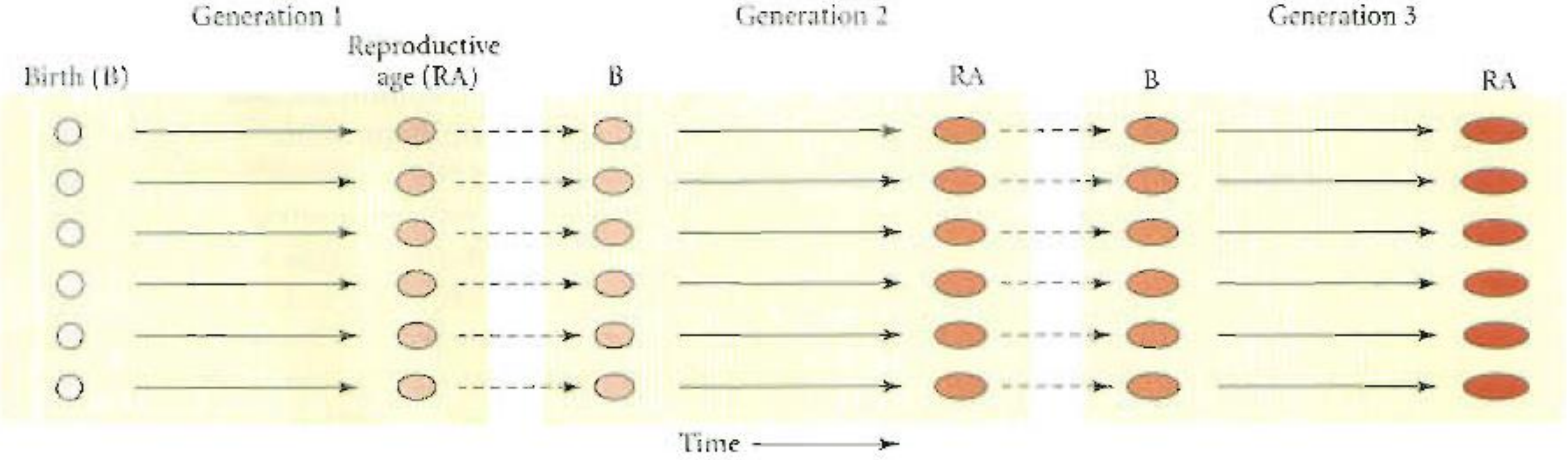
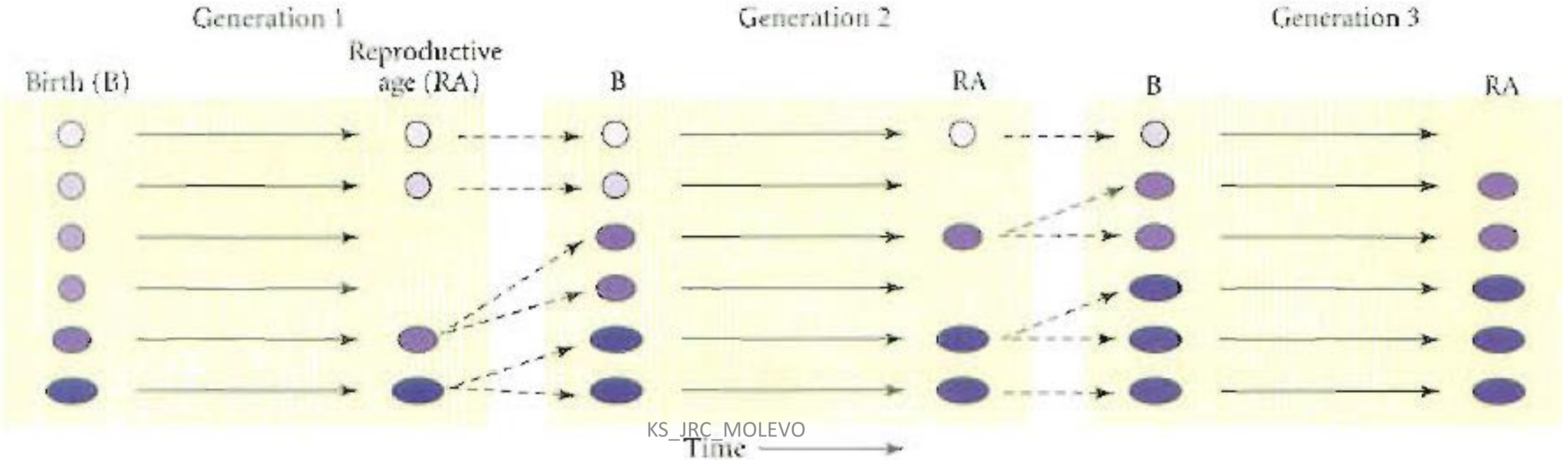
Molecular Evolution

Krishnendu Sinha
Assistant Professor
Department of Zoology
Jhargram Raj College

Transformational evolution = Lamarkism



Variational evolution = Darwinism



Darwinism	Neo-Darwinism
Natural selection (Positive & Negetive)	Natural Selection
	Sexual Selection (Runaway Hypothesis. Handicap Principle etc.)
	Genetic drift (Bottleneck, Founder Effect etc.)
Gradualism	Saltation
	Neutral Selection
Individualism	Population and gene pool

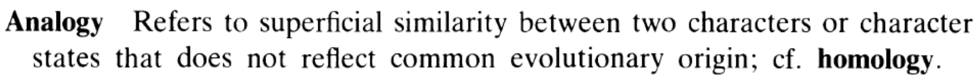
MOLECULAR PHYLOGENY

1. Basic concepts of phylogeny
2. Construction of Phylogenetic Trees
   1. Phylogenetic Inference-distance Methods, Parsimony Methods, Maximum Likelihood Method
   2. Amino Acid Sequences and Phylogeny
   3. Nucleic Acid Phylogeny
3. Nucleotide Sequence Comparisons and Homologies

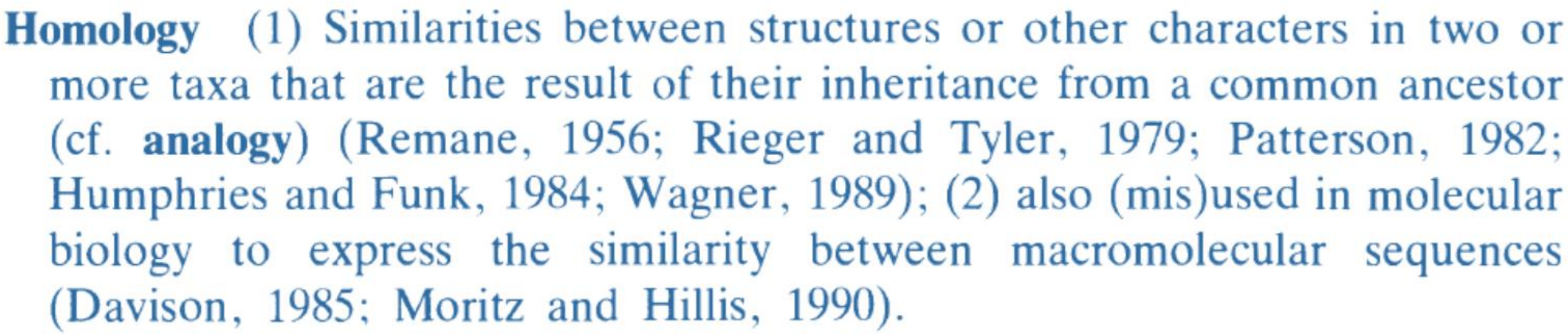
Phylogeny is the history of the evolution of a species or group,

especially in reference to lines of descent and relationships among broad groups of organisms

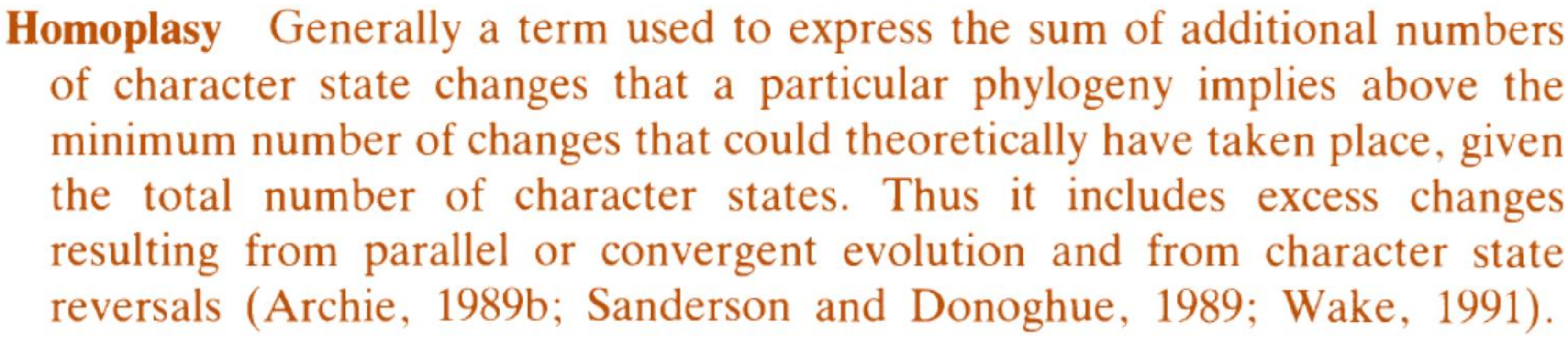
…the evolutionary history of a group of taxa



Analogy Refers to superficial similarity between two characters or character states that does not reflect common evolutionary origin; cf. **homology**.



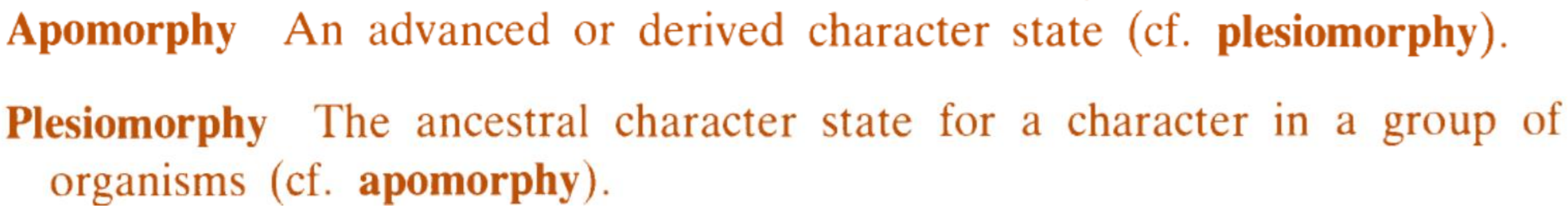
Homology (1) Similarities between structures or other characters in two or more taxa that are the result of their inheritance from a common ancestor (cf. **analogy**) (Remane, 1956; Rieger and Tyler, 1979; Patterson, 1982; Humphries and Funk, 1984; Wagner, 1989); (2) also (mis)used in molecular biology to express the similarity between macromolecular sequences (Davison, 1985; Moritz and Hillis, 1990).



Homoplasy Generally a term used to express the sum of additional numbers of character state changes that a particular phylogeny implies above the minimum number of changes that could theoretically have taken place, given the total number of character states. Thus it includes excess changes resulting from parallel or convergent evolution and from character state reversals (Archie, 1989b; Sanderson and Donoghue, 1989; Wake, 1991).

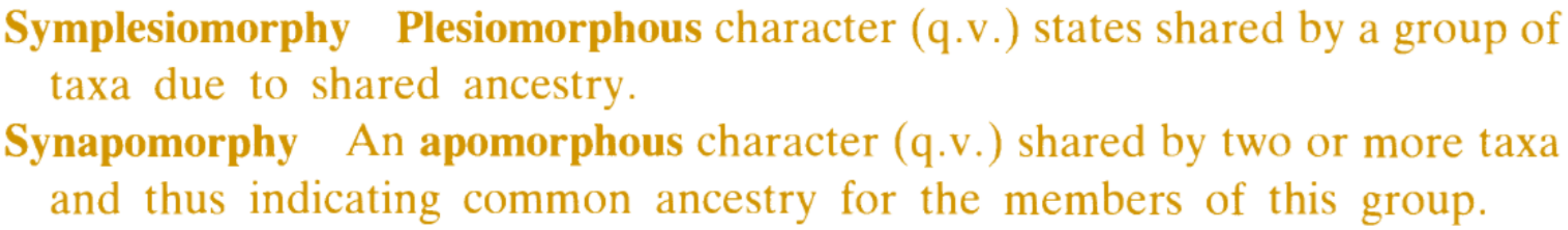
Character Any physical structure (macroscopic, microscopic or molecular) or behavioural system that can have more than one form **(character state q. v.),** the variation in which potentially provides phylogenetic information.

Character state Any of the possible. distinct conditions that a character can display. Sometimes also loosely termed character



Apomorphy An advanced or derived character state (cf. **plesiomorphy**).

Plesiomorphy The ancestral character state for a character in a group of organisms (cf. **apomorphy**).

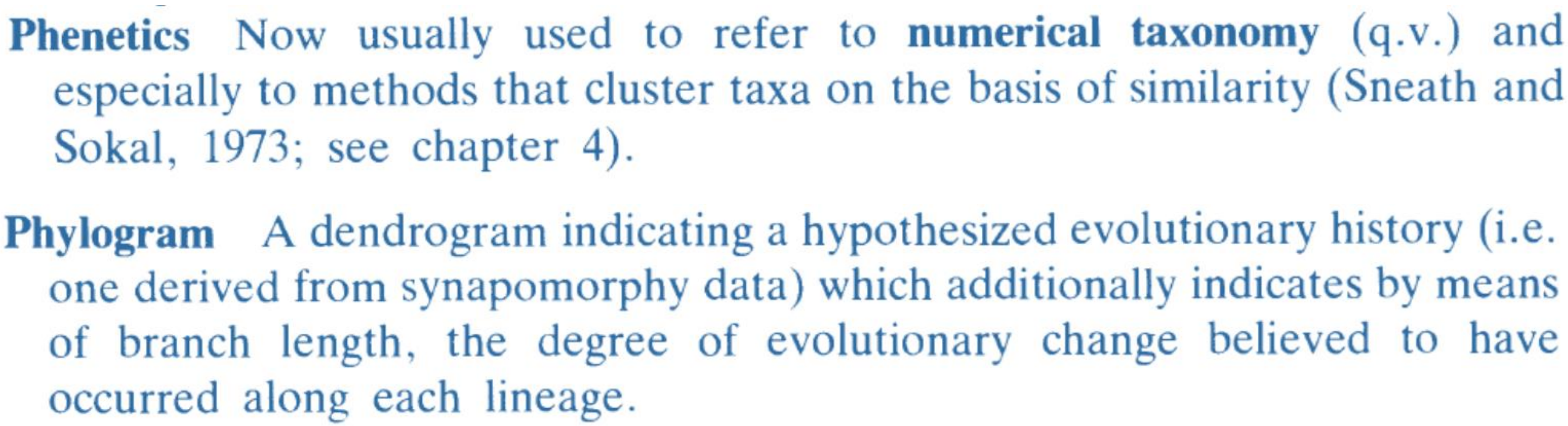


Symplesiomorphy **Plesiomorphous** character (q.v.) states shared by a group of taxa due to shared ancestry.

Synapomorphy An **apomorphous** character (q.v.) shared by two or more taxa and thus indicating common ancestry for the members of this group.

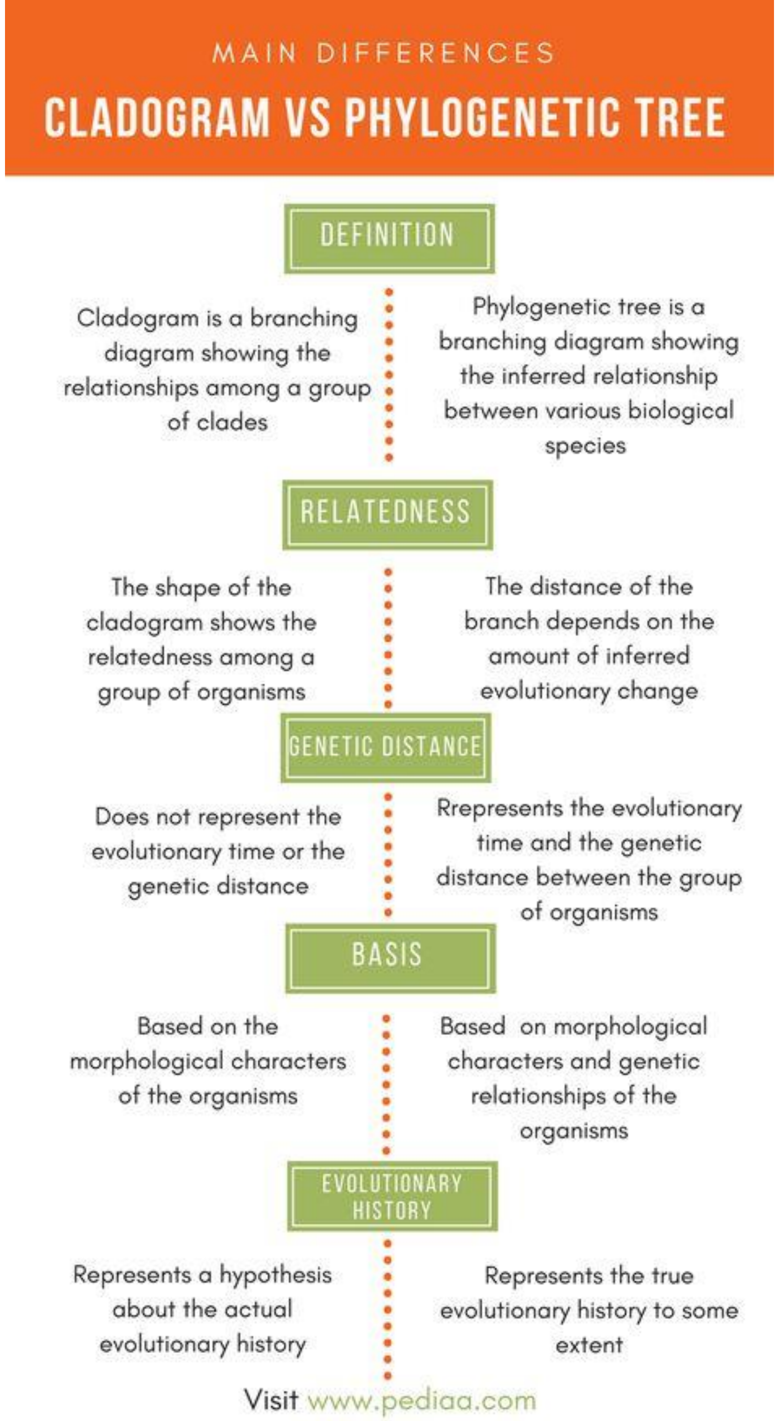
Cladistics In general the process of defining evolutionary relationships between taxa using evidence from extant taxa. Originally formulized by Hennig, now comprising a number of variously related methodologies (Fitch, 1984).

Cladogram A dendrogram (tree diagram) specifically depicting a phylogenetic hypothesis and therefore based on synapomorphies. A cladogram generally only indicates the branching pattern of the evolutionary history, cf. **phylogram (Mayr. 1965).**



Phenetics Now usually used to refer to **numerical taxonomy** (q.v.) and especially to methods that cluster taxa on the basis of similarity (Sneath and Sokal, 1973; see chapter 4).

Phylogram A dendrogram indicating a hypothesized evolutionary history (i.e. one derived from synapomorphy data) which additionally indicates by means of branch length, the degree of evolutionary change believed to have occurred along each lineage.



MAIN DIFFERENCES

CLADOGRAM VS PHYLOGENETIC TREE

DEFINITION

Cladogram is a branching diagram showing the relationships among a group of clades

Phylogenetic tree is a branching diagram showing the inferred relationship between various biological species

RELATEDNESS

The shape of the cladogram shows the relatedness among a group of organisms

The distance of the branch depends on the amount of inferred evolutionary change

GENETIC DISTANCE

Does not represent the evolutionary time or the genetic distance

Rrepresents the evolutionary time and the genetic distance between the group of organisms

BASIS

Based on the morphological characters of the organisms

Based on morphological characters and genetic relationships of the organisms

EVOLUTIONARY HISTORY

Represents a hypothesis about the actual evolutionary history

Represents the true evolutionary history to some extent

Visit www.pediaa.com

Clade Any supposedly monophyletic group of taxa in a phylogenetic hypothesis (Huxley, 1958; Dupuis, 1984).

Monophyletic Two definitions are widely employed. Used here in the same way as Hennig (1966) to mean a taxon or group of taxa all members of which have a common ancestor and which includes all the descendents of that ancestor. This corresponds to 'holophyletic' of Ashlock who also considered **paraphyletic** taxa (q.v.) to be monophyletic (Hennig, 1966; Ashlock, 1971; Platnick, 1976; Hull, 1979; see section 2.4.1).

Paraphyletic and **paraphyly** Several definitions of paraphyly have been proposed (Hennig, 1966; Nelson, 1971; Ashlock, 1971; Farris, 1974; Platnick, 1977a; Wiley, 1981). As used here it designates a group including a common ancestor and whose membership is defined by possession of a uniquely derived character state but one which may have undergone one or several reversals and therefore does not include all descendants of that common ancestor (Farris, 1974; Platnick, 1977a) (cf. **polyphyletic**).

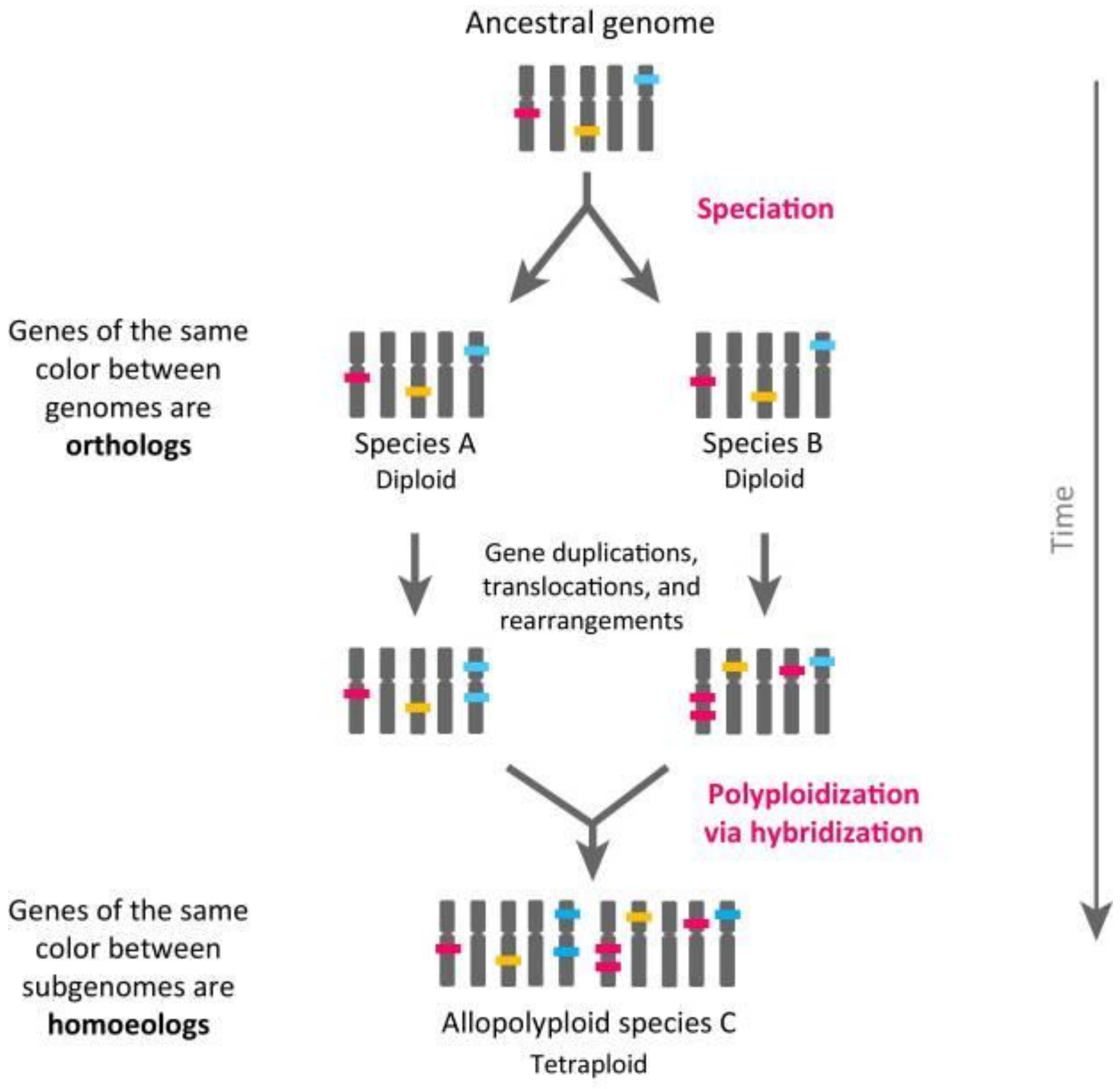
Polyphyletic and **polyphyly** Several definitions of polyphyly have been proposed (Hennig, 1966; Nelson, 1971; Ashlock, 1971; Farris, 1974). As used here it designates a group which does not include the common ancestor of all of its members (Farris, 1974; Platnick, 1977a) (cf. **paraphyletic**; see section 2.4.1).

Outgroup A taxon or group of taxa believed to have a sister group relationship with or by paraphyletic with respect to a grc study (**ingroup** q.v.) and thus by the parsimony criterion the plesiomorphous character state for the ingroup (Watrous and Wheeler, 1981; Maddison *et al*., 1984; chapter 2).

Ingroup The apparently monophyletic group of taxa the relationships of which are under investigation (cf. **outgroup**).

Homologs

	Pairs of genes found in the same species	Pairs of genes found in different species
Genes that originated by a speciation event	**HOMOEOLOGS**	**ORTHOLOGS** Usually same in function
Genes that originated by a duplication event	Whole genome duplication: **OHNOLOGS** Small scale duplication: **PARALOGS**	**PARALOGS** Usually different in function


Ancestral genome
Speciation
Genes of the same color between genomes are **orthologs**
Species A
Diploid
Species B
Diploid
Gene duplications, translocations, and rearrangements
Polyploidization via hybridization
Genes of the same color between subgenomes are **homoeologs**
Allopolyploid species C
Tetraploid
Time

Trends in Plant Science

Evolutionary History of an Allopolyploid. An ancestral genome undergoes a speciation event, resulting in two diploid species. The genes, which descended from a common gene in the ancestor, are orthologs. Evolution occurs after speciation, including structural rearrangements, gene duplications, and gene movement. On polyploidization, genes that were once orthologs are now homoeologs. Homoeologous relationships can be one-to-one, one-to-many, or many-to-many depending on the number of duplications since speciation of the progenitors.

Five Stages of Phylogenetic Analysis

Molecular phylogenetic analyses can be divided into five stages:

(1)selection of sequences for analysis

(2)multiple sequence alignment of homologous protein or nucleic acid sequences

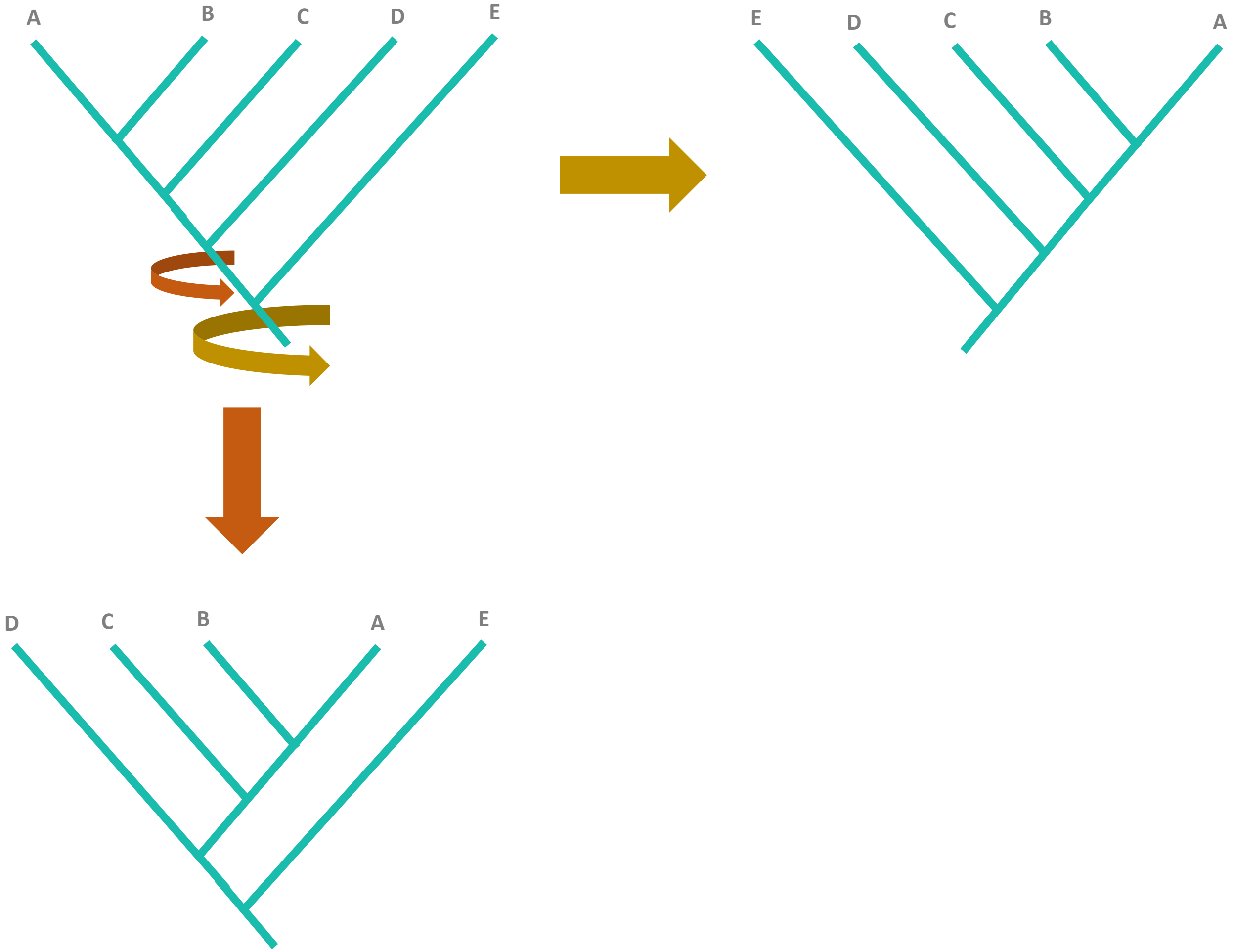
(3)specification of a statistical model of nucleotide or amino acid evolution

(4)tree building

(5)tree evaluation

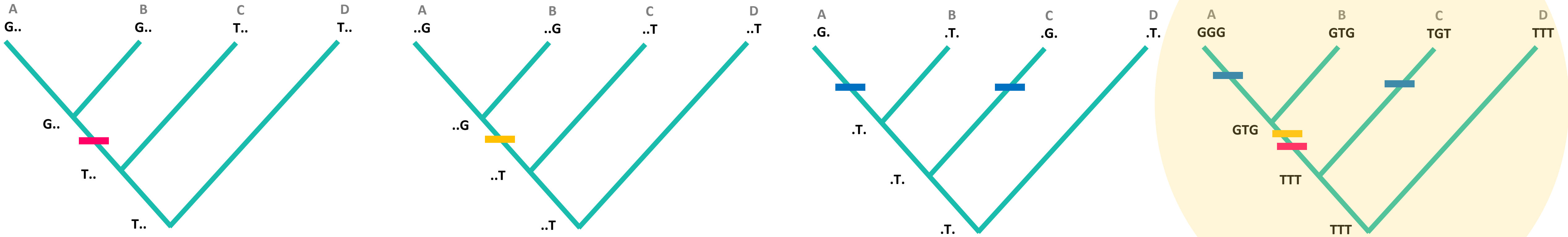
Tree Topology

Tree representation



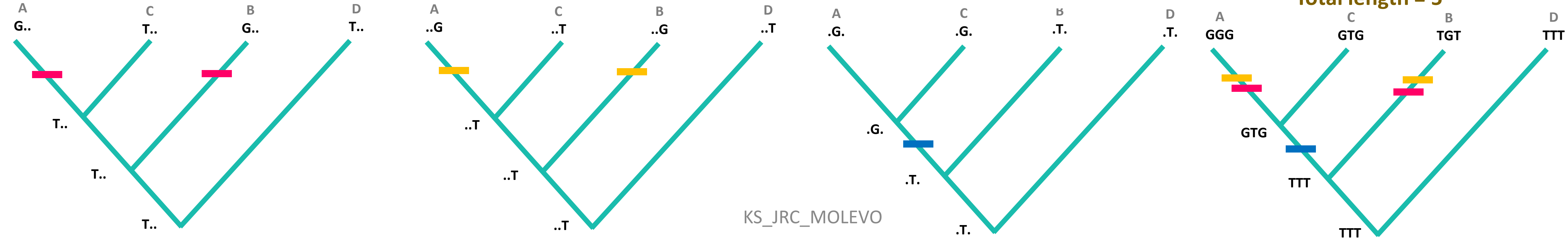
Maximum Parsimony

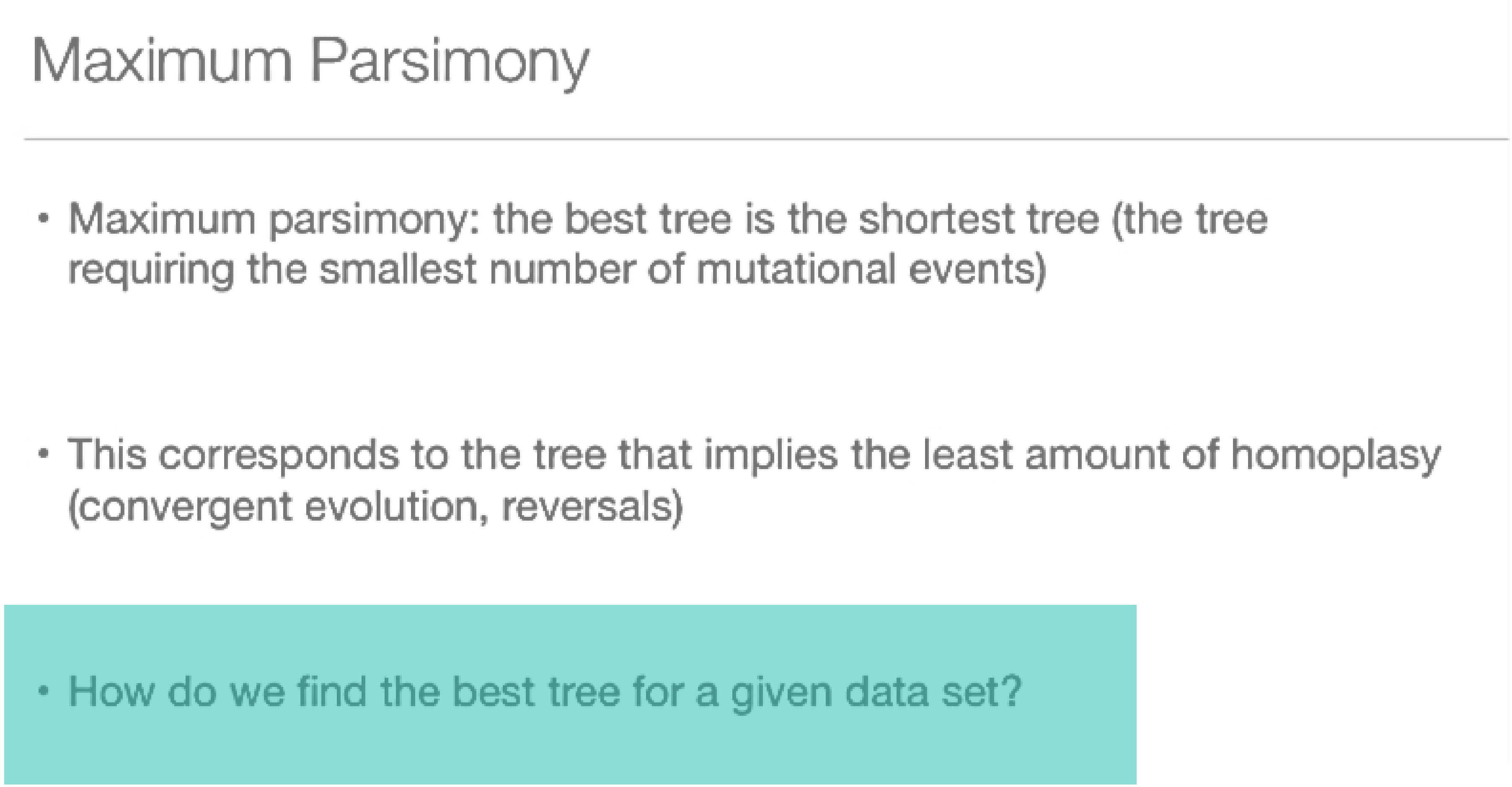
Phylogenetic reconstruction



Total length = 4

Total length = 5



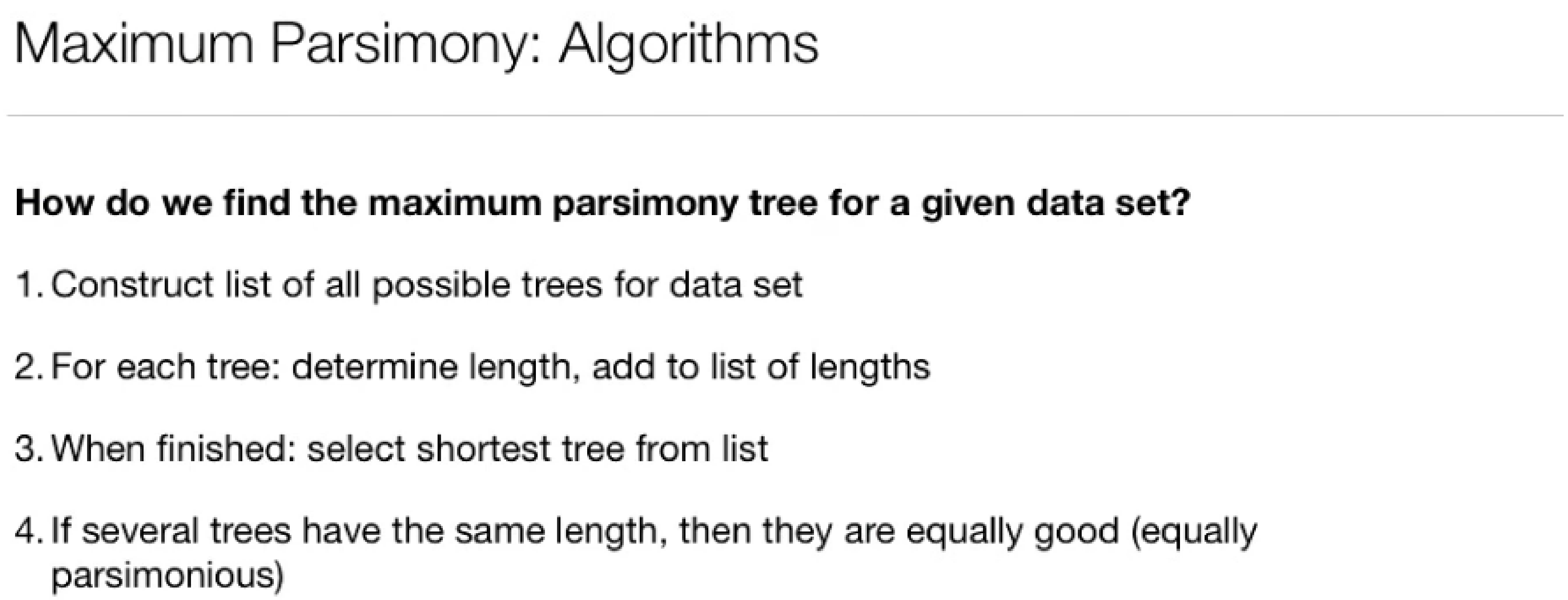


Maximum Parsimony

- Maximum parsimony: the best tree is the shortest tree (the tree requiring the smallest number of mutational events)

- This corresponds to the tree that implies the least amount of homoplasy (convergent evolution, reversals)

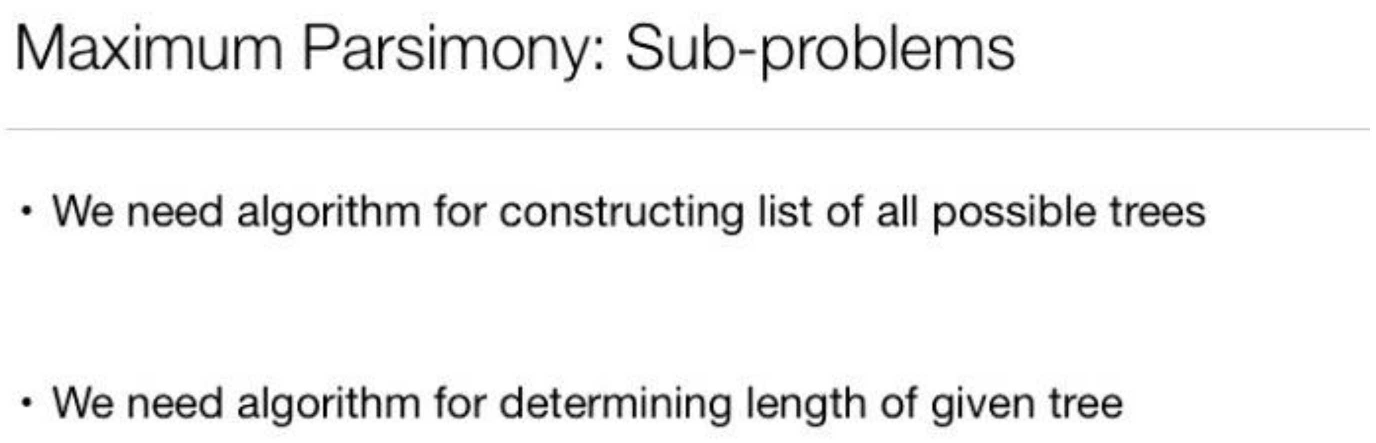
- How do we find the best tree for a given data set?



Maximum Parsimony: Algorithms

How do we find the maximum parsimony tree for a given data set?

1. Construct list of all possible trees for data set
2. For each tree: determine length, add to list of lengths
3. When finished: select shortest tree from list
4. If several trees have the same length, then they are equally good (equally parsimonious)

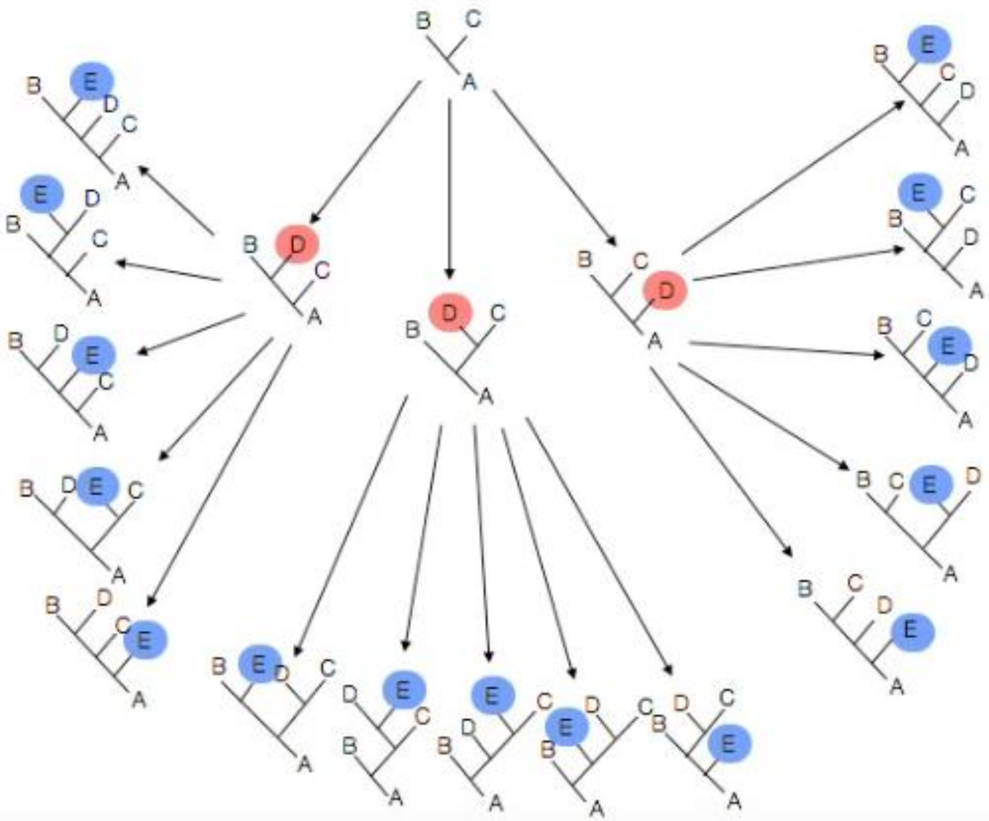


Maximum Parsimony: Sub-problems

- We need algorithm for constructing list of all possible trees

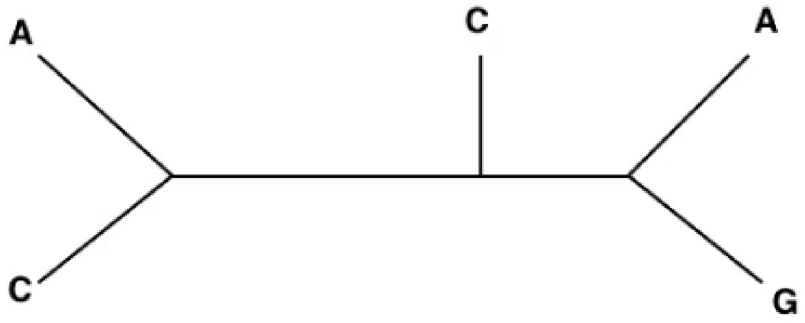
- We need algorithm for determining length of given tree

Constructing list of all possible unrooted trees

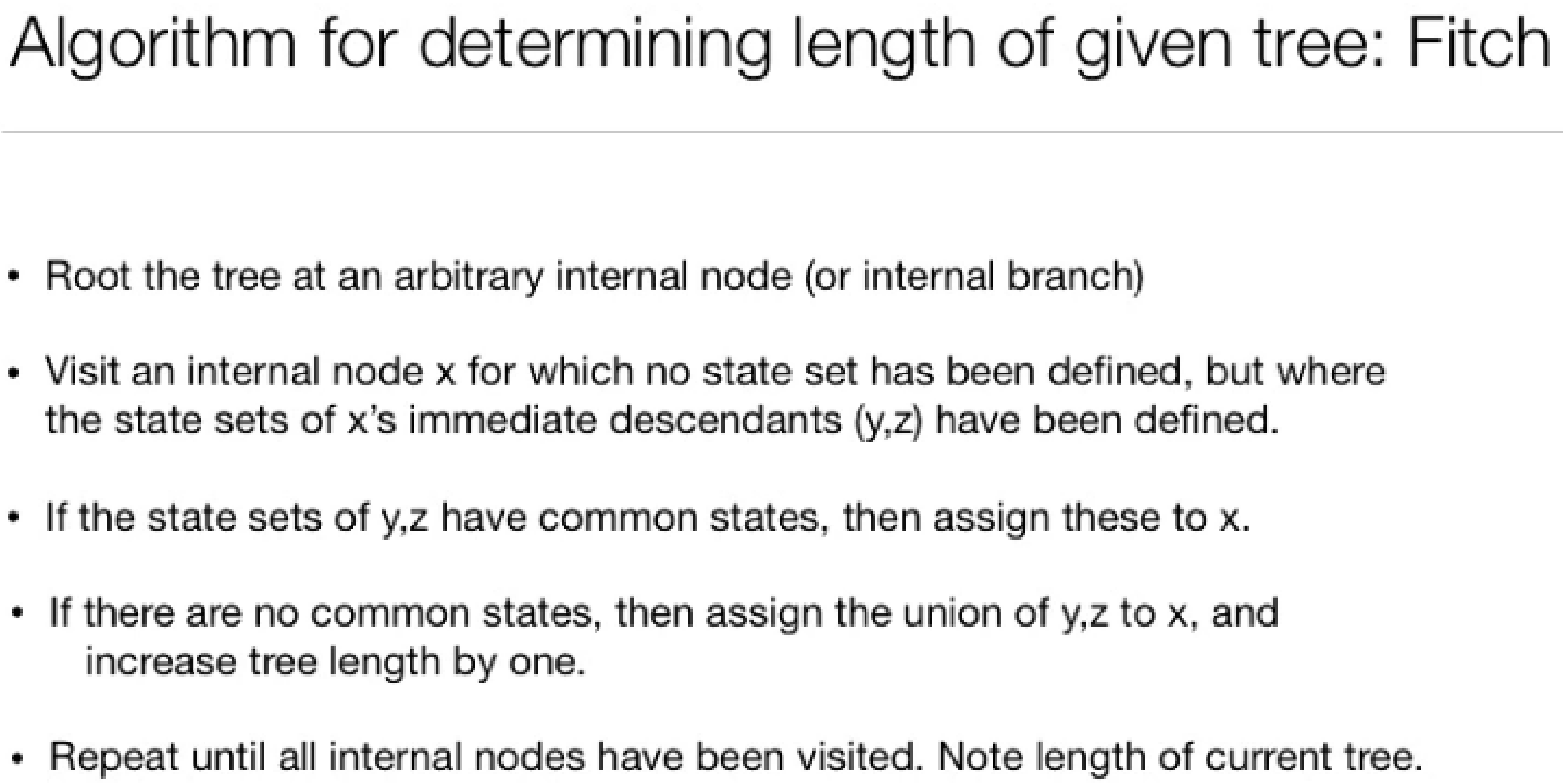


1. Construct unrooted tree from first three taxa. There is only one way of doing this
2. Starting from (1), construct the three possible derived trees by adding taxon 4 to each internal branch
3. From each of the trees constructed in step (2), construct the five possible derived trees by adding taxon 5 to each internal branch.
4. Continue until all taxa have been added in all possible locations

Algorithm for determining length of given tree: Fitch

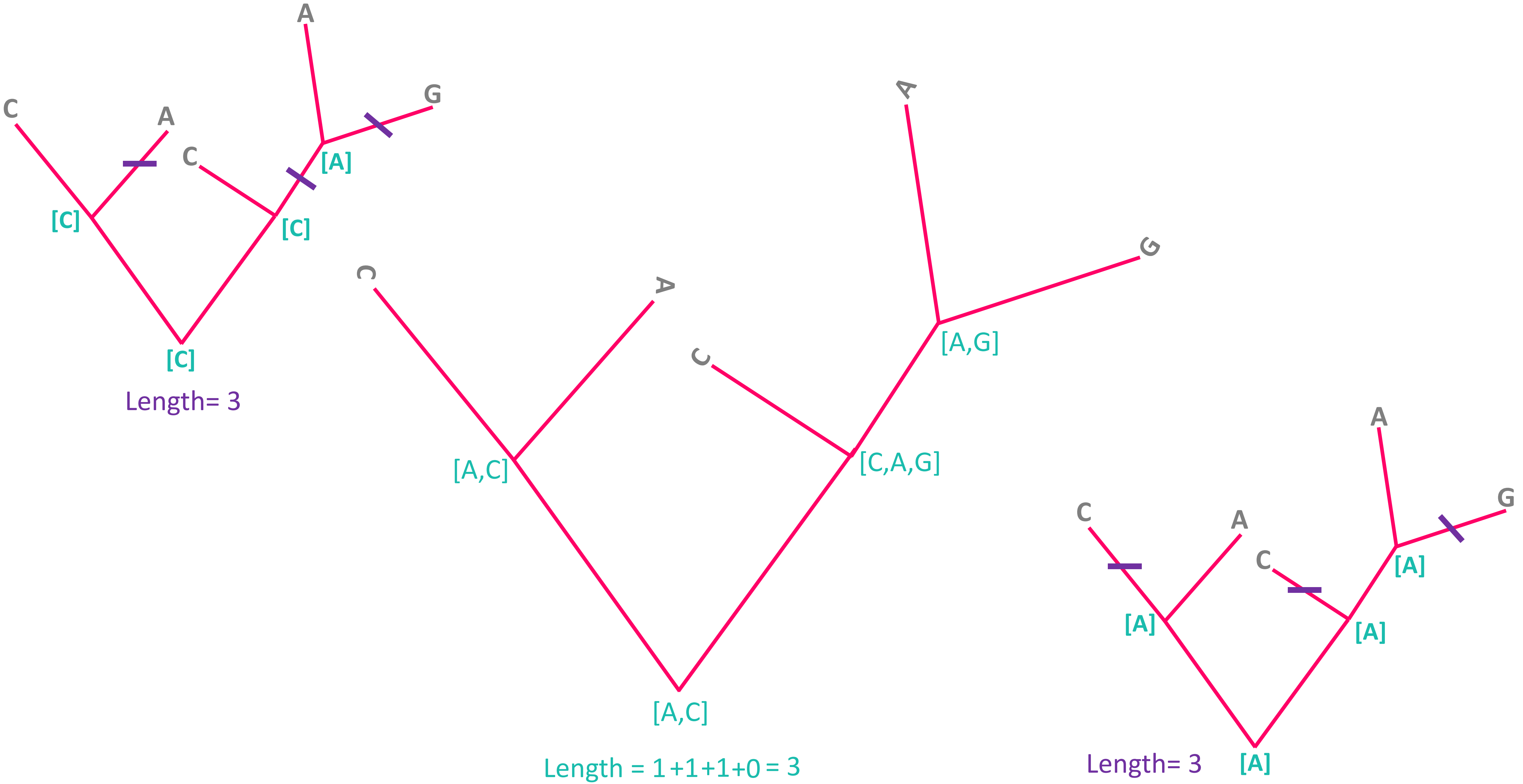


What is the length of this tree? (How many mutational steps are required?)



Algorithm for determining length of given tree: Fitch

- Root the tree at an arbitrary internal node (or internal branch)
- Visit an internal node x for which no state set has been defined, but where the state sets of x's immediate descendants (y,z) have been defined.
- If the state sets of y,z have common states, then assign these to x.
- If there are no common states, then assign the union of y,z to x, and increase tree length by one.
- Repeat until all internal nodes have been visited. Note length of current tree.



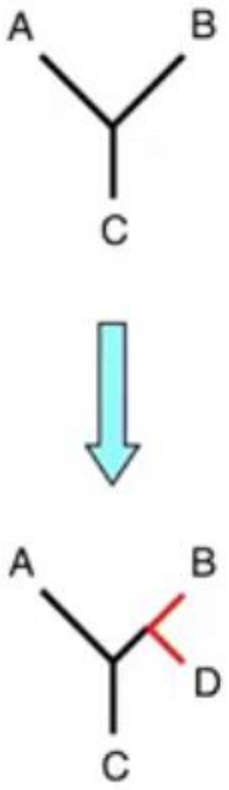
Length= 3

Length = 1+1+1+0 = 3

Length= 3

Tree Space Searching

How many branches are there on an unrooted tree with x tips?



- There is only one way of constructing the first tree. This tree has 3 tips and 3 branches

- Each time an extra taxon is added, two branches are created.

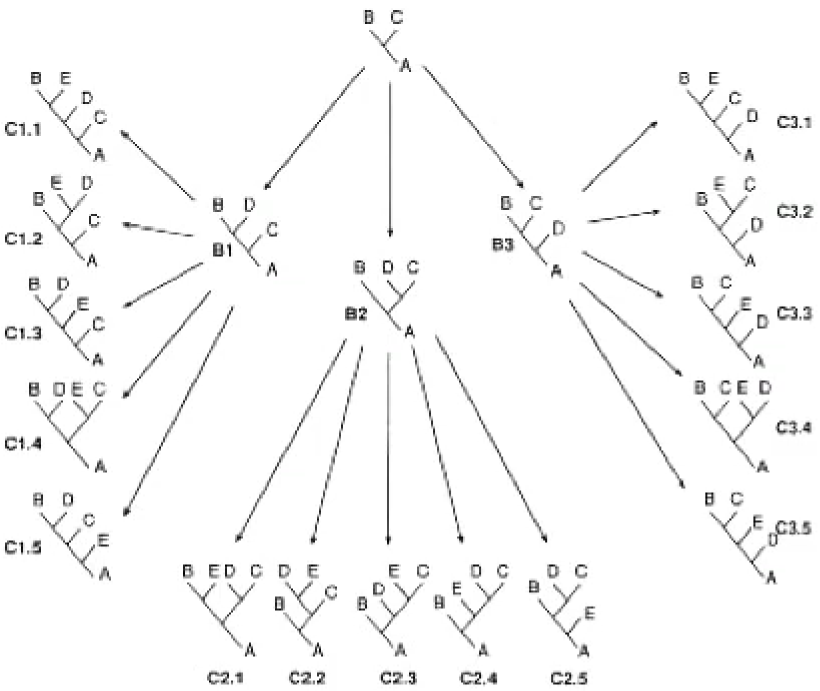
- A tree with x tips will therefore have the following number of branches:

$$n_{\text{branches}} = 3+(x-3)*2$$

$$= 3+2x-6$$

$$= 2x-3$$

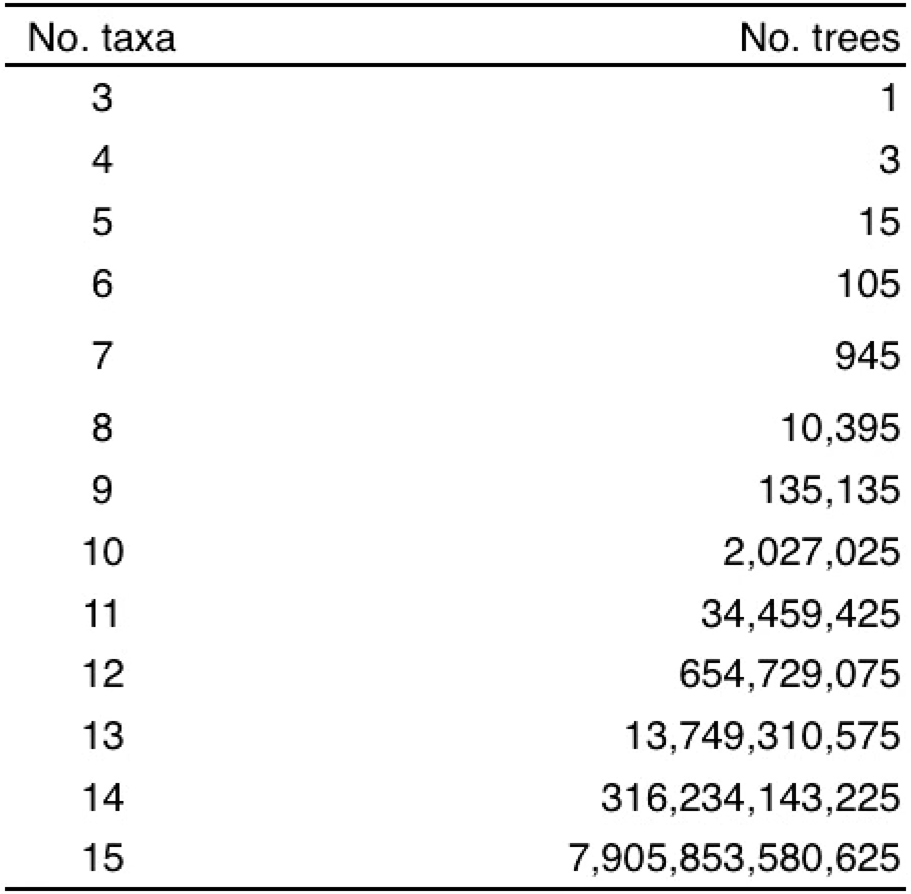
How many unrooted trees are there?



- A tree with x tips has 2x-3 branches

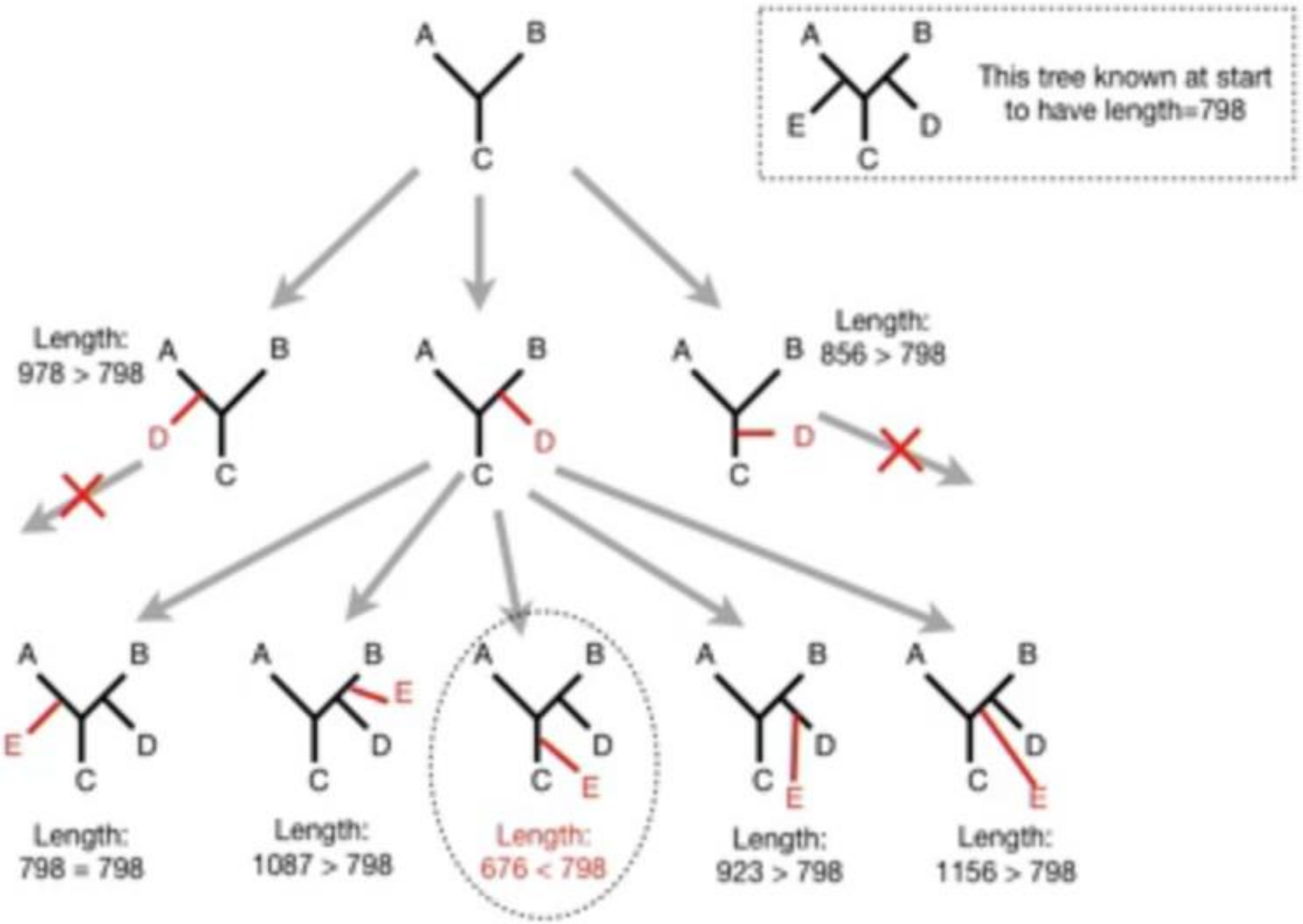
- For each tree with x tips, we can therefore construct 2x-3 derived trees (which each have x+1 tips).

Exhaustive search impossible for large data sets



No. taxa	No. trees
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
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

Branch and bound: shortcut to perfection



Heuristic search

1. Construct initial tree (e.g., sequential addition); determine length

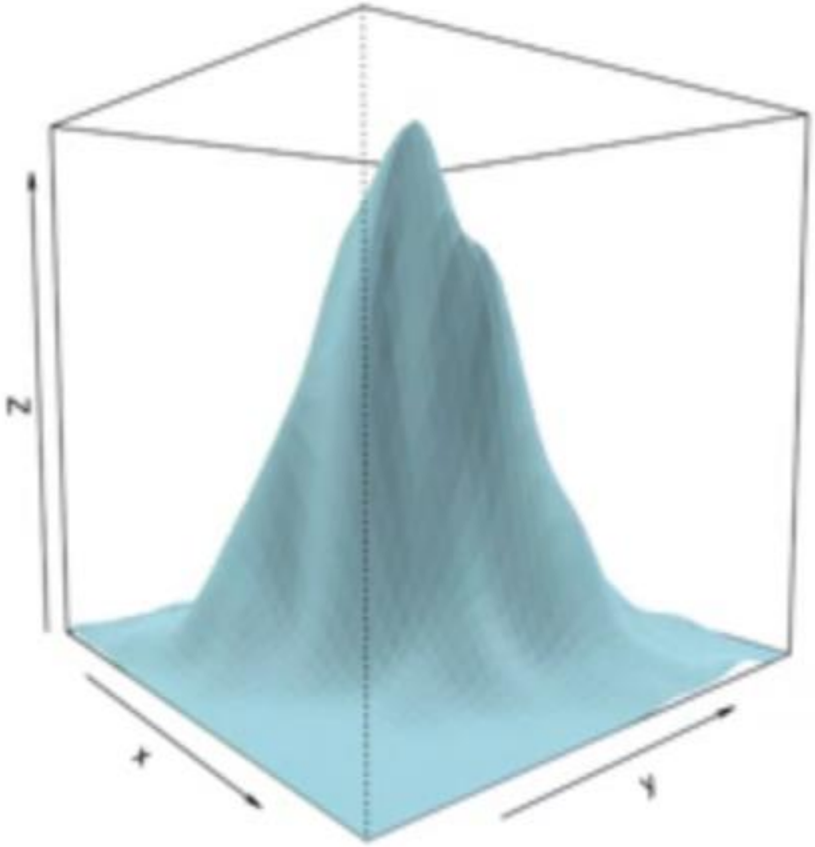
2. Construct set of "neighboring trees" by making small rearrangements of initial tree; determine lengths

3. If any of the neighboring trees are better than the initial tree, then select it/ them and use as starting point for new round of rearrangements. (Possibly several neighbors are equally good)

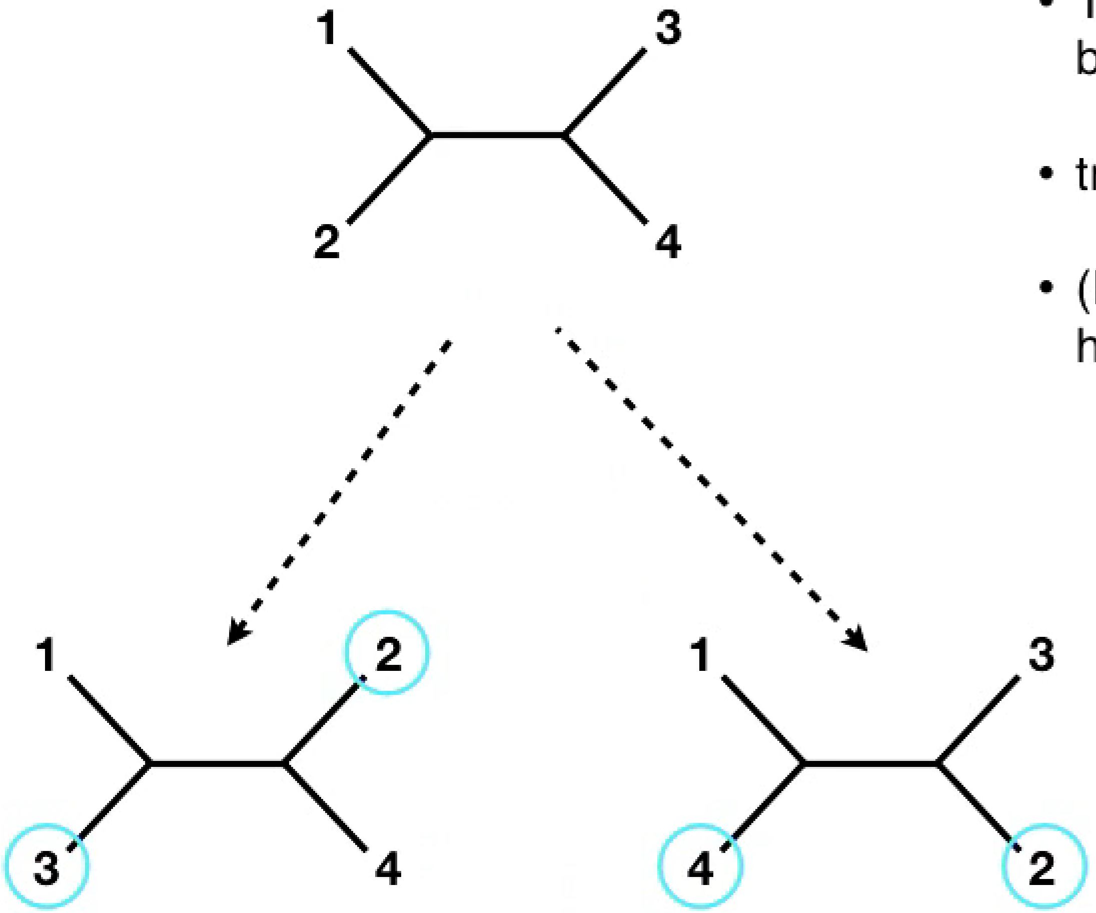
4. Repeat steps 2+3 until you have found a tree that is better than all of its neighbors.

5. This tree is a "local optimum" (not necessarily a global optimum!)

Heuristic search: hill-climbing



Types of rearrangement I: nearest neighbor interchange (NNI)

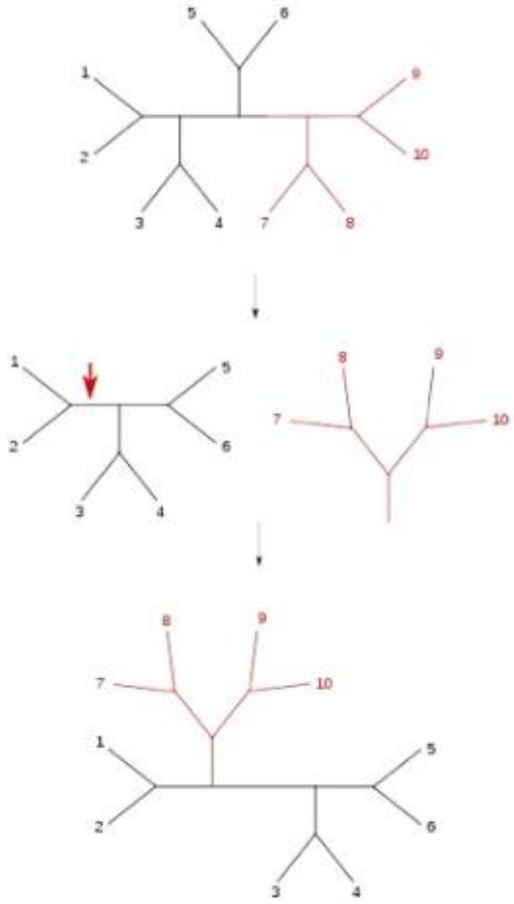


- Two neighboring trees per internal branch:

- tree with n tips has 2(n-3) neighbors

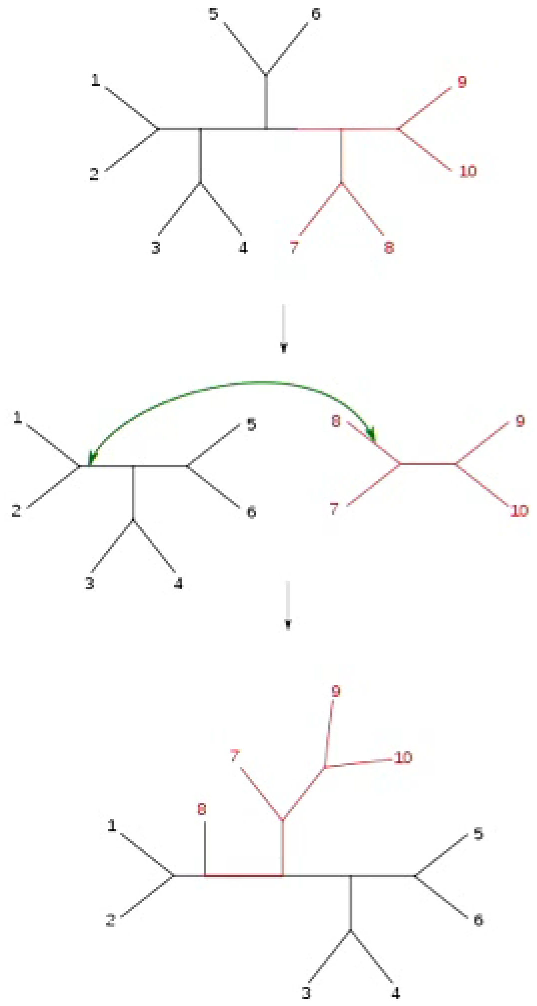
- (For example, a tree with 20 tips has 34 neighbors)

Types of rearrangement II:
subtree pruning and regrafting (SPR)



- Detach subtree
- Re-attach subtree on all branches in other half of tree
- Use cut-point (root of detached subtree) for re-attachment
- NNI is a subset of SPR

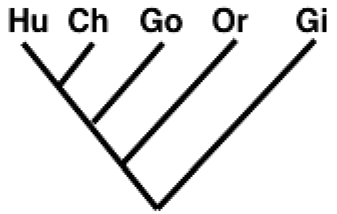
Types of rearrangement III:
tree bisection and reconnection (TBR)



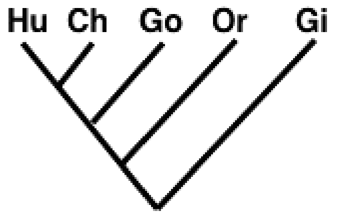
- Divide tree into two parts.
- Reconnect subtrees using every possible pair of branches
- NNI and SPR are subsets of TBR

Consensus Tree

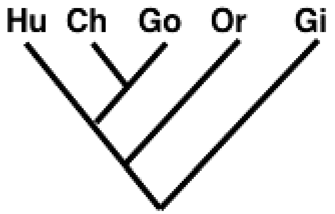
Strict Consensus Tree



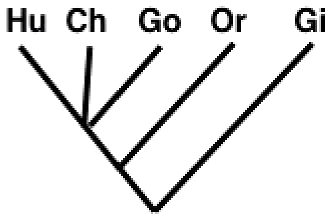
Tree 1



Tree 2

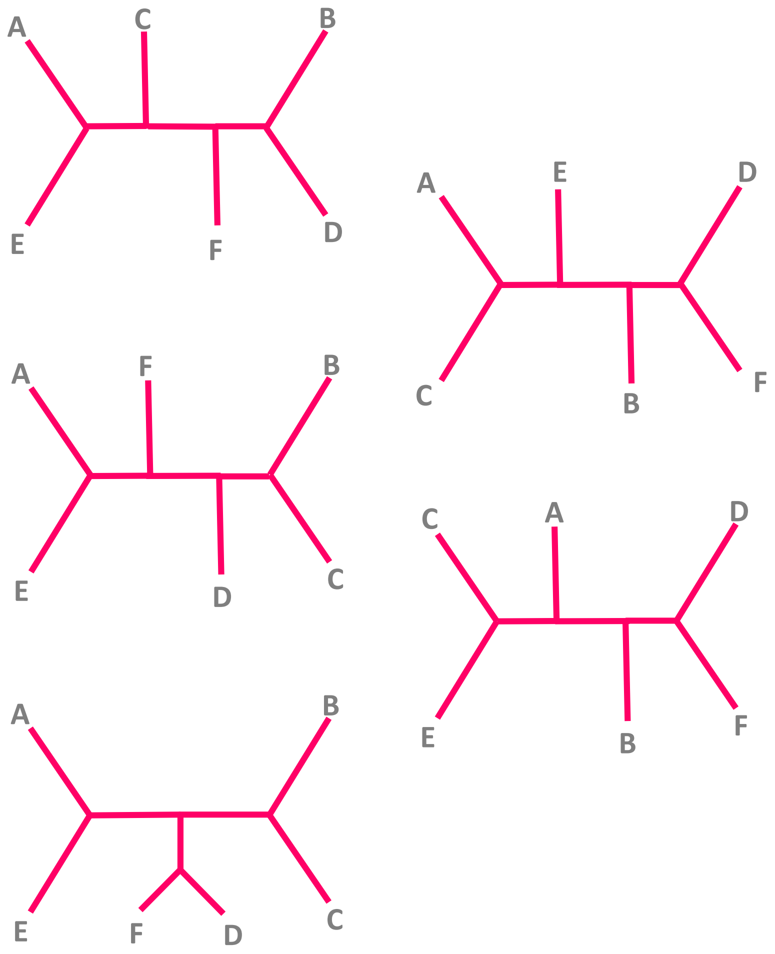


Tree 3



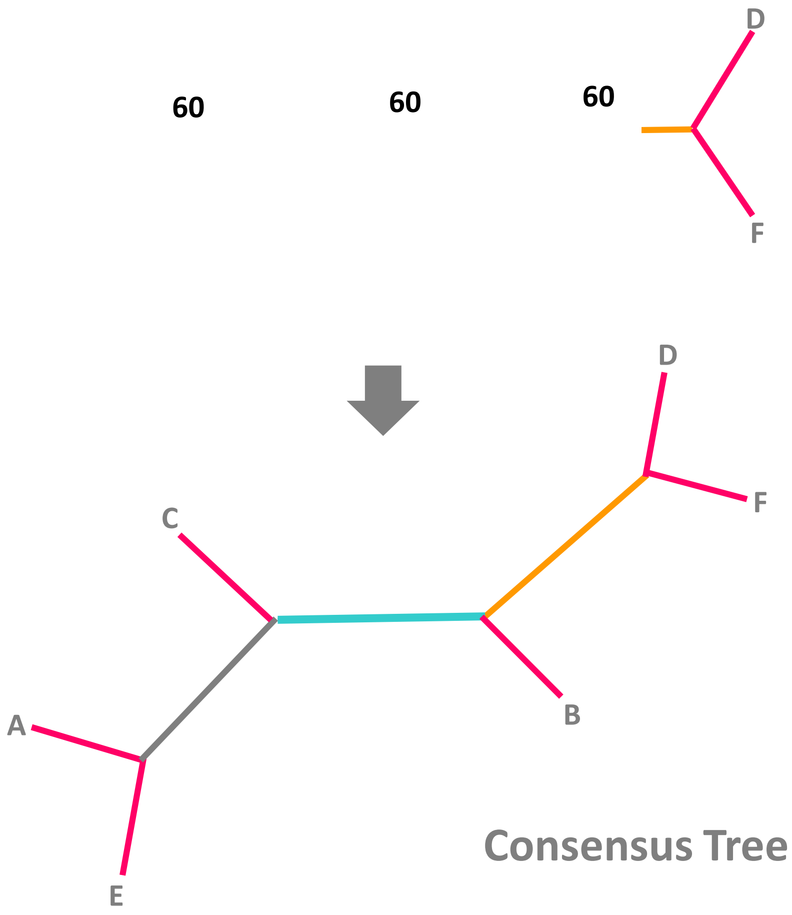
Strict consensus

Majority Rule Consensus Tree



A	B	C	D	E	F	COUNT	FREQUENCY
*	-	-	-	*	-	III	60
*	-	*	-	*	-	III	60
-	*	-	*	-	-	I	20
*	-	*	-	-	-	I	20
-	-	-	*	-	*	III	60
*	-	-	-	*	*	I	20
-	*	*	-	-	-	II	40
-	-	*	-	*	-	I	20

Majority Rule Consensus Tree Construction



Consensus Tree

A	B	C	D	E	F	COUNT	FREQUENCY
*	-	-	-	*	-	III	60
*	-	*	-	*	-	III	60
-	-	-	*	-	*	III	60
-	*	*	-	-	-	II	40
-	*	-	*	-	-	I	20
*	-	*	-	-	-	I	20
*	-	-	-	*	*	I	20
-	-	*	-	*	-	I	20

Distance Matrix Based Methods

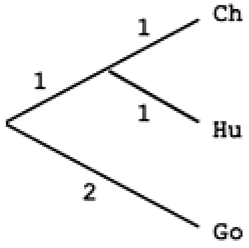
Distance Matrix Methods

```
                ↓↓↓↓
Gorilla   : ACGTCGTA
Human     : ACGTTCCT
Chimpanzee: ACGTTTCG
                 ↑ ↑
```

1. Construct multiple alignment of sequences

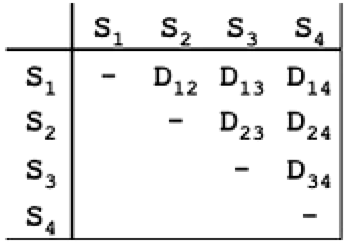
	Go	Hu	Ch
Go	-	4	4
Hu		-	2
Ch			-

2. Construct table listing all pairwise differences (distance matrix)



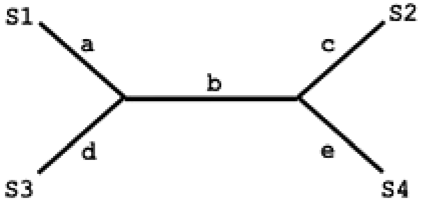
3. Construct tree from pairwise distances

Finding optimal branch lengths



	S_1	S_2	S_3	S_4
S_1	–	D_{12}	D_{13}	D_{14}
S_2		–	D_{23}	D_{24}
S_3			–	D_{34}
S_4				–

Observed distance



Distance along tree
(patristic distance)

D_{12}~d_{12}= a+b+c

D_{13}~d_{13}= a+d

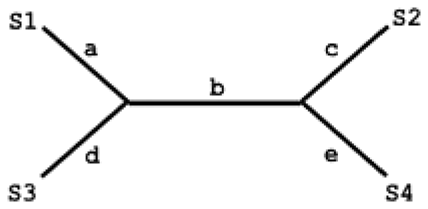
Goal: D_{14}~d_{14}= a+b+e

D_{23}~d_{23}= d+b+c

D_{24}~d_{24}= c+e

D_{34}~d_{34}= d+b+e

Optimal Branch Lengths for a Given Tree: Least Squares



Distance along tree

Goal:

$$\begin{aligned} D_{12} &\approx d_{12} = a + b + c \\ D_{13} &\approx d_{13} = a + d \\ D_{14} &\approx d_{14} = a + b + e \\ D_{23} &\approx d_{23} = d + b + c \\ D_{24} &\approx d_{24} = c + e \\ D_{34} &\approx d_{34} = d + b + e \end{aligned}$$

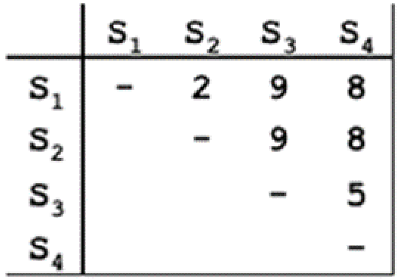
- Fit between given tree and observed distances can be expressed as “sum of squared differences”:
- $$Q = \sum_{j>i} (D_{ij} - d_{ij})^2$$
- Find branch lengths that minimize Q - this is the optimal set of branch lengths for this tree.

- Longer distances associated with larger errors
- Squared deviation may be weighted so longer branches contribute less to Q:

$$Q = \sum_{j>i} \frac{(D_{ij} - d_{ij})^2}{D_{ij}^{\,n}}$$

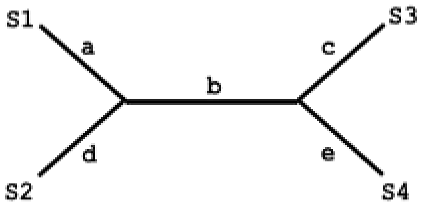
- Power (n) is typically 1 or 2

Optimal Branch Lengths for a Given Tree: Least Squares Example



	S_1	S_2	S_3	S_4
S_1	–	2	9	8
S_2		–	9	8
S_3			–	5
S_4				–

Observed distance



Distance along tree

$d_{12} = a + d$
$d_{13} = a + b + c$
$d_{14} = a + b + e$
$d_{23} = d + b + c$
$d_{24} = d + b + e$
$d_{34} = c + e$

Goal: find branch lengths that minimize Q

$$Q = \sum_{i<j} (D_{ij} - d_{ij})^2$$
$$= (D_{12} - d_{12})^2 + (D_{13} - d_{13})^2 + (D_{14} - d_{14})^2 + (D_{23} - d_{23})^2 + (D_{24} - d_{24})^2 + (D_{34} - d_{34})^2$$

Goal: find branch lengths that minimize Q

	S_1	S_2	S_3	S_4
S_1	–	2	9	8
S_2		–	9	8
S_3			–	5
S_4				–

Observed distance

$$Q = \sum_{i<j}(D_{ij} - d_{ij})^2$$

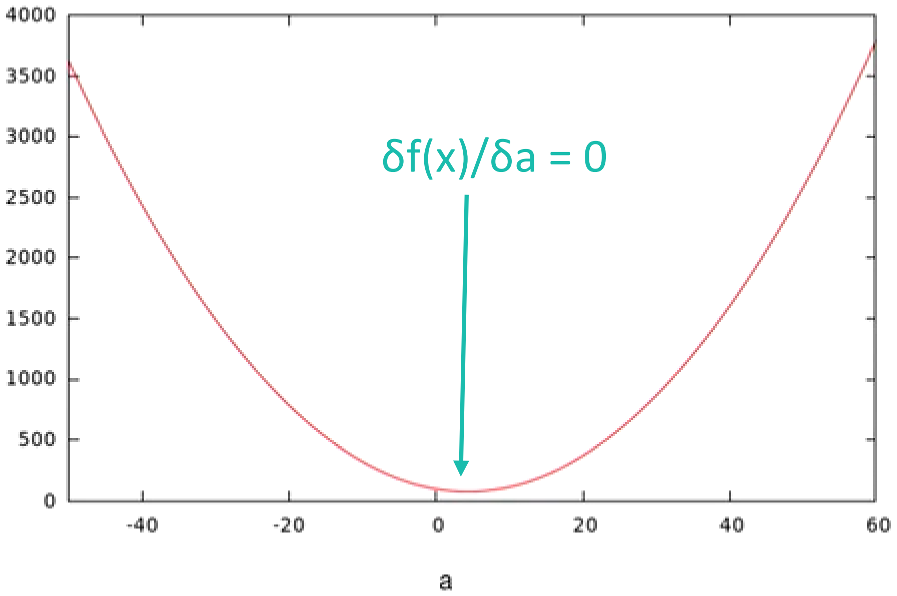
$$= (D_{12} - d_{12})^2 + (D_{13} - d_{13})^2 + (D_{14} - d_{14})^2 + (D_{23} - d_{23})^2 + (D_{24} - d_{24})^2 + (D_{34} - d_{34})^2$$

$$= (D_{12} - a - d)^2 + (D_{13} - a - b - c)^2 + (D_{14} - a - b - e)^2 + (D_{23} - d - b - c)^2 + (D_{24} - d - b - e)^2 + (D_{34} - c - e)^2$$

$$= (2 - a - d)^2 + (9 - a - b - c)^2 + (8 - a - b - e)^2 + (9 - d - b - c)^2 + (8 - d - b - e)^2 + (5 - c - e)^2$$

$$= 319 - 38a - 68b - 42e - 46c - 38d + 2ce + 3a^2 + 2ad + 3d^2 + 4ab + 2ac + 4b^2 + 4bc + 3c^2 + 2ae + 4be + 3e^2 + 4db + 2dc + 2de$$

$$= 3a^2 + (4b + 2c + 2d - 38 + 2e)a + 319 - 68b - 42e - 46c - 38d + 2ce + 3d^2 + 4b^2 + 4bc + 3c^2 + 4be + 3e^2 + 4db + 2dc + 2de$$



$$\frac{\partial Q}{\partial a} = 6a + 4b + 2c + 2d - 38 + 2e = 0$$

$$\frac{\partial Q}{\partial b} = -68 + 4a + 8b + 4c + 4e + 4d = 0$$

$$\frac{\partial Q}{\partial c} = -46 + 2a + 4b + 6c + 2d + 2e = 0$$

$$\frac{\partial Q}{\partial d} = -38 + 2a + 6d + 4b + 2c + 2e = 0$$

$$\frac{\partial Q}{\partial e} = -42 + 2a + 4b + 6e + 2d + 2c = 0$$

- System of 5 linear equations with 5 unknowns
- Can be solved for a, b, c, d, e

$f(x) = ax^2 + bx + c$ (where a, b, and c are real numbers and a ≠ 0)

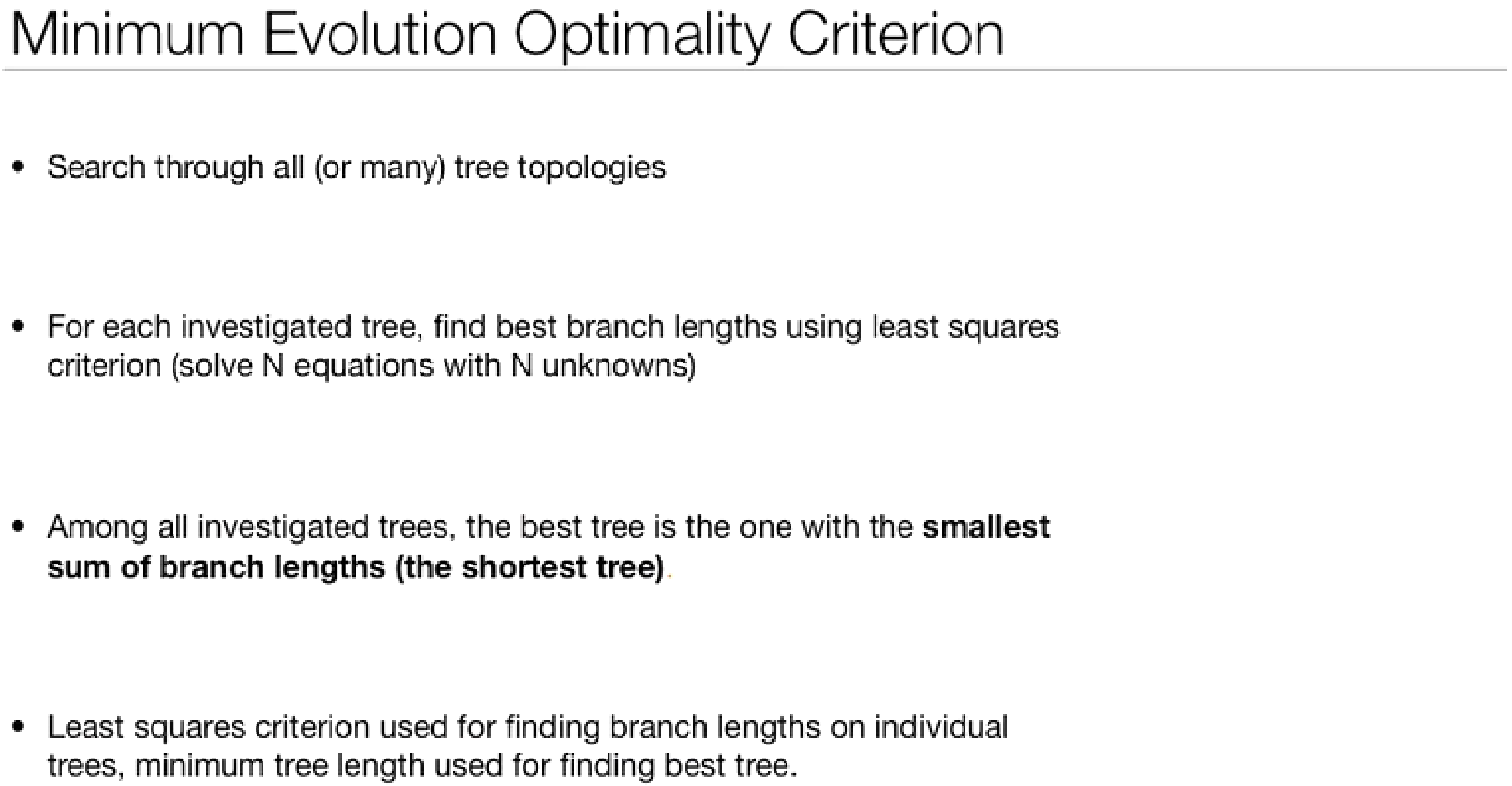
Least Squares Optimality Criterion

- Search through all (or many) tree topologies

- For each investigated tree, find best branch lengths using least squares criterion (solve N equations with N unknowns)

- Among all investigated trees, the best tree is the one with the **smallest sum of squared errors**.

- Least squares criterion used both for finding branch lengths on individual trees, and for finding best tree.



Minimum Evolution Optimality Criterion

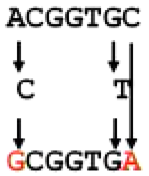
- Search through all (or many) tree topologies

- For each investigated tree, find best branch lengths using least squares criterion (solve N equations with N unknowns)

- Among all investigated trees, the best tree is the one with the **smallest sum of branch lengths (the shortest tree)**.

- Least squares criterion used for finding branch lengths on individual trees, minimum tree length used for finding best tree.

Superimposed Substitutions

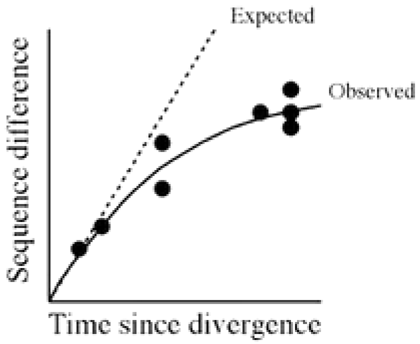


ACGGTGC

C T

GCGGTGA

- Actual number of evolutionary events: 5
- Observed number of differences: 2



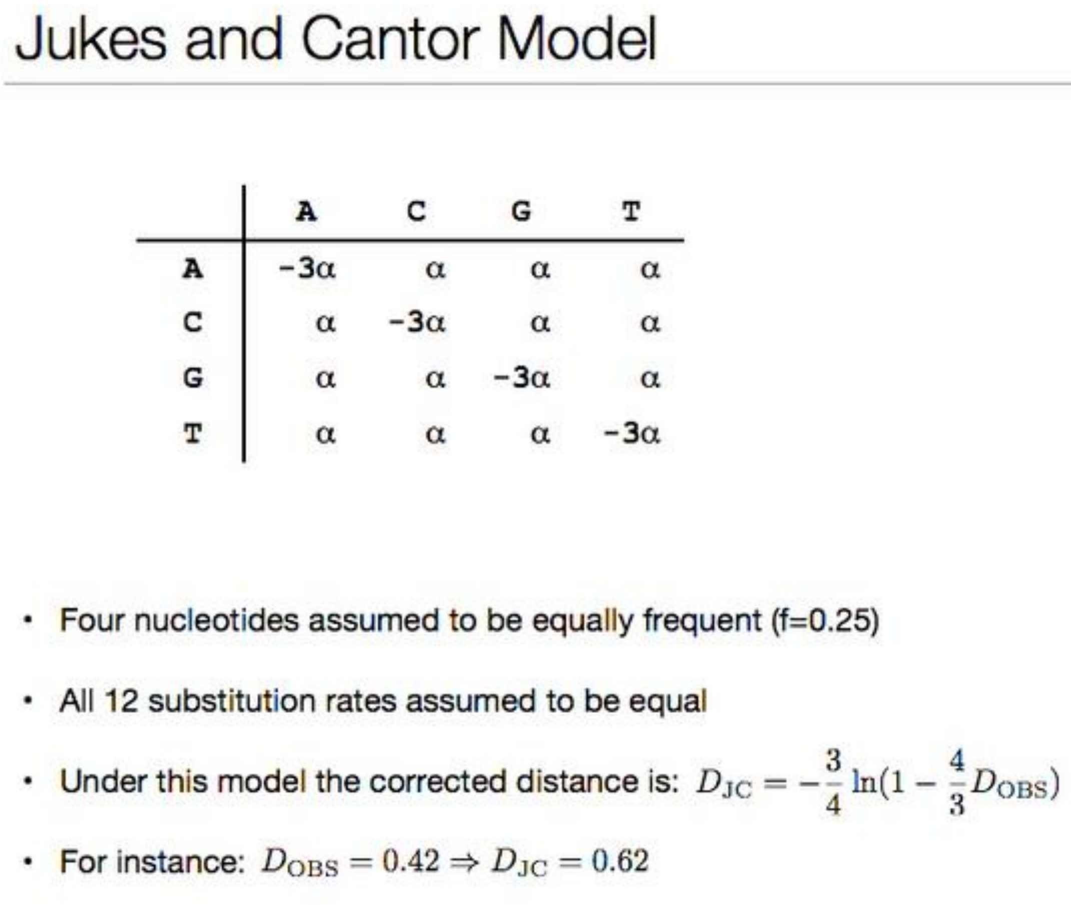
- Distance is (almost) always underestimated

Model Based Correction for Superimposed Substitutions

Goal: try to infer the real number of evolutionary events (the real distance based on

a) Observed data (sequence comparison)

b) A model on how evolution has occurred

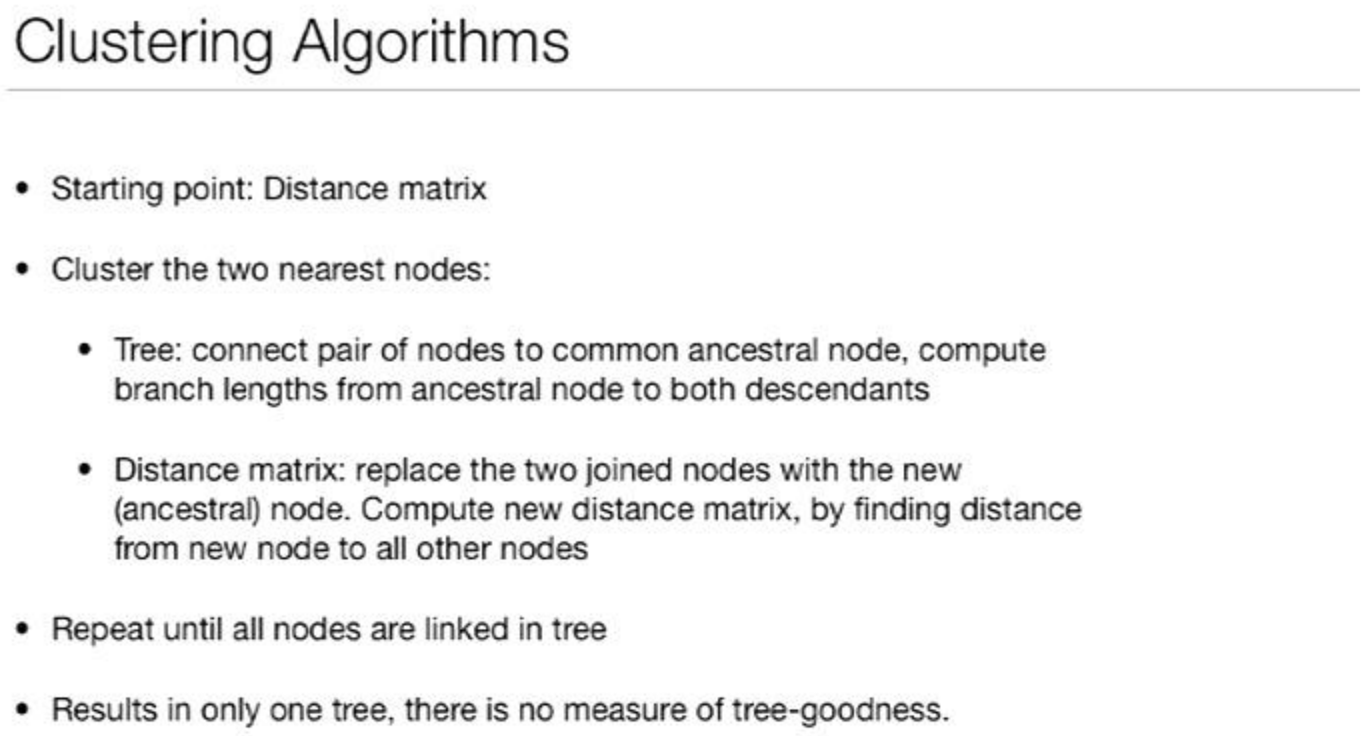


Jukes and Cantor Model

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

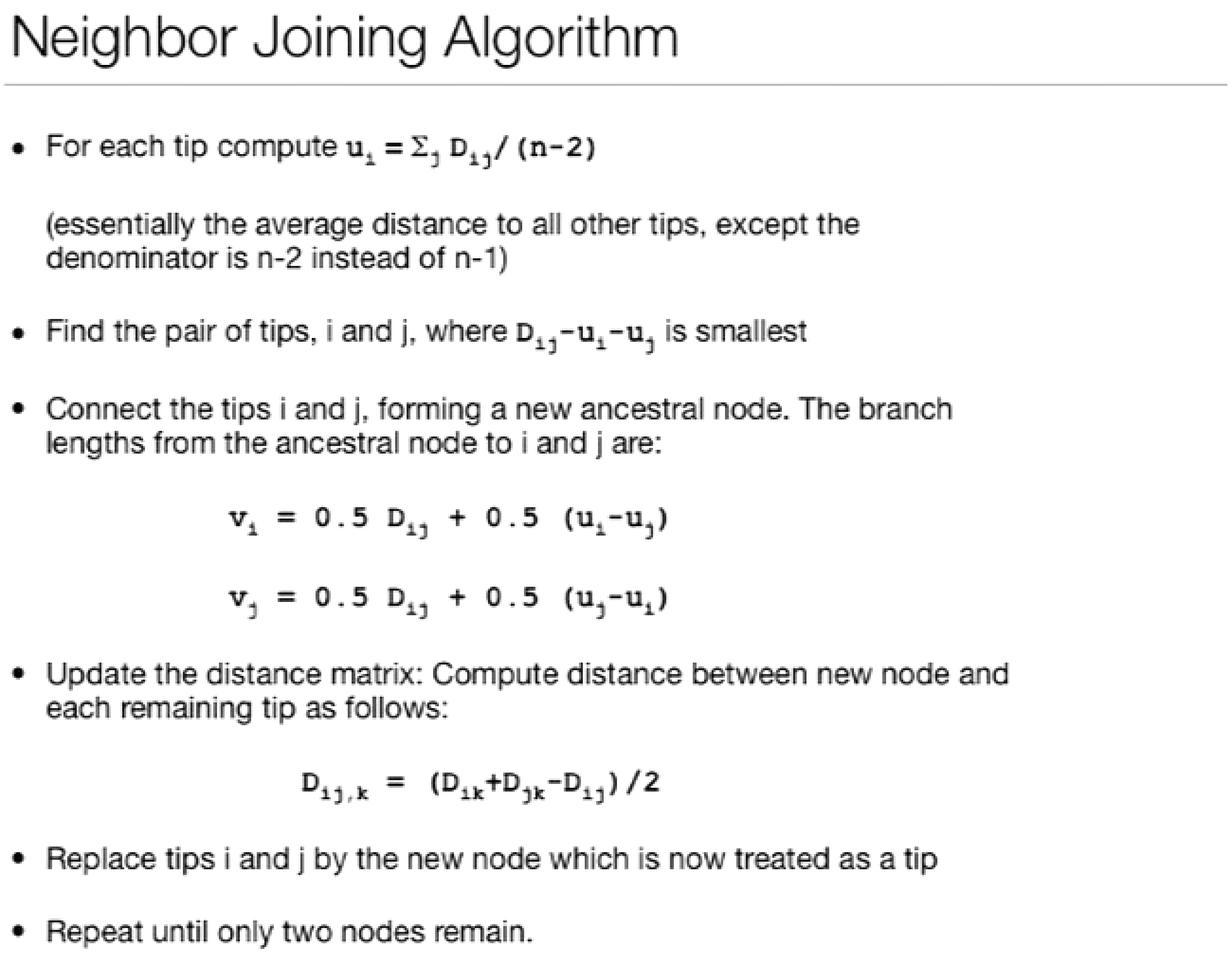
- Four nucleotides assumed to be equally frequent (f=0.25)
- All 12 substitution rates assumed to be equal
- Under this model the corrected distance is: $D_{\mathrm{JC}} = -\frac{3}{4}\ln(1 - \frac{4}{3}D_{\mathrm{OBS}})$
- For instance: $D_{\mathrm{OBS}} = 0.42 \Rightarrow D_{\mathrm{JC}} = 0.62$

Clustering Algorithm: Neighbor Joining (NJ)



Clustering Algorithms

- Starting point: Distance matrix
- Cluster the two nearest nodes:
  - Tree: connect pair of nodes to common ancestral node, compute branch lengths from ancestral node to both descendants
  - Distance matrix: replace the two joined nodes with the new (ancestral) node. Compute new distance matrix, by finding distance from new node to all other nodes
- Repeat until all nodes are linked in tree
- Results in only one tree, there is no measure of tree-goodness.



Neighbor Joining Algorithm

- For each tip compute $u_i = \Sigma_j D_{ij}/(n-2)$

  (essentially the average distance to all other tips, except the denominator is n-2 instead of n-1)

- Find the pair of tips, i and j, where $D_{ij}-u_i-u_j$ is smallest

- Connect the tips i and j, forming a new ancestral node. The branch lengths from the ancestral node to i and j are:

$$v_i = 0.5\ D_{ij} + 0.5\ (u_i-u_j)$$

$$v_j = 0.5\ D_{ij} + 0.5\ (u_j-u_i)$$

- Update the distance matrix: Compute distance between new node and each remaining tip as follows:

$$D_{ij,k} = (D_{ik}+D_{jk}-D_{ij})/2$$

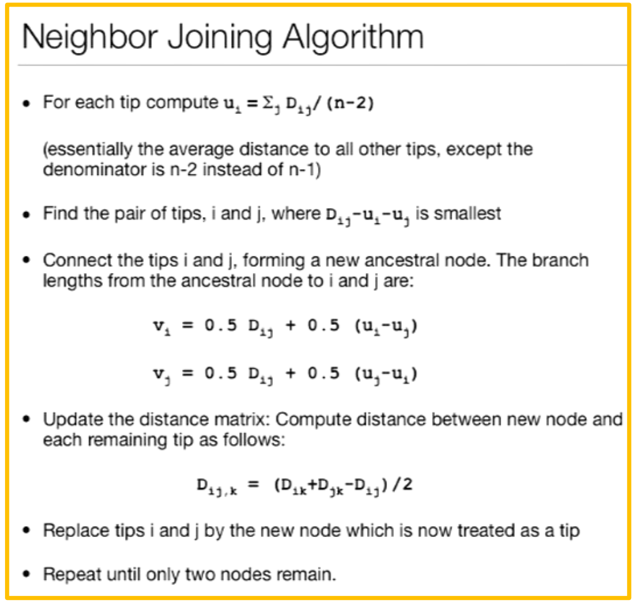
- Replace tips i and j by the new node which is now treated as a tip

- Repeat until only two nodes remain.

Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

i	• $u_i = \Sigma_j D_{ij} / (n-2)$
A	
B	
C	
D	



Neighbor Joining Algorithm

- For each tip compute $u_i = \Sigma_j D_{ij} / (n-2)$

  (essentially the average distance to all other tips, except the denominator is n-2 instead of n-1)

- Find the pair of tips, i and j, where $D_{ij}-u_i-u_j$ is smallest
- Connect the tips i and j, forming a new ancestral node. The branch lengths from the ancestral node to i and j are:

  $$v_i = 0.5\ D_{ij} + 0.5\ (u_i-u_j)$$

  $$v_j = 0.5\ D_{ij} + 0.5\ (u_j-u_i)$$

- Update the distance matrix: Compute distance between new node and each remaining tip as follows:

  $$D_{ij,k} = (D_{ik}+D_{jk}-D_{ij})/2$$

- Replace tips i and j by the new node which is now treated as a tip
- Repeat until only two nodes remain.

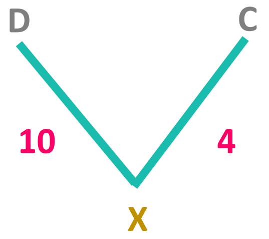
	A	B	C	D
A	-	-39	-35	-35
B		-	-35	-35
C			-	-39
D				-

$D_{ij}-u_i-u_j$

$$v_i = 0.5\ D_{ij} + 0.5\ (u_i-u_j)$$

$$v_j = 0.5\ D_{ij} + 0.5\ (u_j-u_i)$$

$$v_C = 0.5 \times 14 + 0.5 \times (23.5-29.5) = 4$$
$$v_D = 0.5 \times 14 + 0.5 \times (29.5-23.5) = 10$$



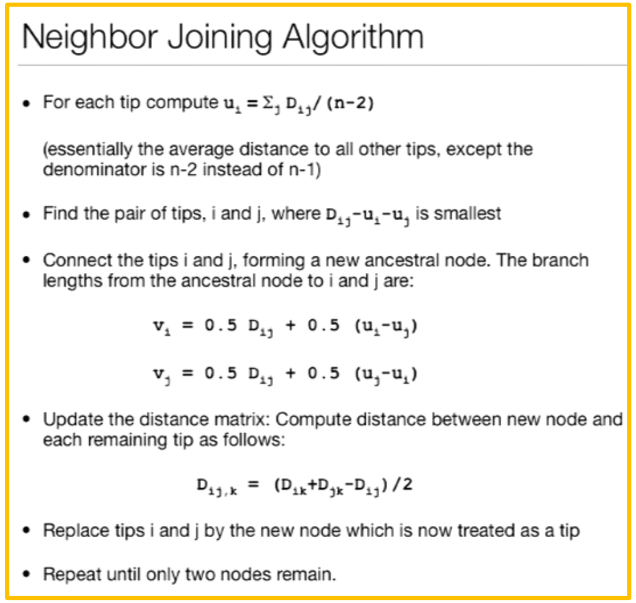
Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

	A	B	C	D	X
A	-	17	21	27	
B		-	12	18	
C			-	14	
D				-	
X					-

	A	B	X
A	-	17	17
B		-	8
X			-

i	• $u_i = \Sigma_j D_{ij}/(n-2)$
A	(17+17)/1 = 34
B	(17+8)/1 = 25
X	(17+8)/1 = 25



Neighbor Joining Algorithm

- For each tip compute $u_i = \Sigma_j D_{ij}/(n-2)$

  (essentially the average distance to all other tips, except the denominator is n-2 instead of n-1)

- Find the pair of tips, i and j, where $D_{ij}-u_i-u_j$ is smallest

- Connect the tips i and j, forming a new ancestral node. The branch lengths from the ancestral node to i and j are:

  $$v_i = 0.5\ D_{ij} + 0.5\ (u_i-u_j)$$

  $$v_j = 0.5\ D_{ij} + 0.5\ (u_j-u_i)$$

- Update the distance matrix: Compute distance between new node and each remaining tip as follows:

  $$D_{ij,k} = (D_{ik}+D_{jk}-D_{ij})/2$$

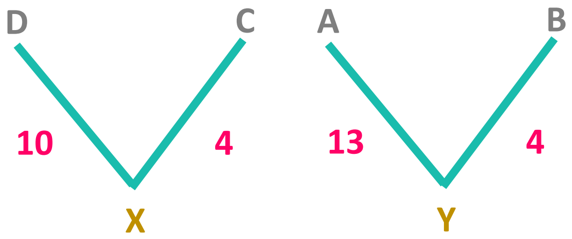
- Replace tips i and j by the new node which is now treated as a tip

- Repeat until only two nodes remain.

	A	B	X
A	-	-42	-42
B		-	-42
X			-

$D_{ij}-u_i-u_j$

$$v_A = 0.5 \times 17 + 0.5 \times (34-25) = 13$$
$$v_D = 0.5 \times 17 + 0.5 \times (25-34) = 4$$



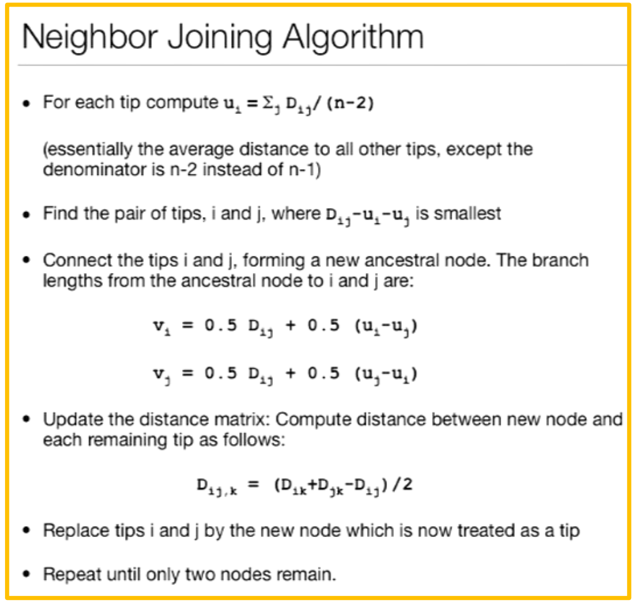
Neighbor Joining Algorithm

	A	B	C	D
A	-	17	21	27
B		-	12	18
C			-	14
D				-

	A	B	C	D	X
A	-	17	21	27	
B		-	12	18	
C			-	14	
D				-	
X					-

	A	B	X
A	-	17	17
B		-	8
X			-

	X	Y
X	-	4
Y		-



Neighbor Joining Algorithm

- For each tip compute $u_i = \Sigma_j D_{ij}/(n-2)$

  (essentially the average distance to all other tips, except the denominator is n-2 instead of n-1)

- Find the pair of tips, i and j, where $D_{ij}-u_i-u_j$ is smallest

- Connect the tips i and j, forming a new ancestral node. The branch lengths from the ancestral node to i and j are:

  $$v_i = 0.5\ D_{ij} + 0.5\ (u_i-u_j)$$

  $$v_j = 0.5\ D_{ij} + 0.5\ (u_j-u_i)$$

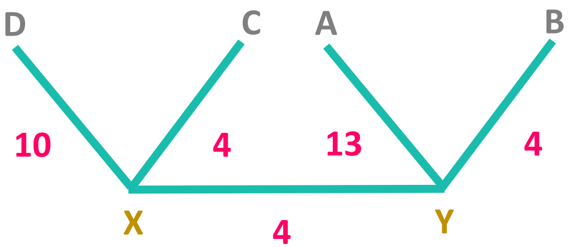
- Update the distance matrix: Compute distance between new node and each remaining tip as follows:

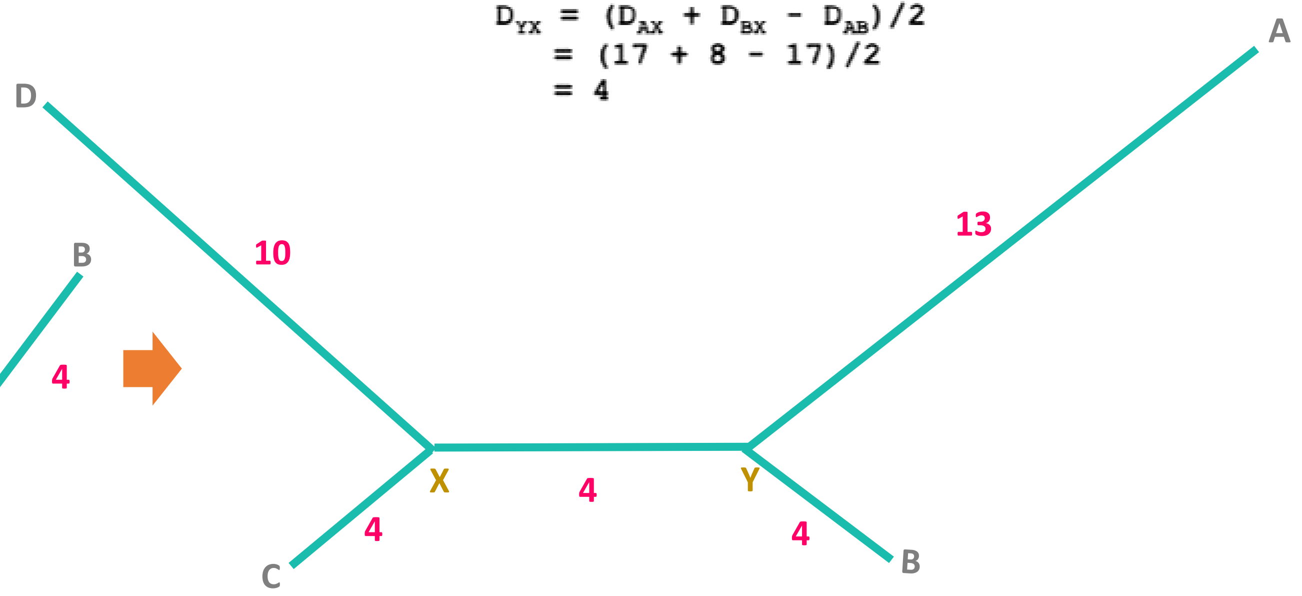
  $$D_{ij,k} = (D_{ik}+D_{jk}-D_{ij})/2$$

- Replace tips i and j by the new node which is now treated as a tip

- Repeat until only two nodes remain.

$$\begin{aligned} D_{YX} &= (D_{AX} + D_{BX} - D_{AB})/2 \\ &= (17 + 8 - 17)/2 \\ &= 4 \end{aligned}$$





Models of Evolotion

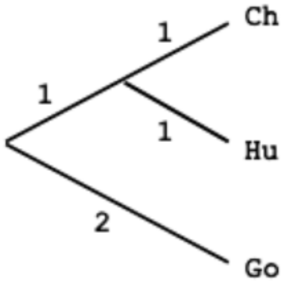
Distance Matrix Methods

```
                     ↓↓↓↓
Gorilla    :  ACGTCGTA
Human      :  ACGTTCCT
Chimpanzee:  ACGTTTCG
                      ↑ ↑
```

1. Construct multiple alignment of sequences

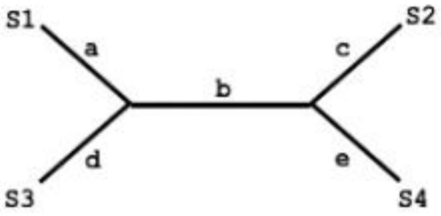
	Go	Hu	Ch
Go	-	4	4
Hu		-	2
Ch			-

2. Construct table listing all pairwise differences (distance matrix)



3. Construct tree from pairwise distances

Optimal Branch Lengths for a Given Tree: Least Squares



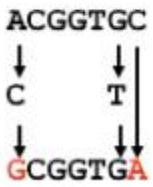
Distance along tree

Goal:

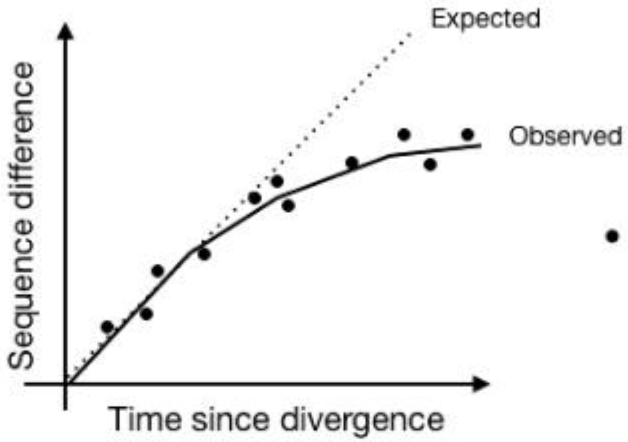
$$\begin{aligned} D_{12} &\approx d_{12} = a + b + c \\ D_{13} &\approx d_{13} = a + d \\ D_{14} &\approx d_{14} = a + b + e \\ D_{23} &\approx d_{23} = d + b + c \\ D_{24} &\approx d_{24} = c + e \\ D_{34} &\approx d_{34} = d + b + e \end{aligned}$$

- Fit between given tree and observed distances can be expressed as “sum of squared differences”:
- $Q = \sum_{j>i} (D_{ij} - d_{ij})^2$
- Find branch lengths that minimize Q - this is the optimal set of branch lengths for this tree.

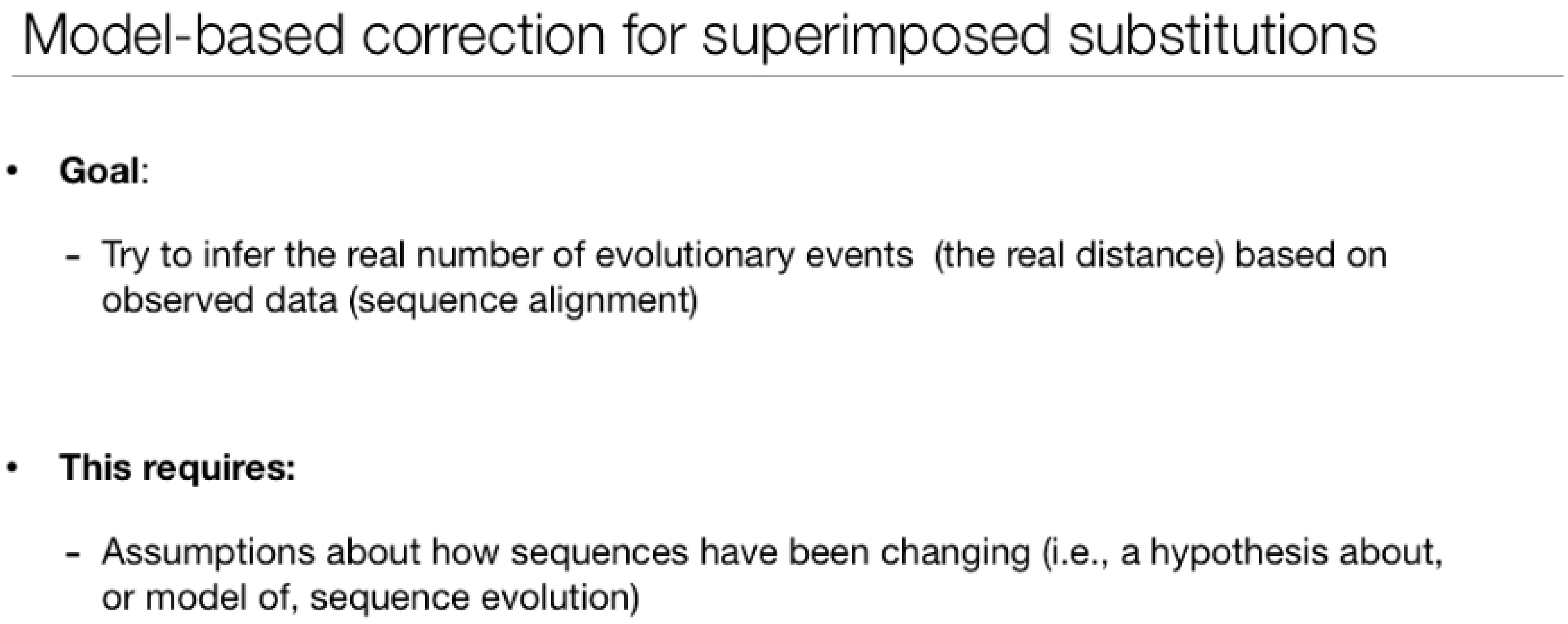
Superimposed Substitutions



- Actual number of evolutionary events: 5
- Observed number of differences: 2



- Distance is (almost) always underestimated



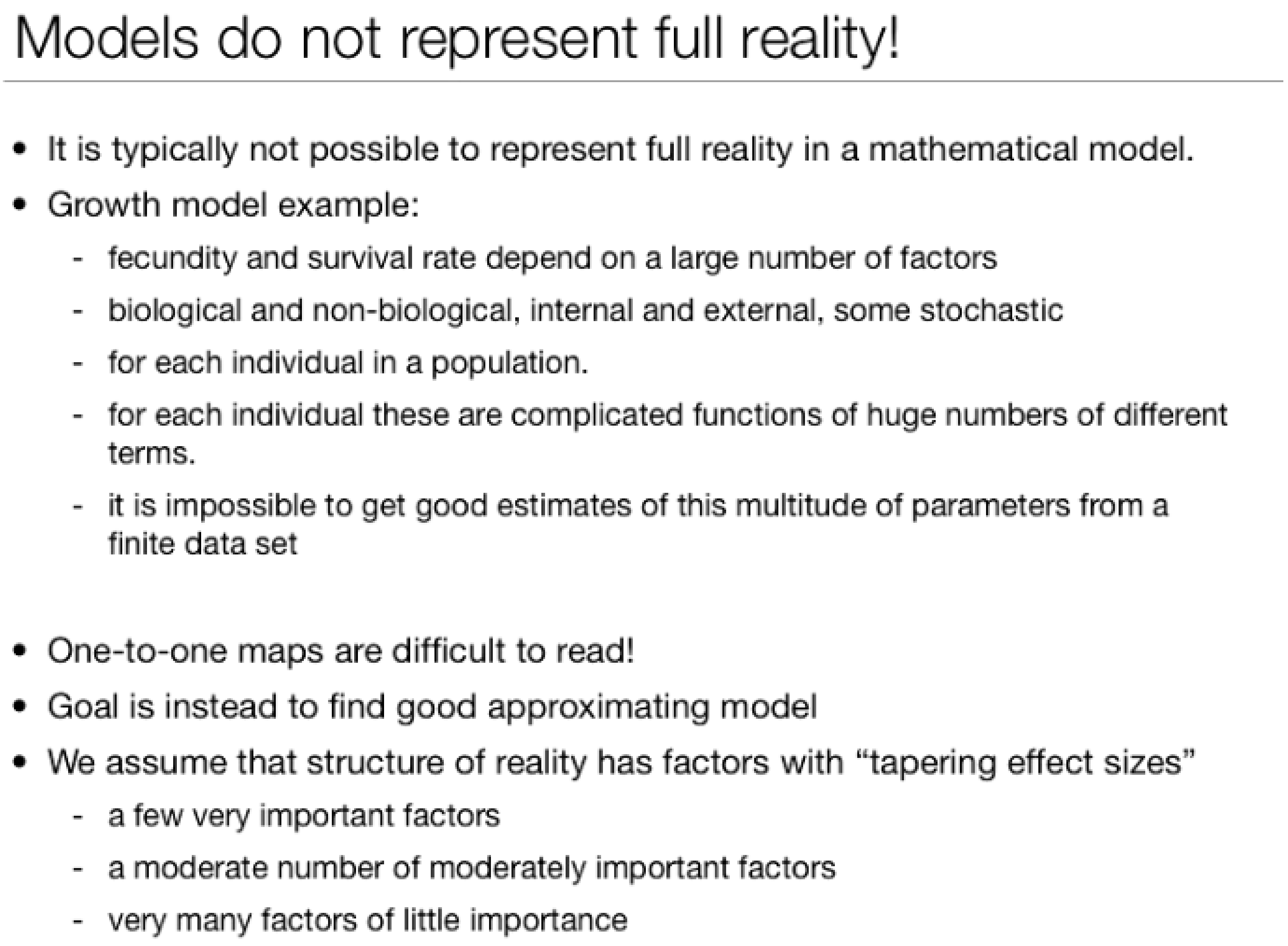
Model-based correction for superimposed substitutions

- **Goal**:
  - Try to infer the real number of evolutionary events (the real distance) based on observed data (sequence alignment)

- **This requires:**
  - Assumptions about how sequences have been changing (i.e., a hypothesis about, or model of, sequence evolution)

What is a Model?

- Model = stringently phrased hypothesis !!!
- Hypothesis (as used in most biological research):
  - Precisely stated, but qualitative
  - Allows you to make qualitative predictions
  - Example: “Population size grows rapidly when there are few individuals, but growth rate declines when resources become limiting.”
- Arithmetic model:
  - Mathematically explicit (parameters)
  - Allows you to make quantitative predictions
  - Example: $N_t = \frac{K}{1 + \left(\frac{K}{N_0} - 1\right) e^{-rt}}$

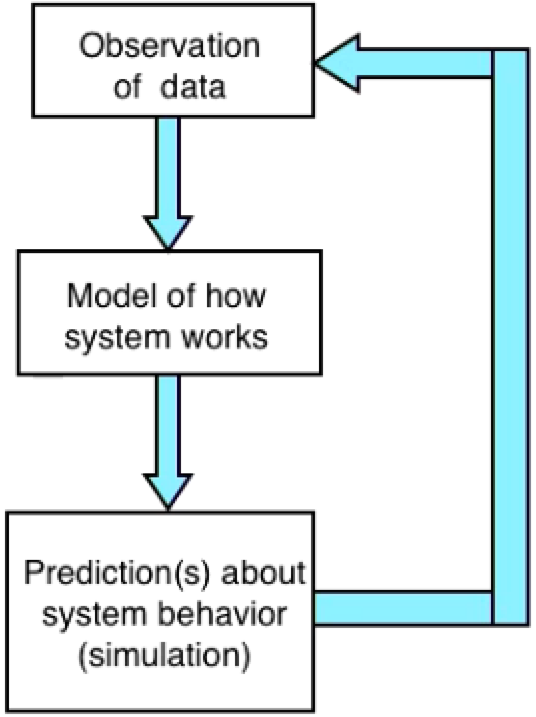


Models do not represent full reality!

- It is typically not possible to represent full reality in a mathematical model.
- Growth model example:
  - fecundity and survival rate depend on a large number of factors
  - biological and non-biological, internal and external, some stochastic
  - for each individual in a population.
  - for each individual these are complicated functions of huge numbers of different terms.
  - it is impossible to get good estimates of this multitude of parameters from a finite data set

- One-to-one maps are difficult to read!
- Goal is instead to find good approximating model
- We assume that structure of reality has factors with “tapering effect sizes”
  - a few very important factors
  - a moderate number of moderately important factors
  - very many factors of little importance

The Scientific Method


Observation of data
Model of how system works
Prediction(s) about system behavior (simulation)

Jukes and Cantor Model of Nucleotide Substitution

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

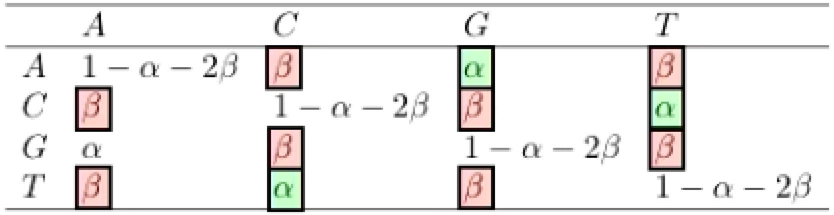
Relative rate matrix

$$\Rightarrow P(t) = e^{Qt} = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix}$$

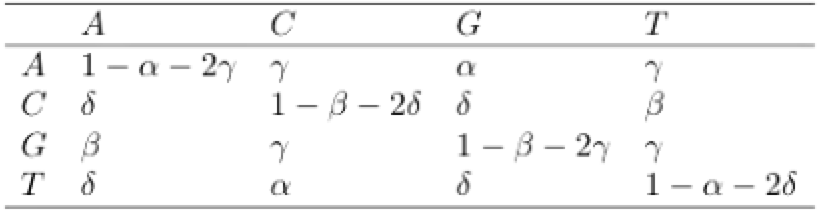
Probability matrix
(function of time)

- Four nucleotides assumed to be equally frequent (f=0.25)
- All 12 substitution rates assumed to be equal
- Under this model the corrected distance is: $D_{\mathrm{JC}} = \frac{3}{4}\ln(1 - \frac{4}{3}D_{\mathrm{OBS}})$
- For instance: $D_{\mathrm{OBS}} = 0.42 \Rightarrow D_{\mathrm{JC}} = 0.64$

Other models of evolution

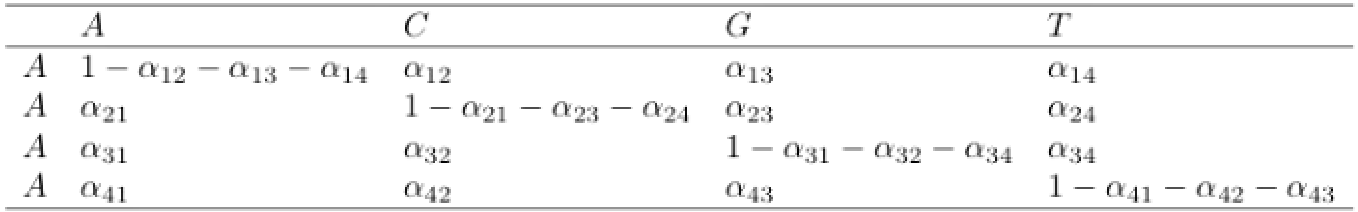


	A	C	G	T
A	$1-\alpha-2\beta$	$\boxed{\beta}$	$\boxed{\alpha}$	$\boxed{\beta}$
C	$\boxed{\beta}$	$1-\alpha-2\beta$	$\boxed{\beta}$	$\boxed{\alpha}$
G	α	$\boxed{\beta}$	$1-\alpha-2\beta$	$\boxed{\beta}$
T	$\boxed{\beta}$	$\boxed{\alpha}$	$\boxed{\beta}$	$1-\alpha-2\beta$

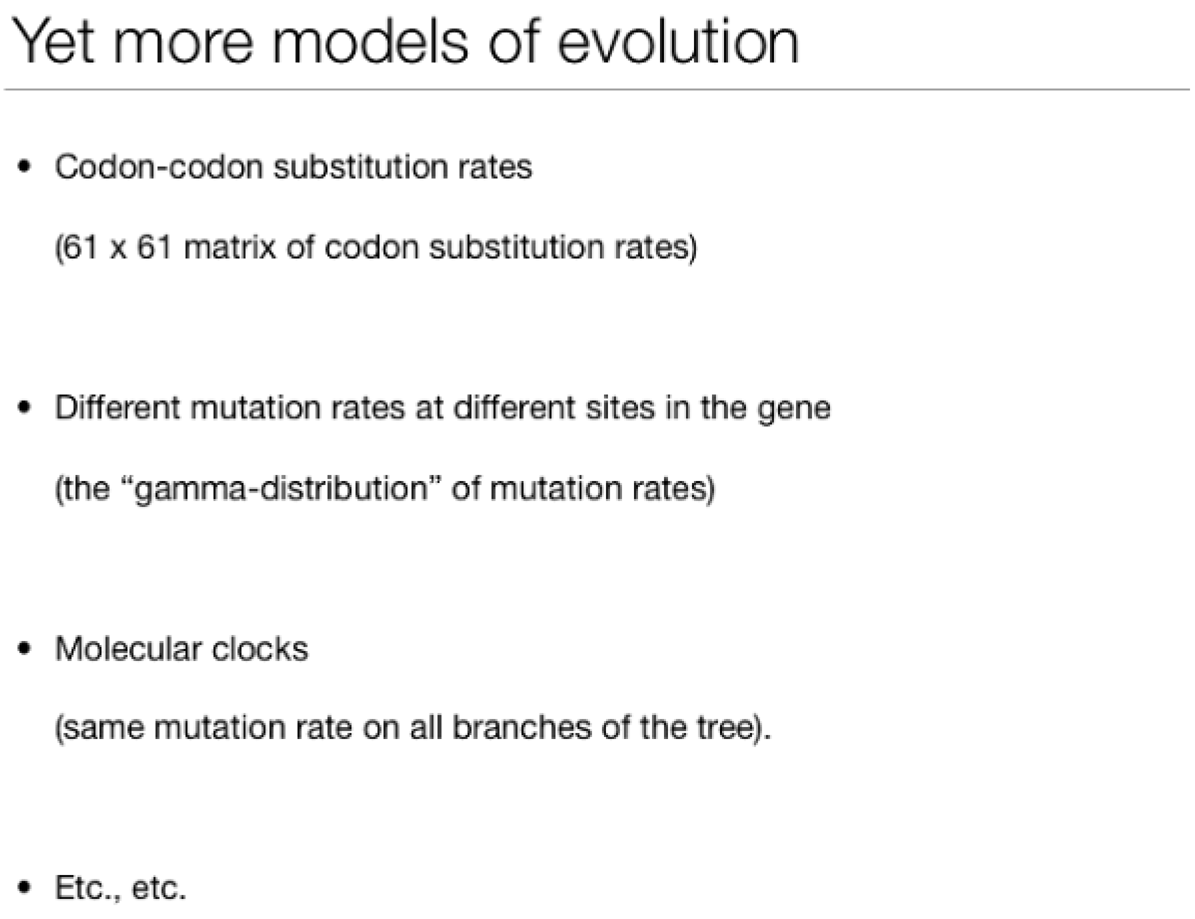


	A	C	G	T
A	$1-\alpha-2\gamma$	γ	α	γ
C	δ	$1-\beta-2\delta$	δ	β
G	β	γ	$1-\beta-2\gamma$	γ
T	δ	α	δ	$1-\alpha-2\delta$

$\vdots$



	A	C	G	T
A	$1-\alpha_{12}-\alpha_{13}-\alpha_{14}$	α_{12}	α_{13}	α_{14}
A	α_{21}	$1-\alpha_{21}-\alpha_{23}-\alpha_{24}$	α_{23}	α_{24}
A	α_{31}	α_{32}	$1-\alpha_{31}-\alpha_{32}-\alpha_{34}$	α_{34}
A	α_{41}	α_{42}	α_{43}	$1-\alpha_{41}-\alpha_{42}-\alpha_{43}$



Yet more models of evolution

- Codon-codon substitution rates

  (61 x 61 matrix of codon substitution rates)

- Different mutation rates at different sites in the gene

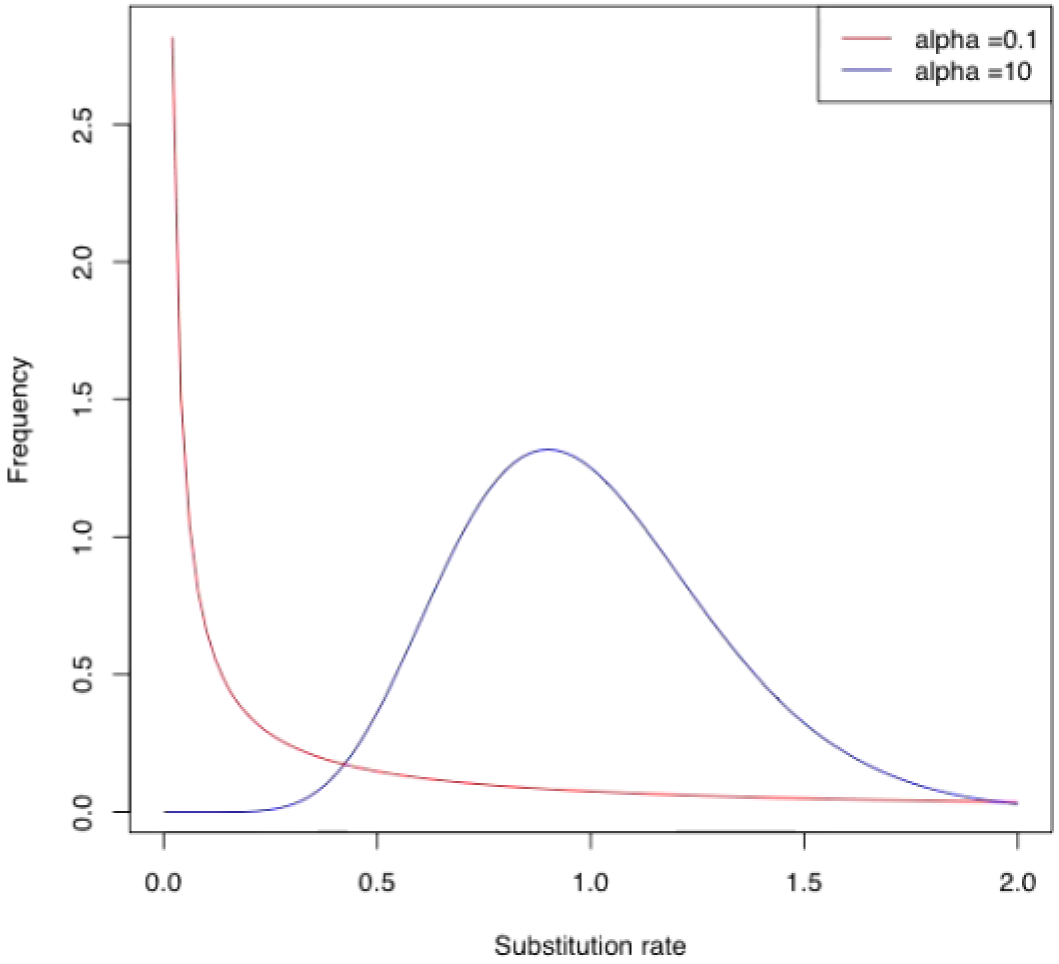
  (the “gamma-distribution” of mutation rates)

- Molecular clocks

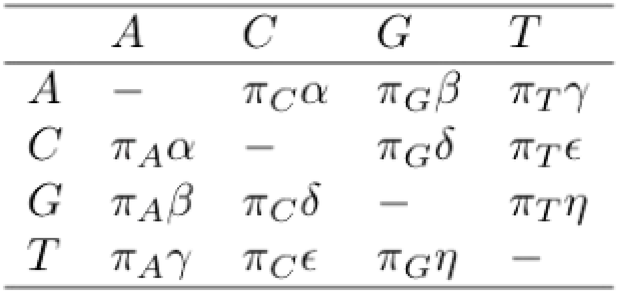
  (same mutation rate on all branches of the tree).

- Etc., etc.

Different rates at different sites: the gamma distribution



General Time Reversible Model

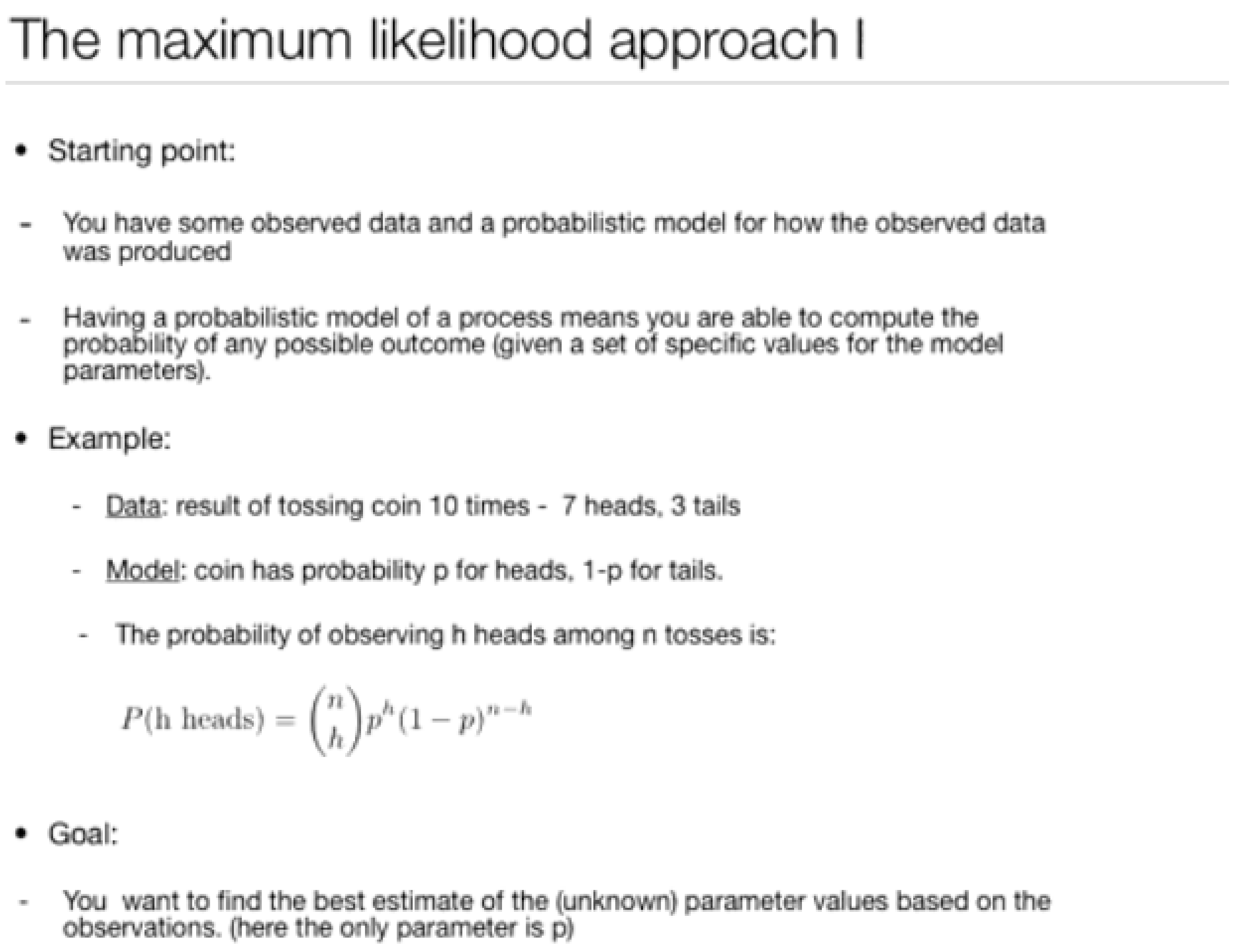


	A	C	G	T
A	$-$	$\pi_C\alpha$	$\pi_G\beta$	$\pi_T\gamma$
C	$\pi_A\alpha$	$-$	$\pi_G\delta$	$\pi_T\epsilon$
G	$\pi_A\beta$	$\pi_C\delta$	$-$	$\pi_T\eta$
T	$\pi_A\gamma$	$\pi_C\epsilon$	$\pi_G\eta$	$-$

- Time-reversibility:
  The amount of change from state x to y is equal to the amount of change from y to x

$$\pi_A \times \mathrm{rate}_{AG} = \pi_G \times \mathrm{rate}_{GA} \Leftrightarrow \pi_A\pi_G\beta = \pi_G\pi_A\beta$$

Maximum Likelihood



The maximum likelihood approach I

- Starting point:

- You have some observed data and a probabilistic model for how the observed data was produced

- Having a probabilistic model of a process means you are able to compute the probability of any possible outcome (given a set of specific values for the model parameters).

- Example:

  - Data: result of tossing coin 10 times - 7 heads, 3 tails

  - Model: coin has probability p for heads, 1-p for tails.

  - The probability of observing h heads among n tosses is:

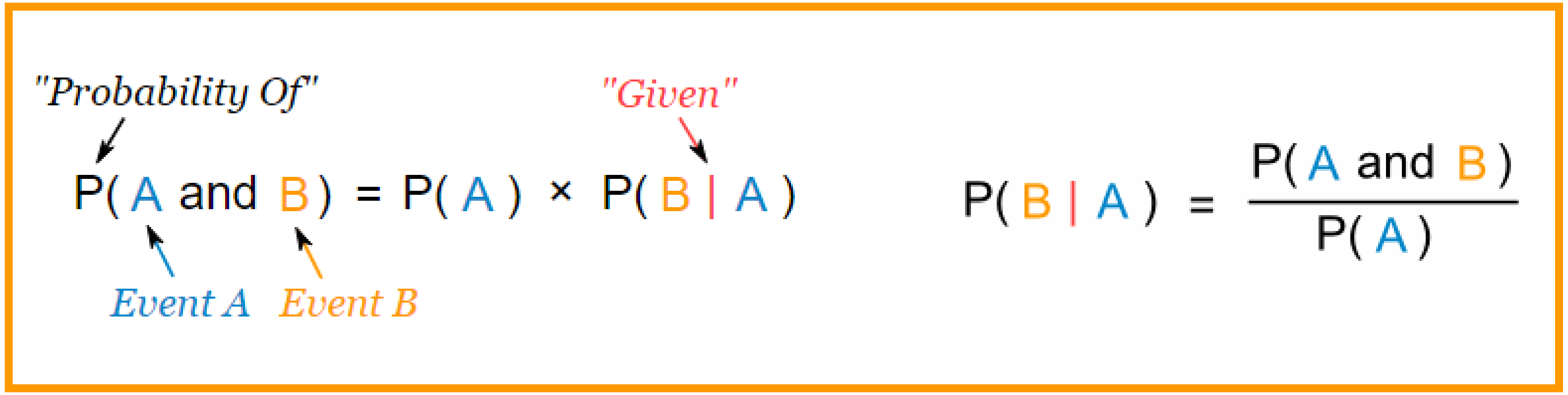
$$P(\text{h heads}) = \binom{n}{h} p^h (1-p)^{n-h}$$

- Goal:

- You want to find the best estimate of the (unknown) parameter values based on the observations. (here the only parameter is p)

The maximum likelihood approach II

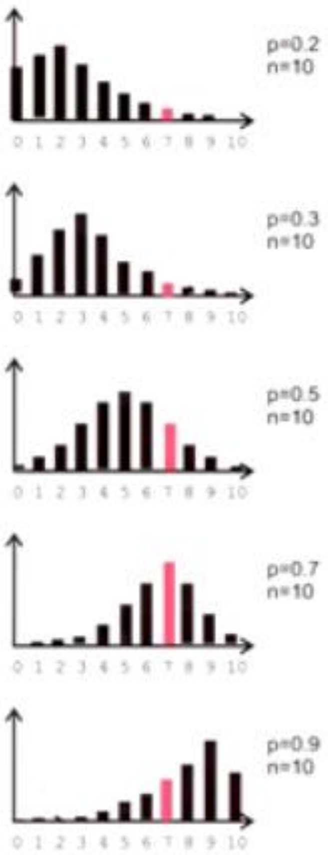
- Likelihood (Model) = Probability (Data I Model)
- Maximum likelihood: Best estimate is the set of parameter values which gives the highest possible likelihood.



$$P(\,A \text{ and } B\,) = P(\,A\,) \times P(\,B \mid A\,)$$

$$P(\,B \mid A\,) = \frac{P(\,A \text{ and } B\,)}{P(\,A\,)}$$

Maximum likelihood: coin tossing example



Probability distribution for possible outcomes when value of p-parameter=0.2 and n=10 tosses of coin.

Probabilities sum to 1.

Likelihood of p having the value 0.2 given that we observed x=7 heads:

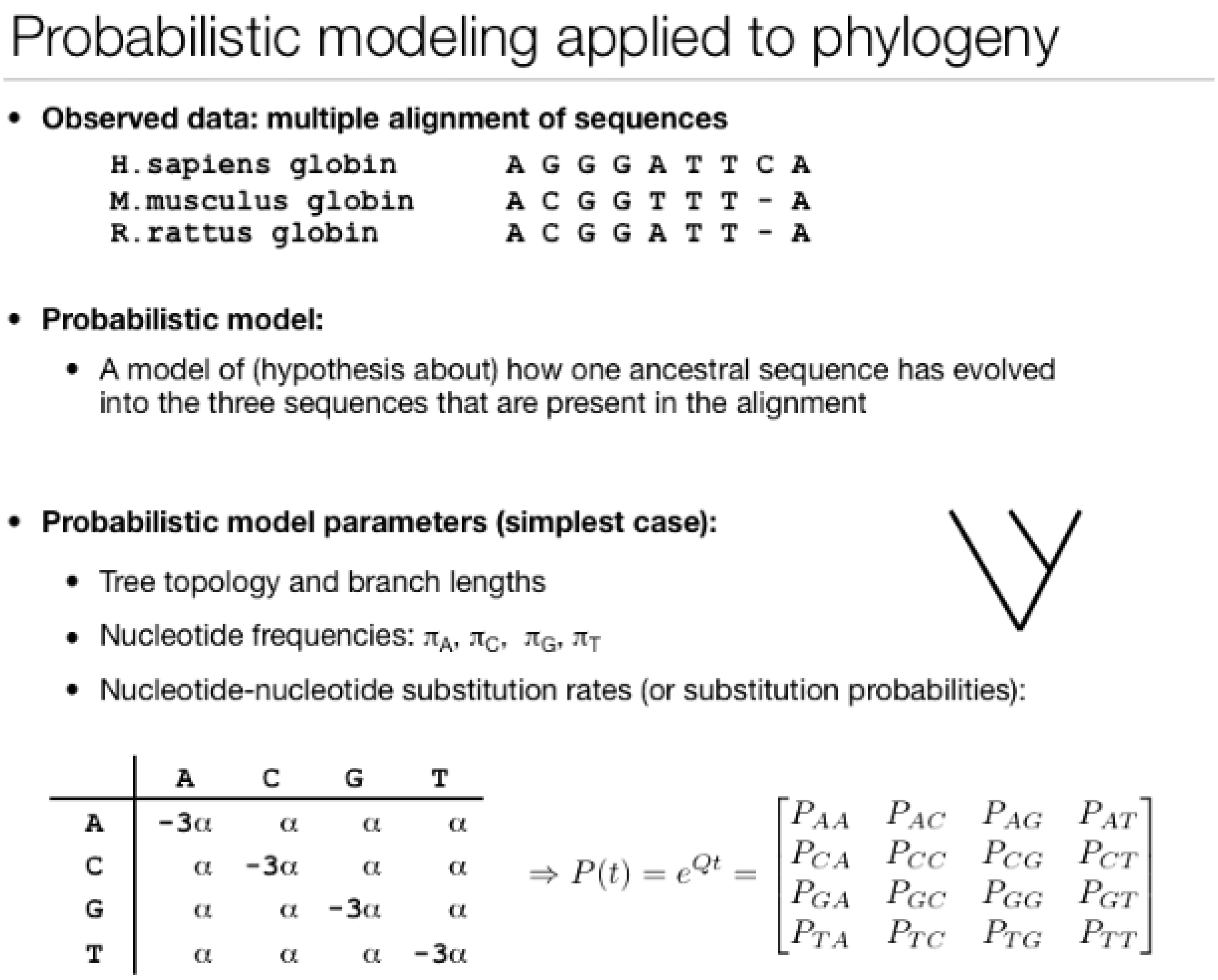
L(p=0.2 | x=7) =

Pr(x=7 | p=0.2) = 0.001

p=0.7 is the maximum likelihood estimate of p given that we observed x=7 heads.

Note that the likelihoods L(p | x=7) do not necessarily sum to 1

- Data: result of tossing coin 10 times - 7 heads, 3 tails
- Model: coin has probability p for heads, 1-p for tails.



Probabilistic modeling applied to phylogeny

- **Observed data: multiple alignment of sequences**

```
H.sapiens globin      A G G G A T T C A
M.musculus globin     A C G G T T T - A
R.rattus globin       A C G G A T T - A
```

- **Probabilistic model:**
  - A model of (hypothesis about) how one ancestral sequence has evolved into the three sequences that are present in the alignment

- **Probabilistic model parameters (simplest case):**
  - Tree topology and branch lengths
  - Nucleotide frequencies: π_A, π_C, π_G, π_T
  - Nucleotide-nucleotide substitution rates (or substitution probabilities):

	A	C	G	T
A	-3α	α	α	α
C	α	-3α	α	α
G	α	α	-3α	α
T	α	α	α	-3α

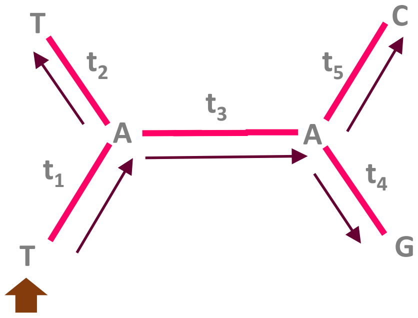
$$\Rightarrow P(t) = e^{Qt} = \begin{bmatrix} P_{AA} & P_{AC} & P_{AG} & P_{AT} \\ P_{CA} & P_{CC} & P_{CG} & P_{CT} \\ P_{GA} & P_{GC} & P_{GG} & P_{GT} \\ P_{TA} & P_{TC} & P_{TG} & P_{TT} \end{bmatrix}$$

Computing the probability of one column in an alignment given tree topology and other parameters

```
A T G G A T T C A
A T G G T T T - A
A C G G A T T - A
A G G G T T T - A
```

- Columns in alignment contain homologous nucleotides

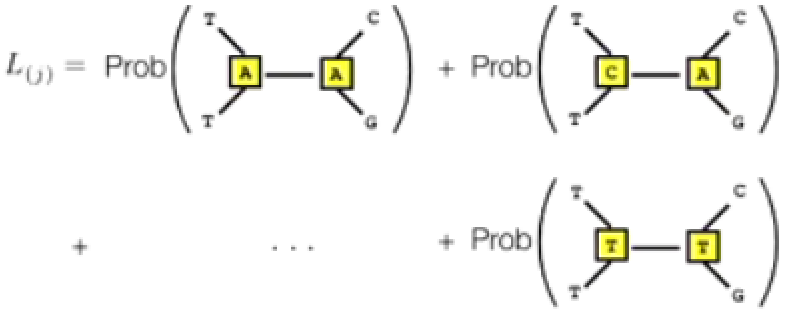
- Assume tree topology, branch lengths, and other parameters are given. For now, assume ancestral states were A and A (we'll get to the full computation on next slide). Start computation at any internal or external node. Arrows indicate "direction" of computations ("flowing" away from the starting point).



$$Pr = \pi_T \, P_{TA}(t_1) \, P_{AT}(t_2) \, P_{AA}(t_3) \, P_{AG}(t_4) \, P_{AC}(t_5)$$

Computing the probability of an entire alignment given tree topology and other parameters

```
A T G G A T T C A
A T G G T T T - A
A C G G A T T - A
A G G G T T T - A
  j
```

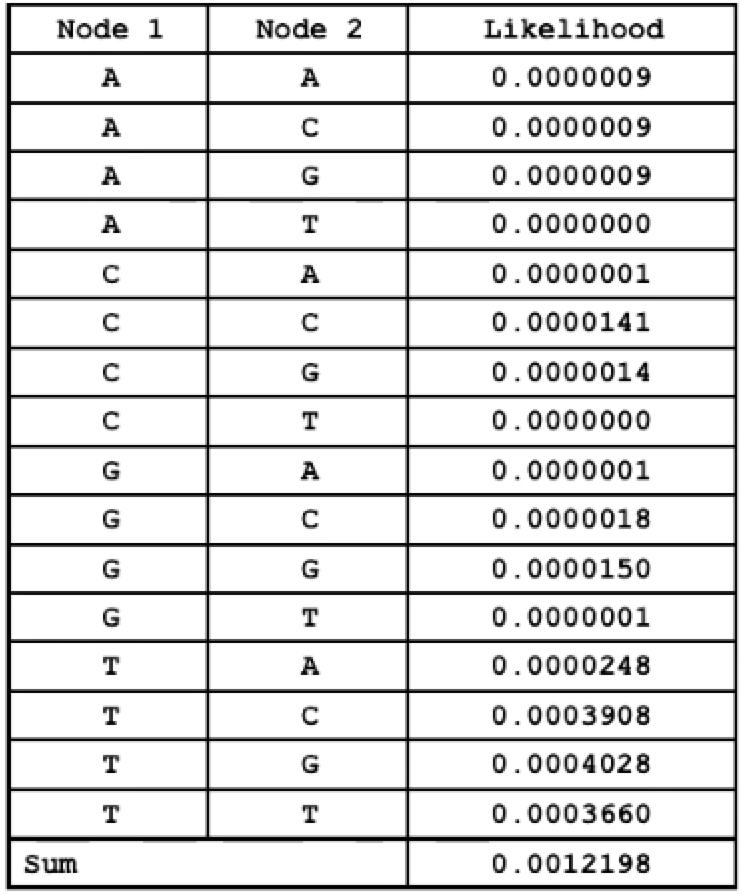


$$L = L_{(1)} \cdot L_{(2)} \cdots L_{(N)} = \prod_{j=1}^{N} L_{(j)}$$

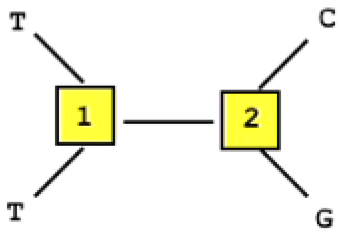
$$\ln(L) = \ln(L_{(1)}) + \ln(L_{(2)}) + \cdots + \ln(L_{(N)}) = \sum_{j=1}^{N} \ln(L_{(j)})$$

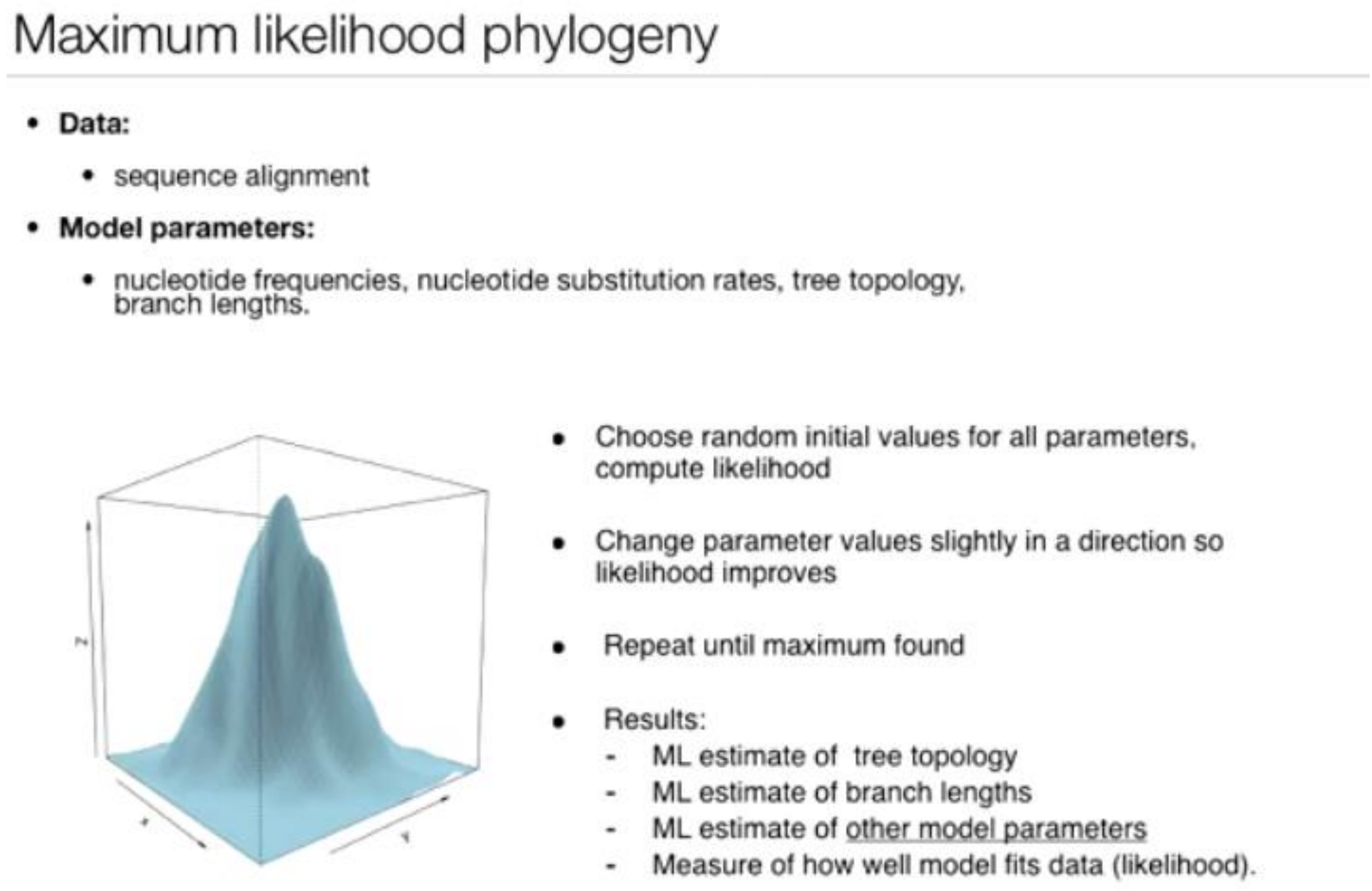
- Probability must be summed over all possible combinations of ancestral nucleotides.
- Here we have two internal nodes giving 16 possible combinations
- Probability of individual columns are multiplied to give the overall probability of the alignment, i.e., the likelihood of the model.
- In phylogeny software these computations are done using summation of the logs of the probabilities ("log likelihoods"), because multiplication of the large number of probability terms may lead to underflow (computer problems caused by very small numbers).

Likelihood of column in alignment:
sum over all possible pairs of ancestral nucleotides



Node 1	Node 2	Likelihood
A	A	0.0000009
A	C	0.0000009
A	G	0.0000009
A	T	0.0000000
C	A	0.0000001
C	C	0.0000141
C	G	0.0000014
C	T	0.0000000
G	A	0.0000001
G	C	0.0000018
G	G	0.0000150
G	T	0.0000001
T	A	0.0000248
T	C	0.0003908
T	G	0.0004028
T	T	0.0003660
Sum		0.0012198





Maximum likelihood phylogeny

- **Data:**
  - sequence alignment
- **Model parameters:**
  - nucleotide frequencies, nucleotide substitution rates, tree topology, branch lengths.

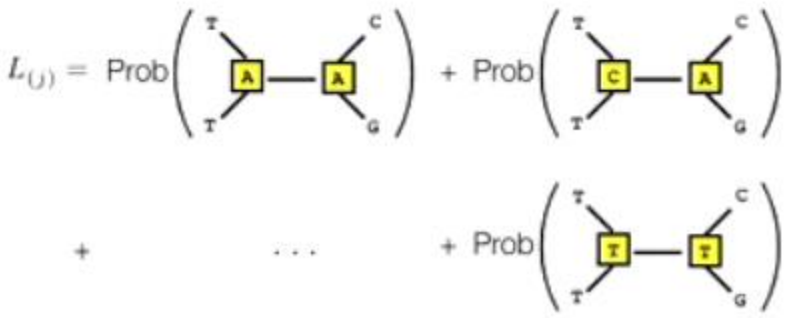
- Choose random initial values for all parameters, compute likelihood
- Change parameter values slightly in a direction so likelihood improves
- Repeat until maximum found
- Results:
  - ML estimate of tree topology
  - ML estimate of branch lengths
  - ML estimate of other model parameters
  - Measure of how well model fits data (likelihood).

Ancestral Reconstruction

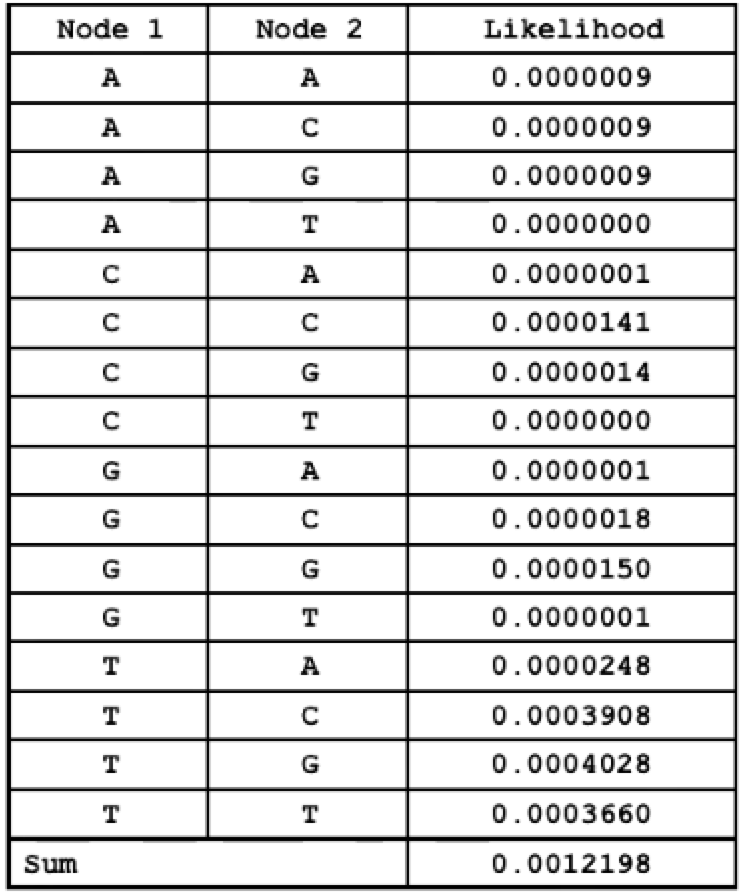
Likelihood of column in alignment:
sum over all possible pairs of ancestral nucleotides

```
A T G G A T T C A
A T G G T T T - A
A C G G A T T - A
A G G G T T T - A
  j
```

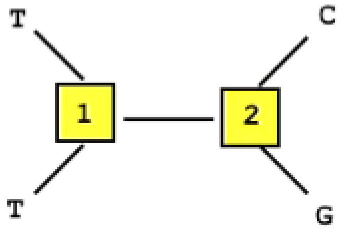
- Probability must be summed over all possible combinations of ancestral nucleotides.
- Here we have two internal nodes giving 16 possible combinations



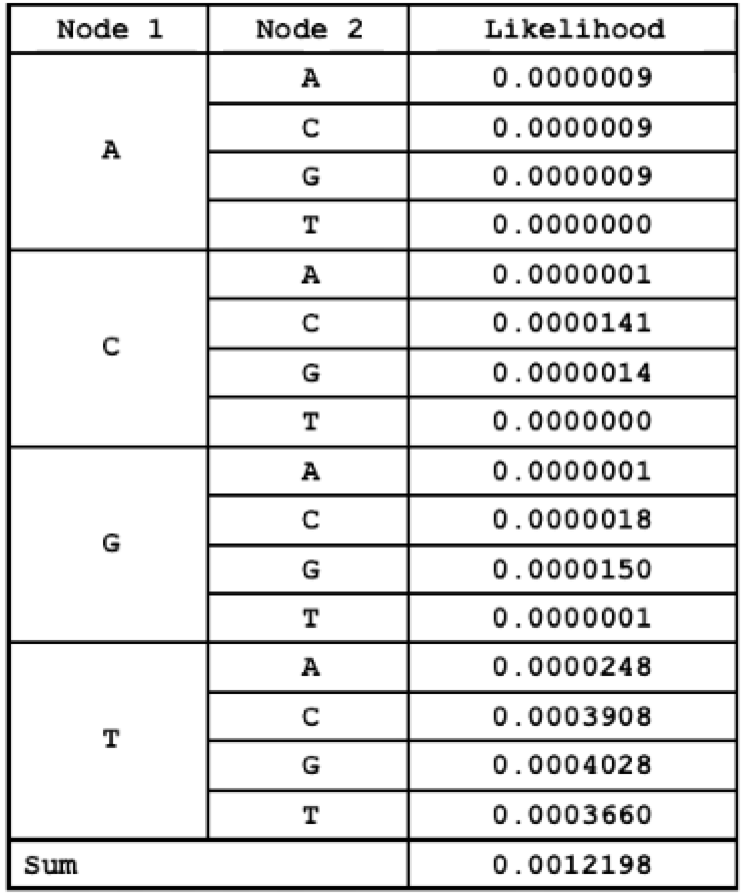
Likelihood of column in alignment:
sum over all possible pairs of ancestral nucleotides



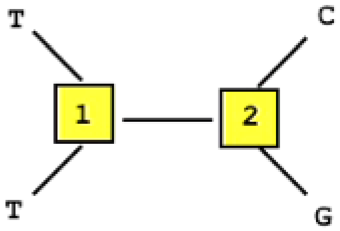
Node 1	Node 2	Likelihood
A	A	0.0000009
A	C	0.0000009
A	G	0.0000009
A	T	0.0000000
C	A	0.0000001
C	C	0.0000141
C	G	0.0000014
C	T	0.0000000
G	A	0.0000001
G	C	0.0000018
G	G	0.0000150
G	T	0.0000001
T	A	0.0000248
T	C	0.0003908
T	G	0.0004028
T	T	0.0003660
Sum		0.0012198



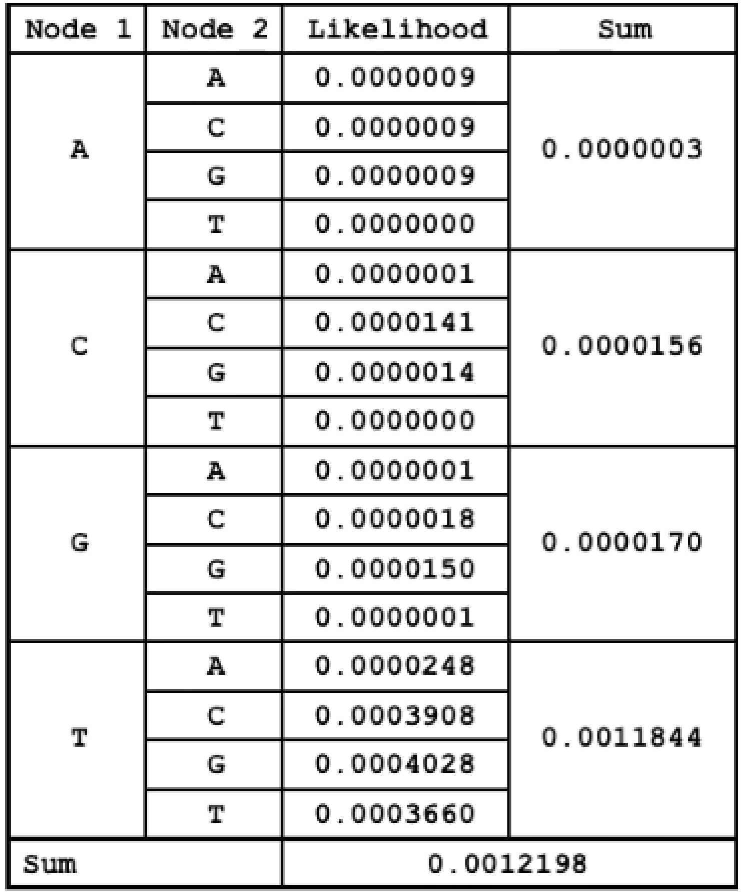
Likelihood of column in alignment: sum over all possible pairs of ancestral nucleotides



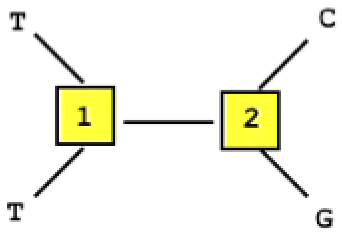
Node 1	Node 2	Likelihood
A	A	0.0000009
	C	0.0000009
	G	0.0000009
	T	0.0000000
C	A	0.0000001
	C	0.0000141
	G	0.0000014
	T	0.0000000
G	A	0.0000001
	C	0.0000018
	G	0.0000150
	T	0.0000001
T	A	0.0000248
	C	0.0003908
	G	0.0004028
	T	0.0003660
Sum		0.0012198



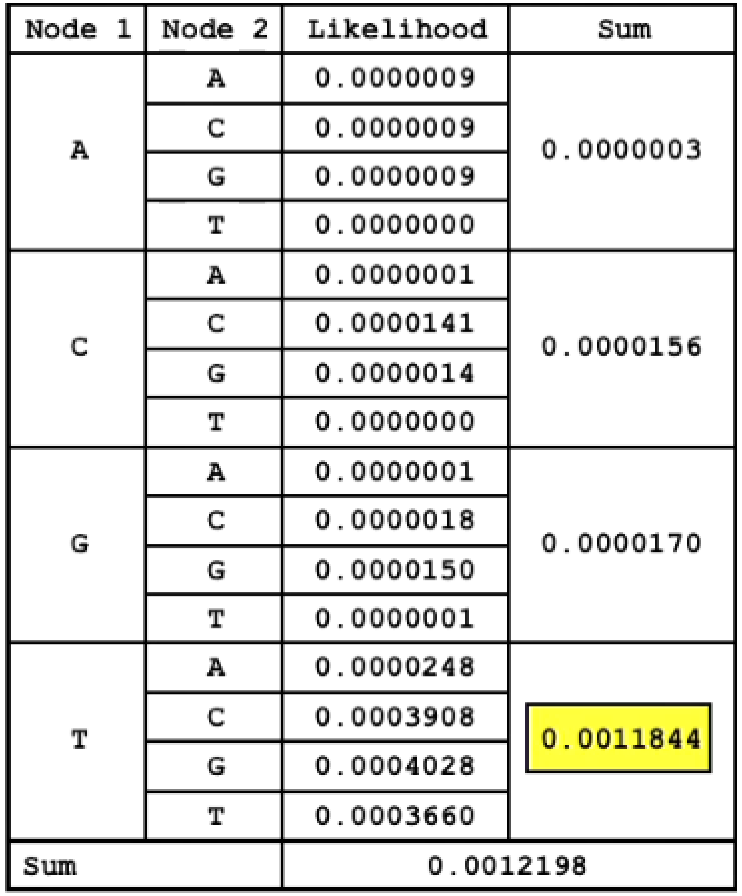
Ancestral Reconstruction:



Node 1	Node 2	Likelihood	Sum
A	A	0.0000009	0.0000003
	C	0.0000009	
	G	0.0000009	
	T	0.0000000	
C	A	0.0000001	0.0000156
	C	0.0000141	
	G	0.0000014	
	T	0.0000000	
G	A	0.0000001	0.0000170
	C	0.0000018	
	G	0.0000150	
	T	0.0000001	
T	A	0.0000248	0.0011844
	C	0.0003908	
	G	0.0004028	
	T	0.0003660	
Sum		0.0012198	



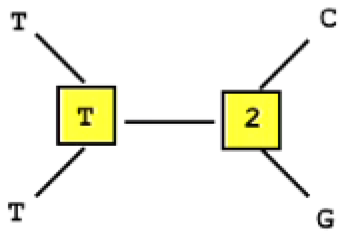
Ancestral Reconstruction:

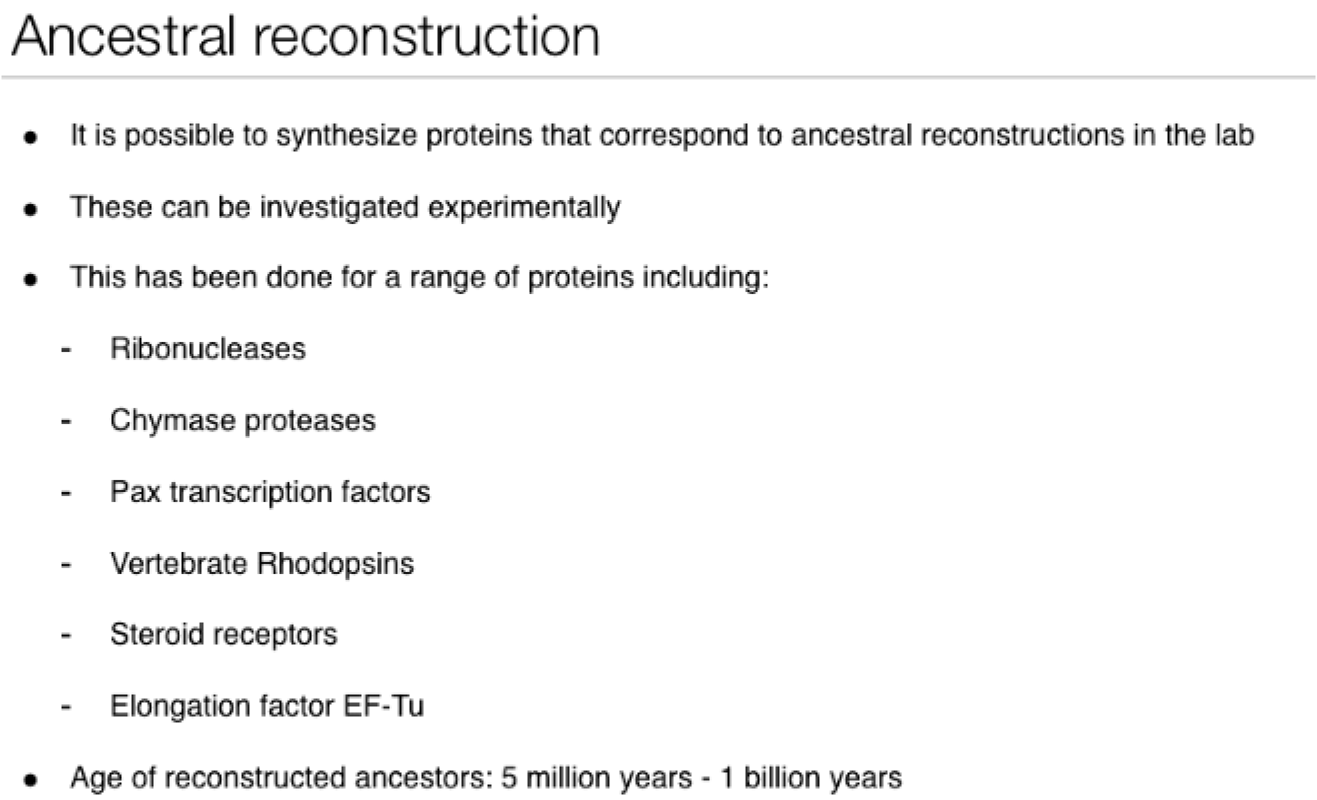


Node 1	Node 2	Likelihood	Sum
A	A	0.0000009	0.0000003
	C	0.0000009	
	G	0.0000009	
	T	0.0000000	
C	A	0.0000001	0.0000156
	C	0.0000141	
	G	0.0000014	
	T	0.0000000	
G	A	0.0000001	0.0000170
	C	0.0000018	
	G	0.0000150	
	T	0.0000001	
T	A	0.0000248	0.0011844
	C	0.0003908	
	G	0.0004028	
	T	0.0003660	
Sum		0.0012198	

Ancestral reconstruction:

Node 1 = T





Ancestral reconstruction

- It is possible to synthesize proteins that correspond to ancestral reconstructions in the lab
- These can be investigated experimentally
- This has been done for a range of proteins including:
    - Ribonucleases
    - Chymase proteases
    - Pax transcription factors
    - Vertebrate Rhodopsins
    - Steroid receptors
    - Elongation factor EF-Tu
- Age of reconstructed ancestors: 5 million years - 1 billion years

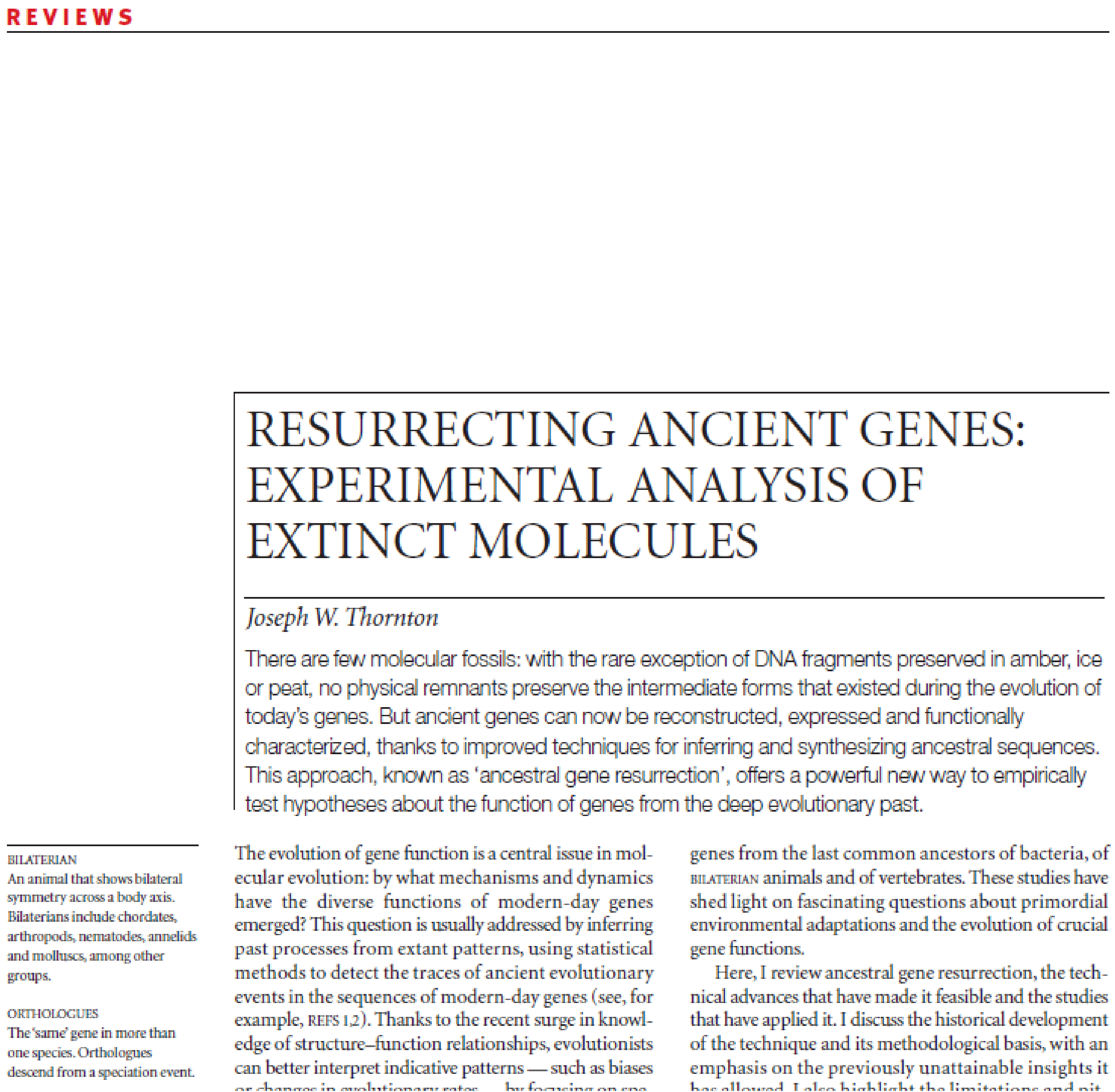
Ancestral reconstruction: dinosaur night vision

Despite its great age, the ancestral rhodopsin functioned well, carrying out all the individual steps that are required for visual function in dim light as effectively as the extant proteins in mammals, which generally have good night vision.

Specifically, the ancestral protein bound the visual chromophore 11-cis-retinal and, when exposed to light, activated the G-protein transducin at a rate similar to that of bovine rhodopsin.

These results are consistent with the hypothesis that the ancestral archosaur possessed the ability — at the molecular level at least — to see well in dim light, and might have been active at night. This insight, of course, could never have been drawn from fossils or any other non-molecular evidence about the behaviour of ancient dinosaurs.

Resurrecting ancient genes: experimental analysis of extinct molecules, Nature Reviews Genetics 5, 366-375 (May 2004). Joseph W. Thornton



REVIEWS

RESURRECTING ANCIENT GENES: EXPERIMENTAL ANALYSIS OF EXTINCT MOLECULES

Joseph W. Thornton

There are few molecular fossils: with the rare exception of DNA fragments preserved in amber, ice or peat, no physical remnants preserve the intermediate forms that existed during the evolution of today's genes. But ancient genes can now be reconstructed, expressed and functionally characterized, thanks to improved techniques for inferring and synthesizing ancestral sequences. This approach, known as 'ancestral gene resurrection', offers a powerful new way to empirically test hypotheses about the function of genes from the deep evolutionary past.

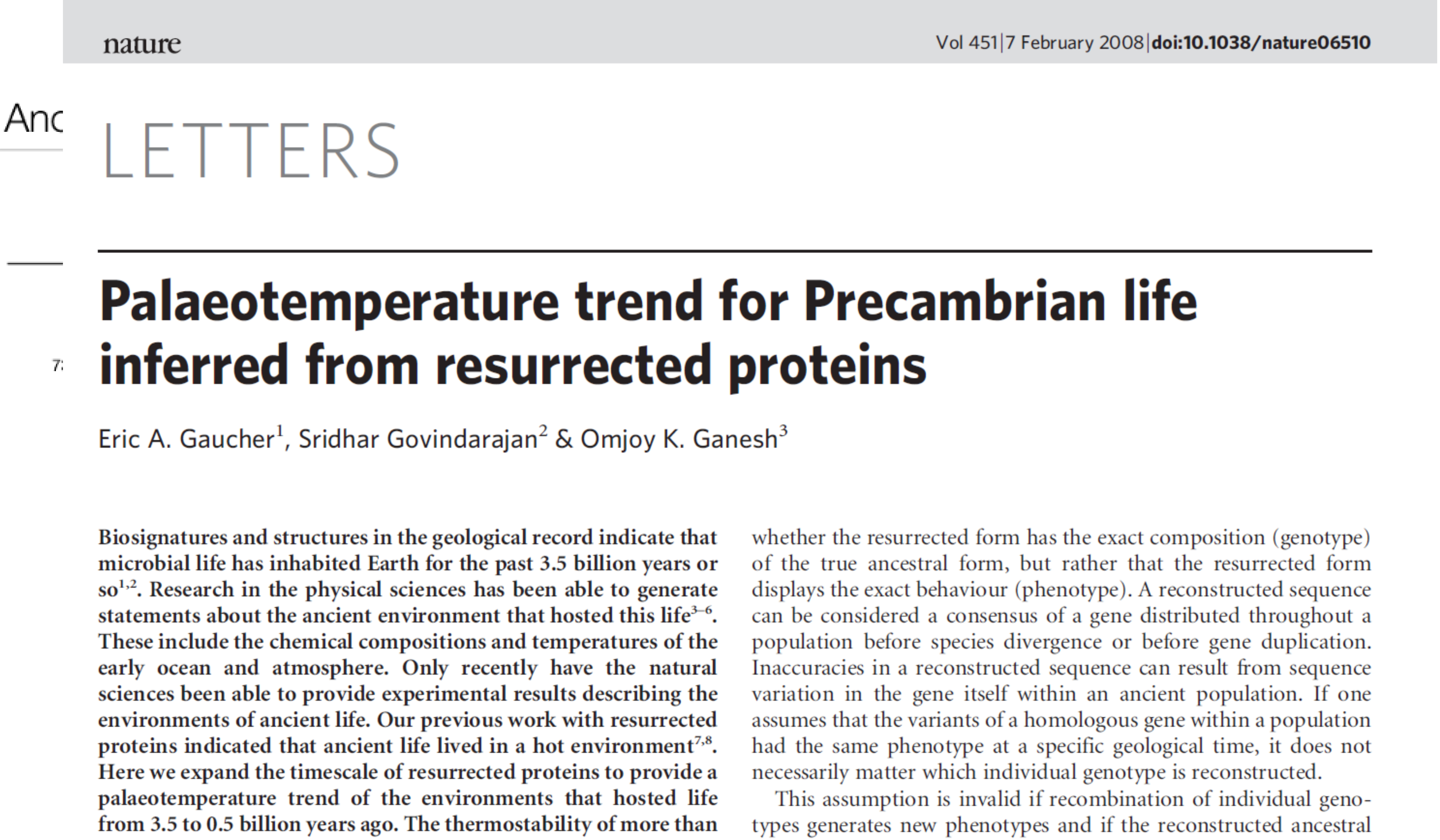
BILATERIAN
An animal that shows bilateral symmetry across a body axis. Bilaterians include chordates, arthropods, nematodes, annelids and molluscs, among other groups.

ORTHOLOGUES
The 'same' gene in more than one species. Orthologues descend from a speciation event.

The evolution of gene function is a central issue in molecular evolution: by what mechanisms and dynamics have the diverse functions of modern-day genes emerged? This question is usually addressed by inferring past processes from extant patterns, using statistical methods to detect the traces of ancient evolutionary events in the sequences of modern-day genes (see, for example, REFS 1,2). Thanks to the recent surge in knowledge of structure–function relationships, evolutionists can better interpret indicative patterns — such as biases or changes in evolutionary rates — by focusing on spe

genes from the last common ancestors of bacteria, of BILATERIAN animals and of vertebrates. These studies have shed light on fascinating questions about primordial environmental adaptations and the evolution of crucial gene functions.

Here, I review ancestral gene resurrection, the technical advances that have made it feasible and the studies that have applied it. I discuss the historical development of the technique and its methodological basis, with an emphasis on the previously unattainable insights it has allowed. I also highlight the limitations and pit



nature Vol 451|7 February 2008|doi:10.1038/nature06510

Anc

LETTERS

Palaeotemperature trend for Precambrian life inferred from resurrected proteins

Eric A. Gaucher[1], Sridhar Govindarajan[2] & Omjoy K. Ganesh[3]

Biosignatures and structures in the geological record indicate that microbial life has inhabited Earth for the past 3.5 billion years or so[1,2]. Research in the physical sciences has been able to generate statements about the ancient environment that hosted this life[3–6]. These include the chemical compositions and temperatures of the early ocean and atmosphere. Only recently have the natural sciences been able to provide experimental results describing the environments of ancient life. Our previous work with resurrected proteins indicated that ancient life lived in a hot environment[7,8]. Here we expand the timescale of resurrected proteins to provide a palaeotemperature trend of the environments that hosted life from 3.5 to 0.5 billion years ago. The thermostability of more than

whether the resurrected form has the exact composition (genotype) of the true ancestral form, but rather that the resurrected form displays the exact behaviour (phenotype). A reconstructed sequence can be considered a consensus of a gene distributed throughout a population before species divergence or before gene duplication. Inaccuracies in a reconstructed sequence can result from sequence variation in the gene itself within an ancient population. If one assumes that the variants of a homologous gene within a population had the same phenotype at a specific geological time, it does not necessarily matter which individual genotype is reconstructed.

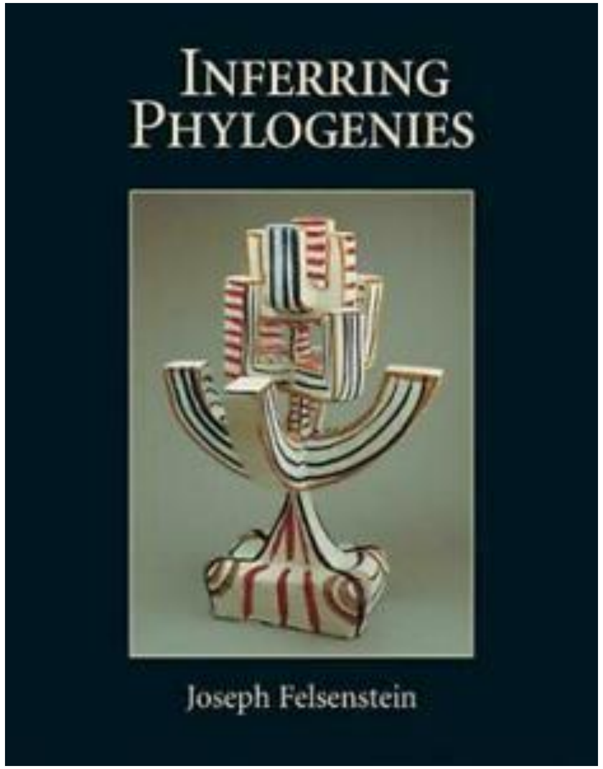
This assumption is invalid if recombination of individual genotypes generates new phenotypes and if the reconstructed ancestral

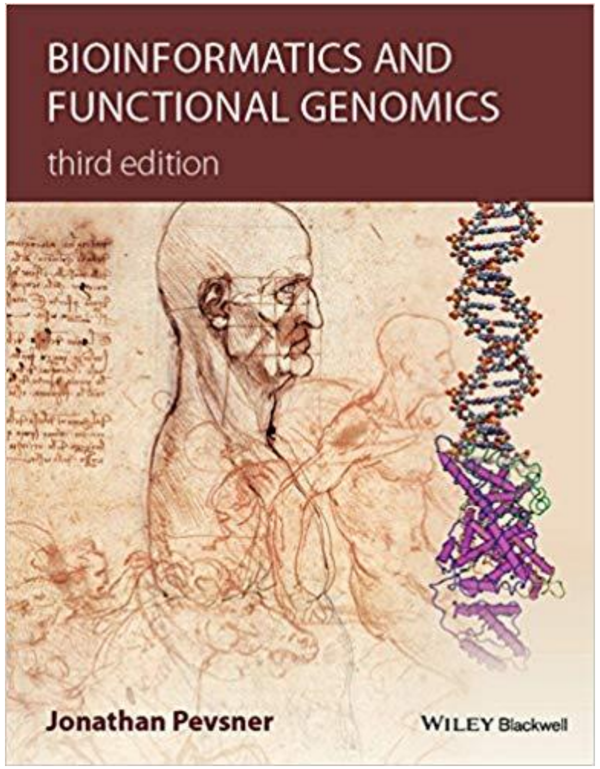
Palaeotemperature trend for Precambrian life inferred from resurrected proteins, Eric A. Gaucher, Sridhar Govindarajan & Omjoy K. Ganesh, Nature 451, 704-707, 2008

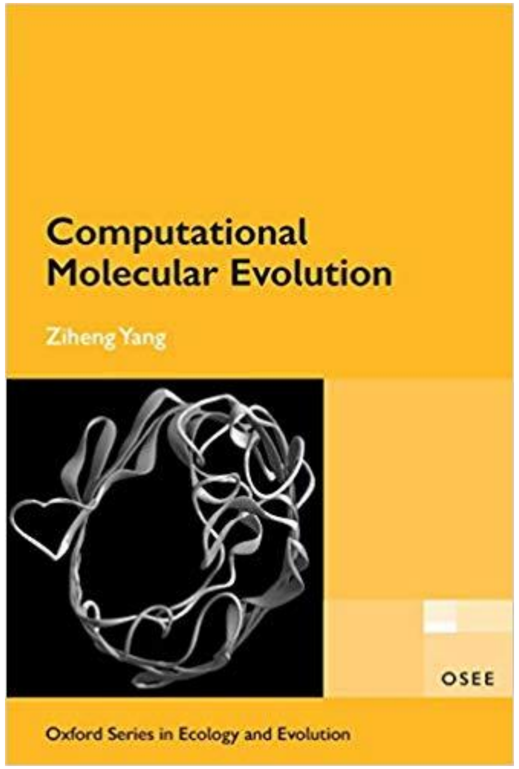
References

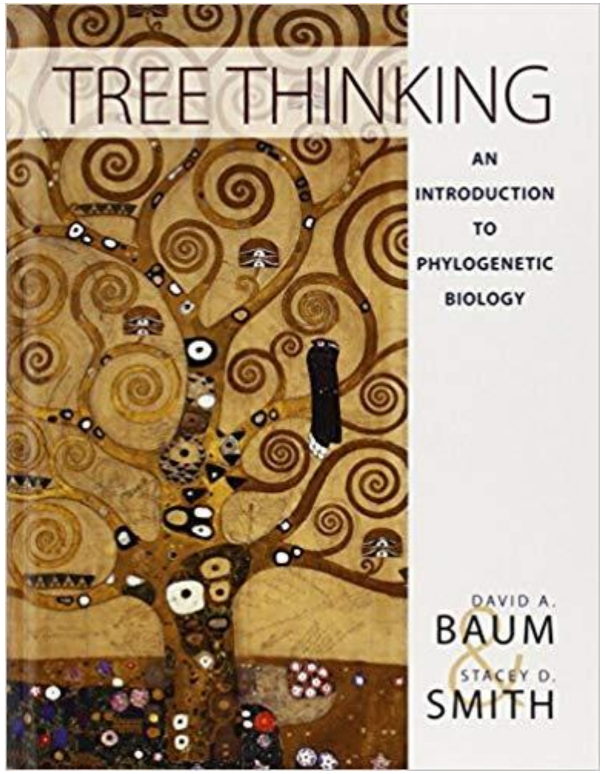
Lecture series of Dr. Anders Gorm Pedersen, Head of Section & Professor, Department of Health Technology, Technical University of Denmark, Kemitorvet on Molecular Evolution at Coursera

&






Computational Molecular Evolution
Ziheng Yang
OSEE
Oxford Series in Ecology and Evolution



That's all Folks!