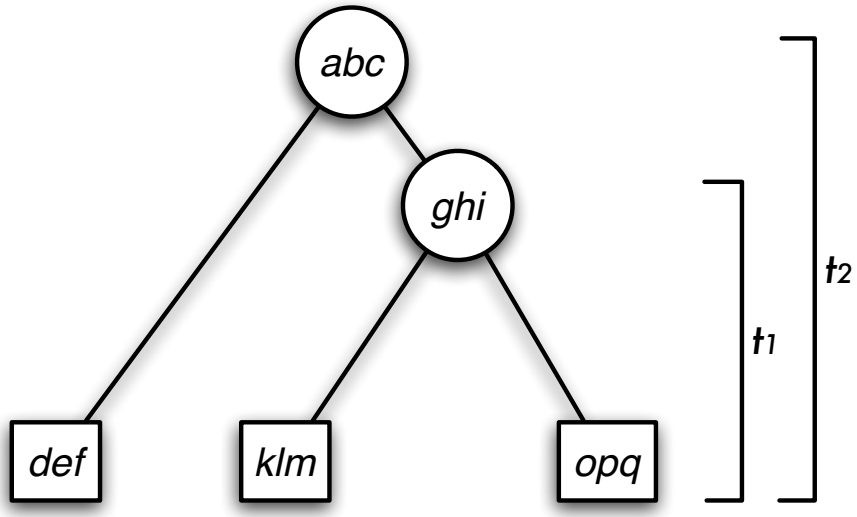
where α is the mutation rate parameter.

Probability of a tree: model of evolution

In addition to the base model, most methods allow for relaxations:

- Variable rates across positions ($+\Gamma$)
- Variable rates across lineages ($+\mathcal{I}$)

Probability of a tree: model of evolution



HKY85+Γ + $\mathcal{I}$ implies variable base composition, transition rate $\neq$ transversion rate, variable transition and transversion rates, that vary across sites and lineages.

Probability of a tree: model of evolution

- These models are also used to estimate pairwise distances for building phylogenies using distance-based approaches (e.g. Neighbour-joining).

Maximum likelihood methods

- Let L be the **likelihood** of a phylogenetic tree. L is defined as the probability of the data for a given tree, $P(\text{observed sequences}|\text{tree})$.
- A **maximum likelihood** approach finds a tree, amongst all possible trees, with the largest value of L. Such tree explains best the data.
- **Furthermore, the length of the branches are unknown and must be estimated as part of this process.**
- Finding an exact solution to this problem is impractical when the number of input sequences is large, say 5 sequences/species.
- Heuristic techniques have been developed to explore the tree space.

Maximum likelihood methods

- Generate an initial tree topology $\mathcal{T}$ (e.g. using NJ)
- Calculate its likelihood $\mathcal{L}$
- For a fixed number of iterations
  - From $\mathcal{T}$, generate new trees using NNI, SPR or TBR
  - For each new tree $\mathcal{T}'$, calculate its likelihood $\mathcal{L}'$
  - $\mathcal{L} = \mathcal{L}'$ and $\mathcal{T} = \mathcal{T}'$ if $\mathcal{L}' > \mathcal{L}$

Remarks: Parsimony vs Maximum Likelihood

- For a given tree topology, maximum parsimony considers all the reconstructions that lead to the same optimal score
- Maximum parsimony under-estimates the number of evolutionary events (because of the multiple substitutions along the branches of the tree)
- Maximum likelihood, through its evolutionary models, takes into account multiple substitutions, rate variations amongst sites and lineages, etc.
- For a given tree topology, maximum likelihood considers all the reconstructions (and not only the most parsimonious ones)
- This is the most time consuming approach of all three

Other issues: informative sites

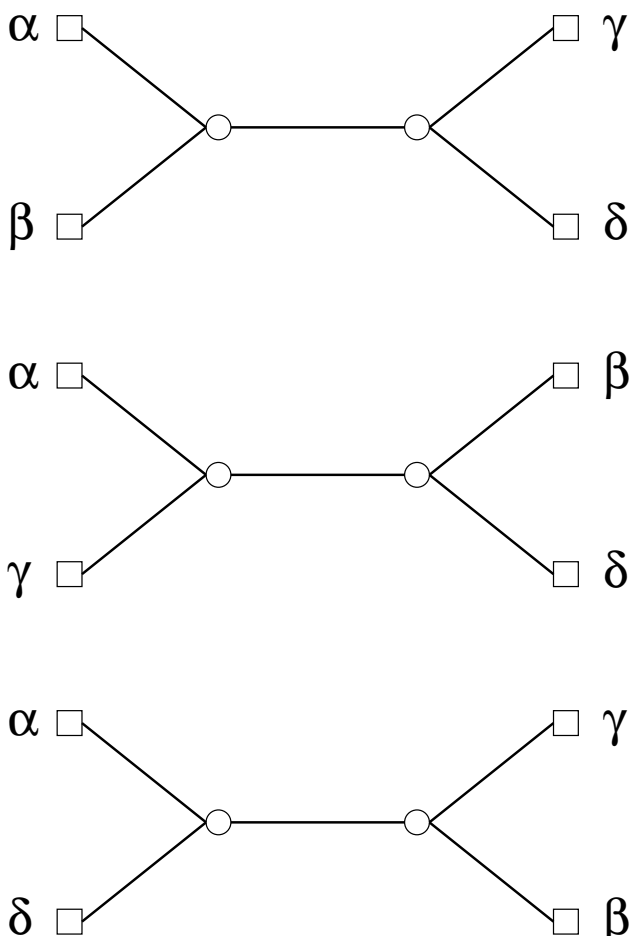
- Intuitively, the sites (columns) of an alignment that contain a single nucleotide type (**invariant** sites) provide no useful information for building a phylogenetic tree using a character-based approach.
- A site is **informative** if it allows to discriminate between trees, i.e. the minimum parsimony scores for at least two trees are different for that site, otherwise the site is **uninformative**.
- Clearly, invariant sites are uninformative.

Informative sites

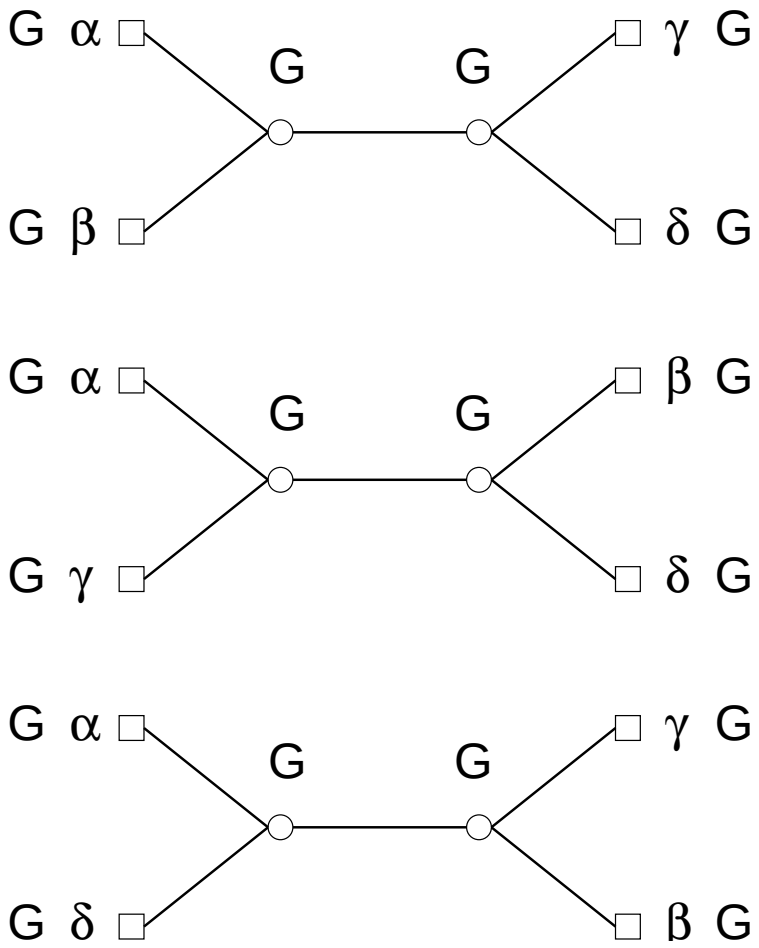
	Site					
Species	**1**	**2**	**3**	**4**	**5**	**6**
α	G	G	G	G	G	G
β	G	G	G	A	G	T
γ	G	G	A	T	A	G
δ	G	A	T	C	A	T

$\Rightarrow$ Adapted from [5, pages 99–101].

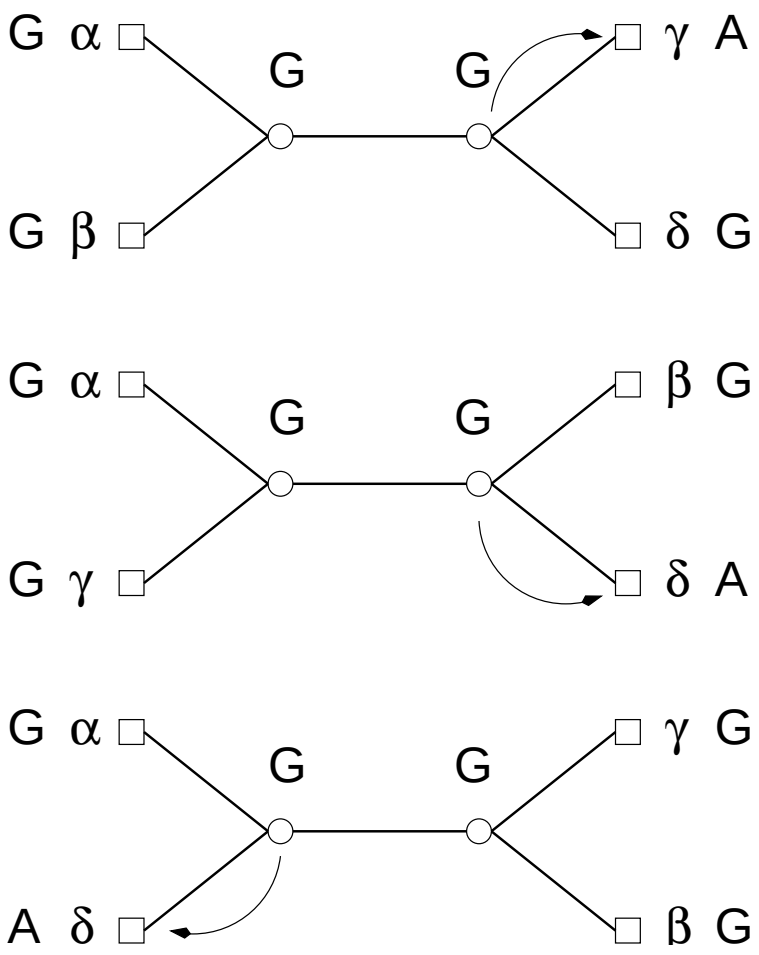
Given 4 species, there are 3 possible unrooted trees



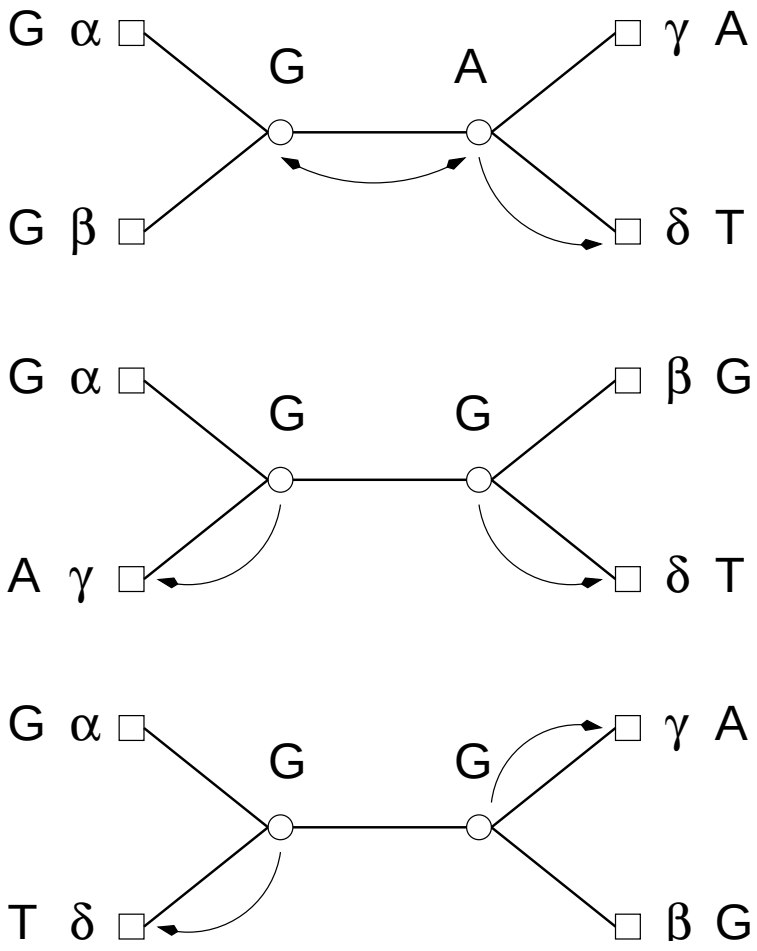
Site 1



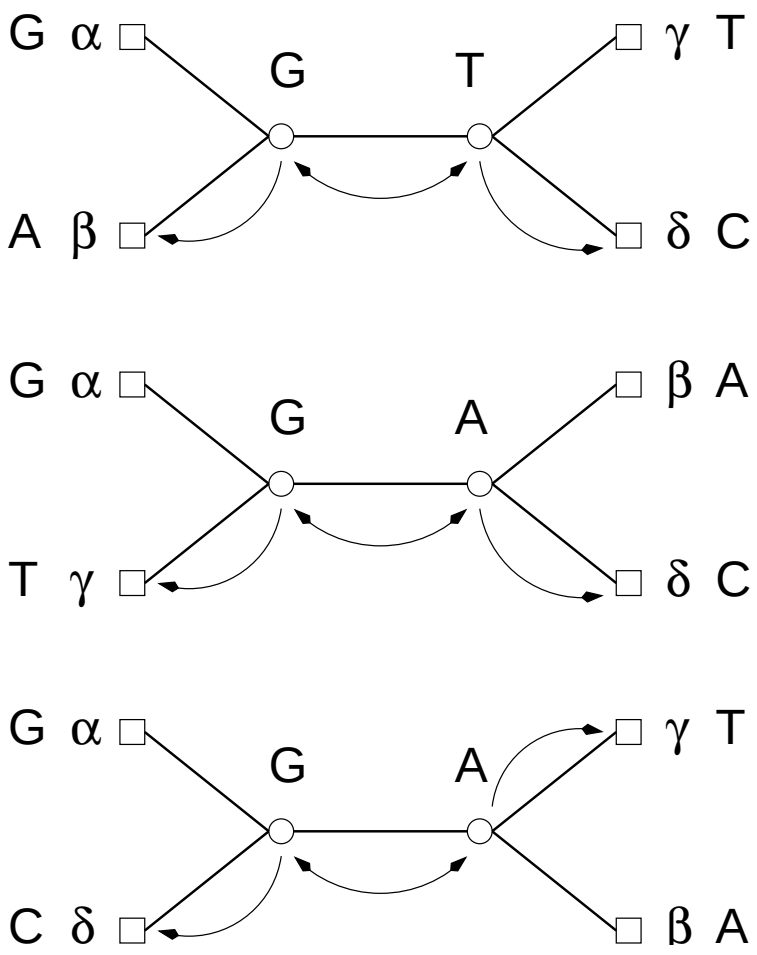
Site 2



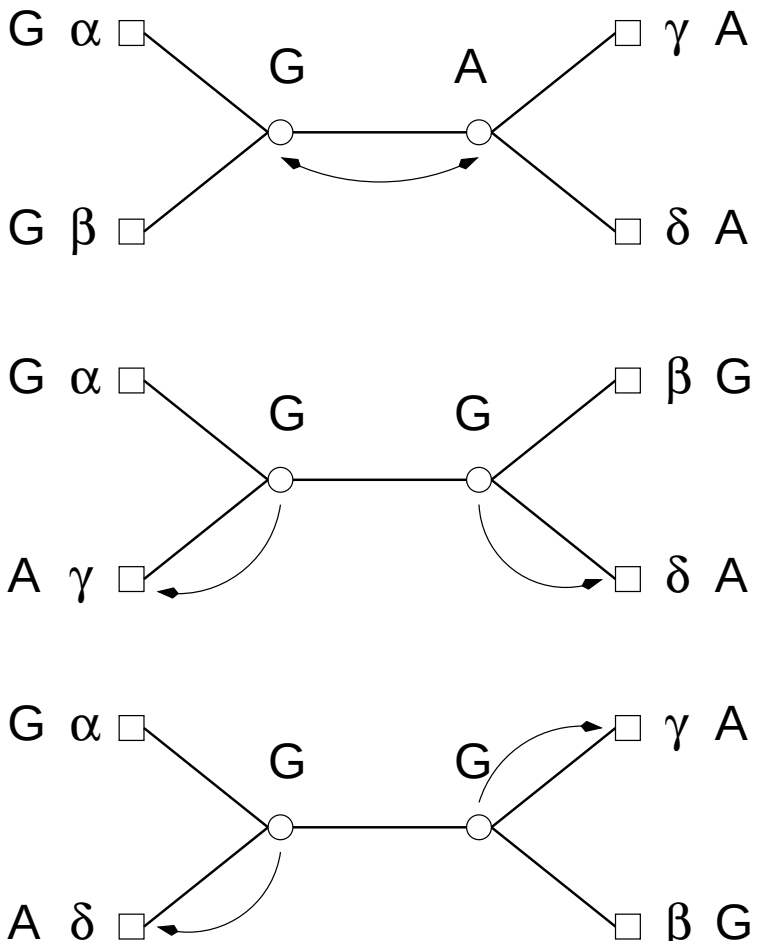
Site 3



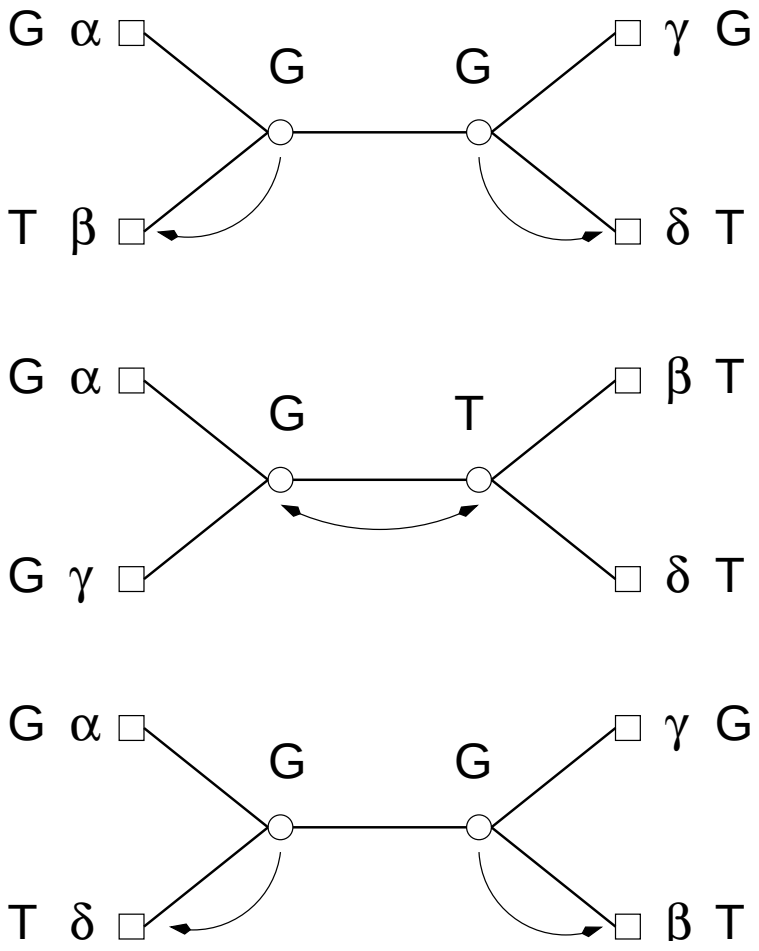
Site 4



Site 5



Site 6



Informative sites

- Fortunately, there is a simple rule to identify informative sites. There has be at least two types that are occur at least twice at that site.
- Uninformative sites are discarded prior to the inference of the tree.
- Notice that those sites are typically kept by distance-based methods. This partly explains why the methods are producing different results.

Further readings

- Bayesian inference of phylogenetic trees
- Quartet methods
- Evolutionary networks (as opposed to evolutionary trees)
- Bootstraps, consensus, comparing trees
- Matching interior nodes (taxonomic units) with paleontological information, so as to assign time to events

Bibliography


R. Durbin, S. Eddy, A. Krogh, and G. Mitchison.
Biological Sequence Analysis.
Cambridge University Press, 1998.


J. Felsenstein.
Inferring Phylogenies.
Sinauer, 2004.


D. Gusfield.
Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.
Cambridge University Press, 1997.


N. C. Jones and P. A. Pevzner.
An introduction to bioinformatics algorithm.
MIT Press, 2004.

Bibliography (cont.)


D. E. Krane and M. L. Raymer.
Fundamental Concepts of Bioinformatics.
Benjamin Cummings, 2003.


W.-H. Li and D. Graur.
Fundamentals of Molecular Evolution.
Sinauer, 1991.

Please don't print these lecture notes unless you really need to!