Phylogeny reconstruction

Eva Stukenbrock & Julien Dutheil

eva.stukenbrock@mpi-marburg.mpg.de

julien.dutheil@mpi-marburg.mpg.de

Max Planck Institute for Terrestrial Microbiology – Marburg

19 Janvier 2011

Why studying phylogenetics?

Phylogenetics

The study of the evolution of (taxonomic) groups of organisms (*e.g.* species, populations).

Why studying phylogenetics?

Phylogenetics

The study of the evolution of (taxonomic) groups of organisms (*e.g.* species, populations).

- Infer the history of taxa
  - Taxonomy
  - Phylogeography
- Necessary to comparative analysis, as species are not “independent”

Phylogenetics applications

Phylogenetic tree

A graph depicting the ancestor-descendant relationships between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree connect the tips to their (unobservable) ancestral sequences (Holder 2003)

Phylogenetics applications

Phylogenetic tree

A graph depicting the ancestor-descendant relationships between organisms or gene sequences. The sequences are the tips of the tree. Branches of the tree connect the tips to their (unobservable) ancestral sequences (Holder 2003)

Some applications:

- Resolving orthology and paralogy
- Datation: estimate divergence times
- Reconstruct ancestral sequences
- Exhibit sites under positive selection
- Predict the structure of a molecule...

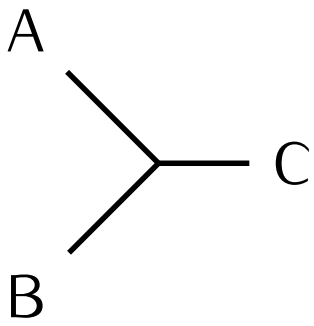
What is a tree?

2 leaves

What is a tree?

2 leaves

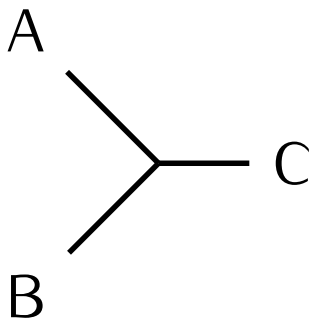
3 leaves



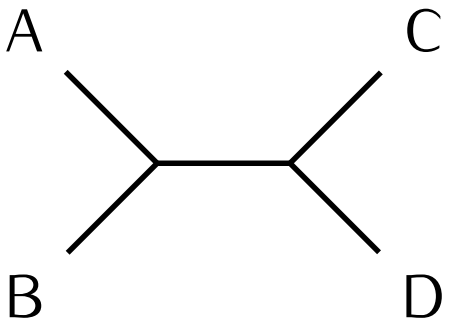
What is a tree?

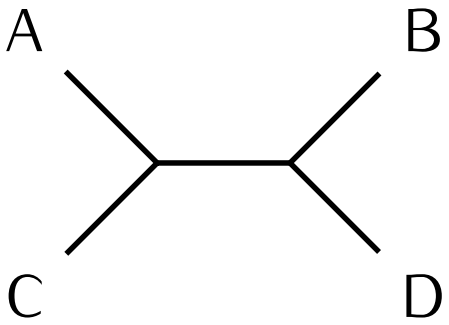
2 leaves

3 leaves



4 leaves

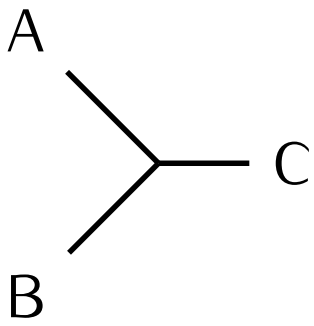




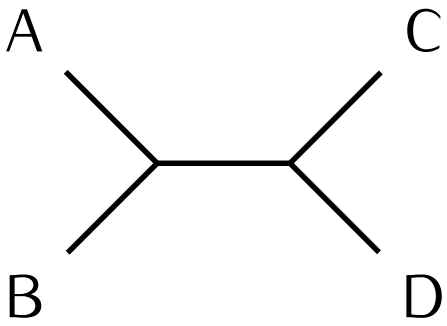
What is a tree?

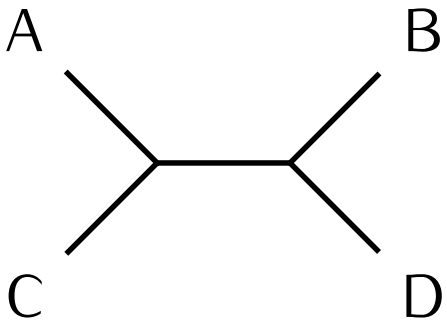
2 leaves

3 leaves



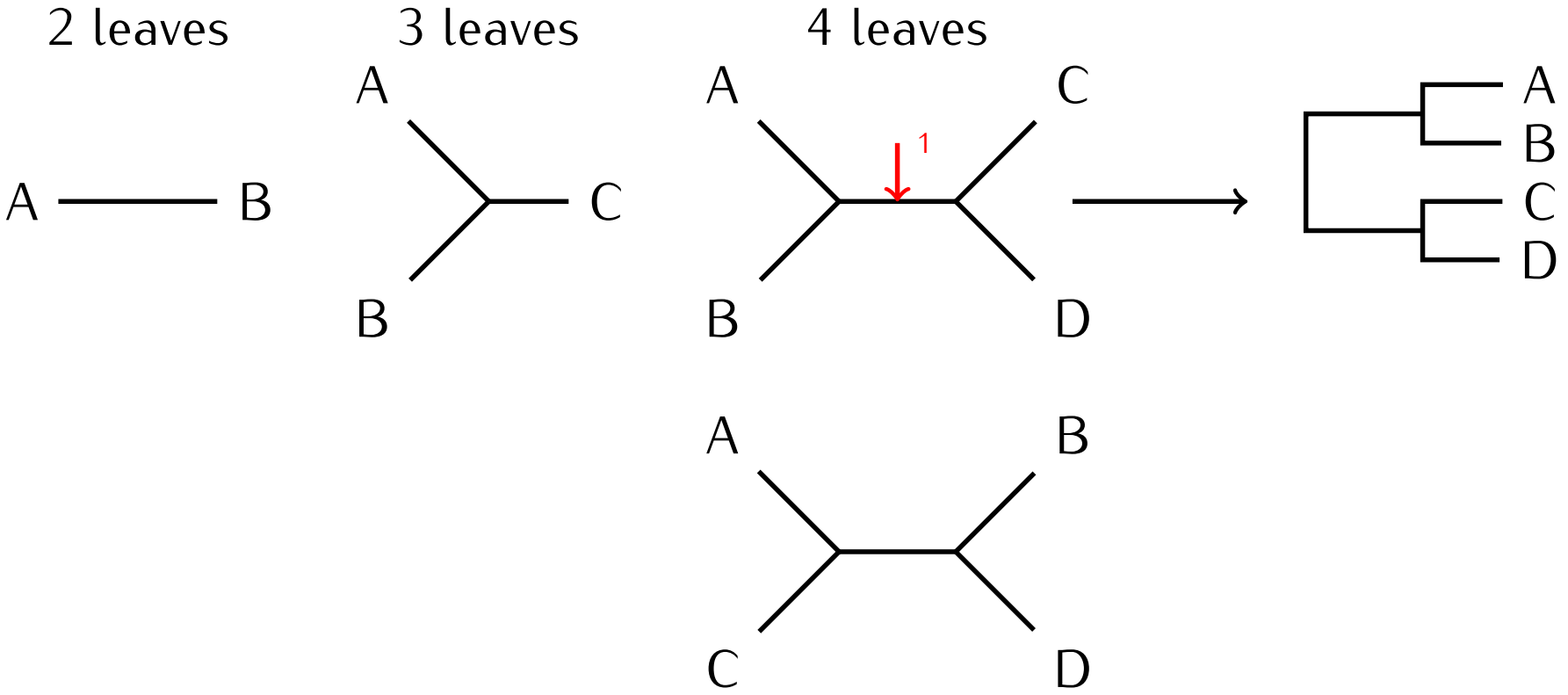
4 leaves





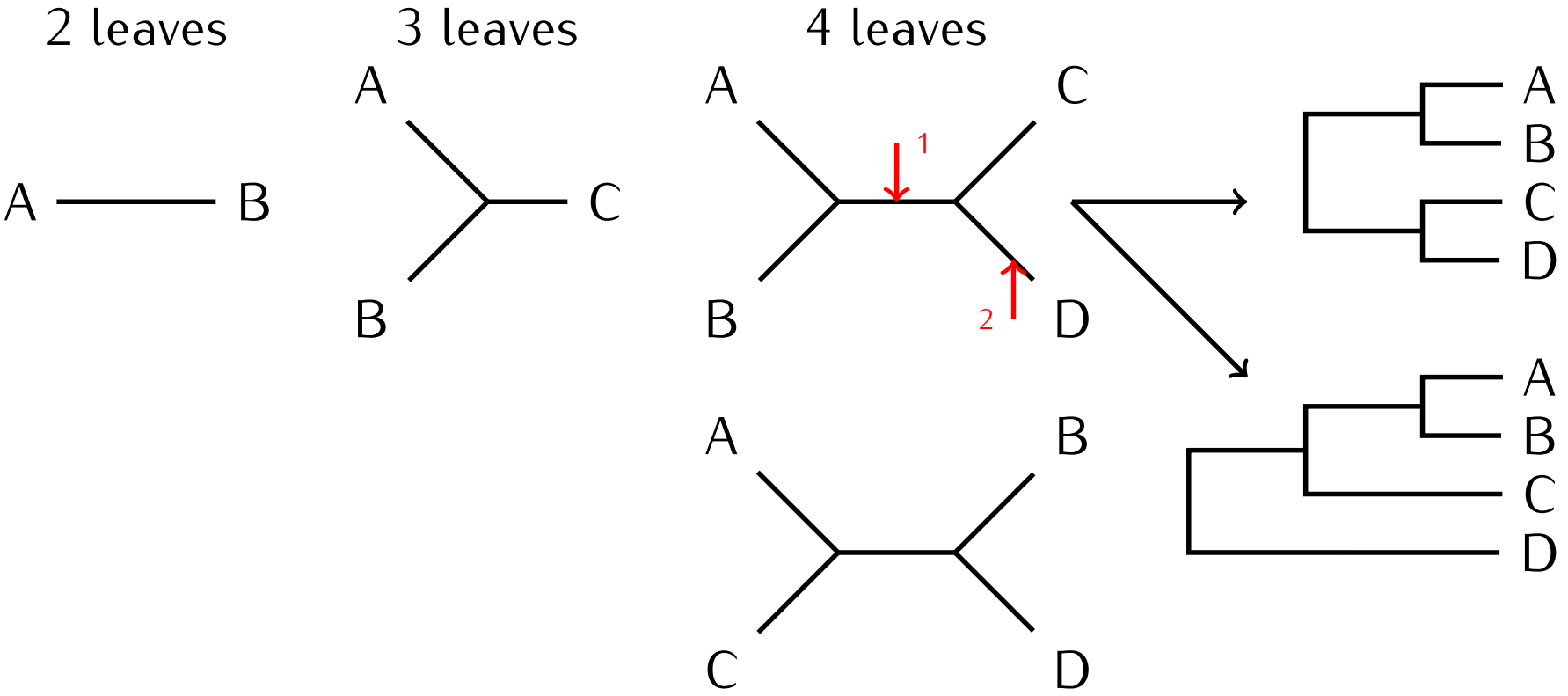
$$\frac{(2n-5)!}{2^{n-3}(n-2)!}$$
distinct unrooted trees

What is a tree?



$$\frac{(2n-5)!}{2^{n-3}(n-2)!}$$
distinct unrooted trees

What is a tree?

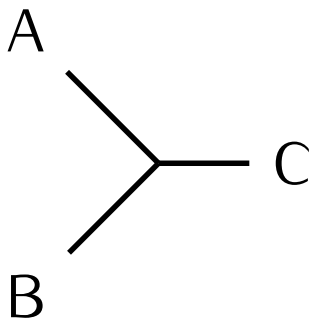


$$\frac{(2n-5)!}{2^{n-3}(n-2)!}$$
distinct unrooted trees

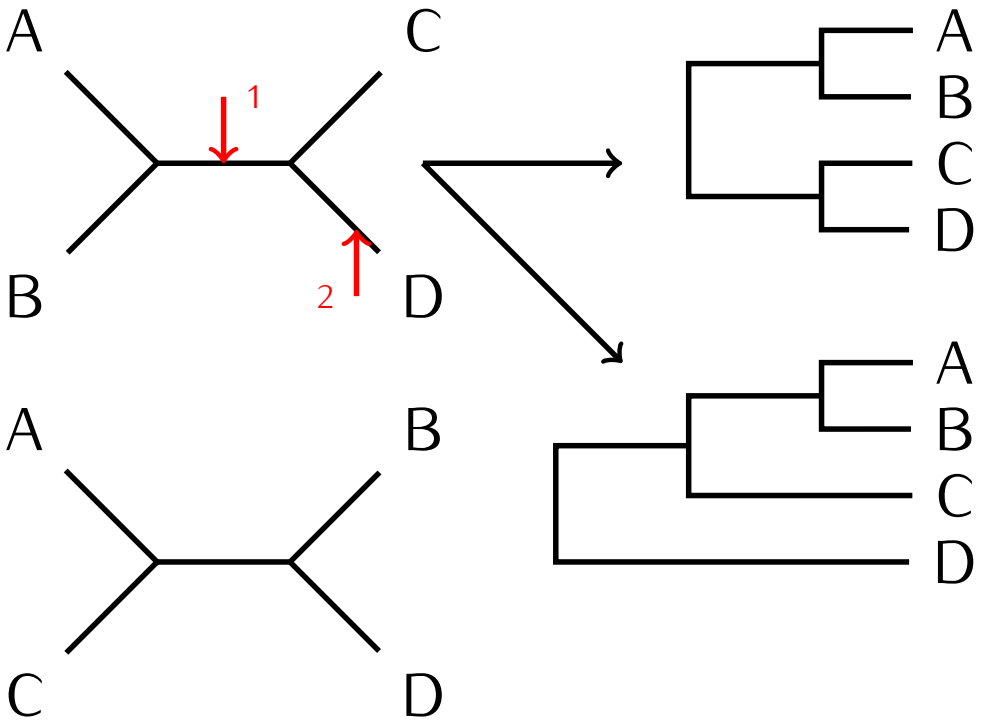
What is a tree?

2 leaves

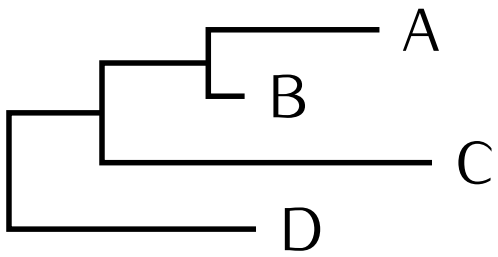
3 leaves



4 leaves

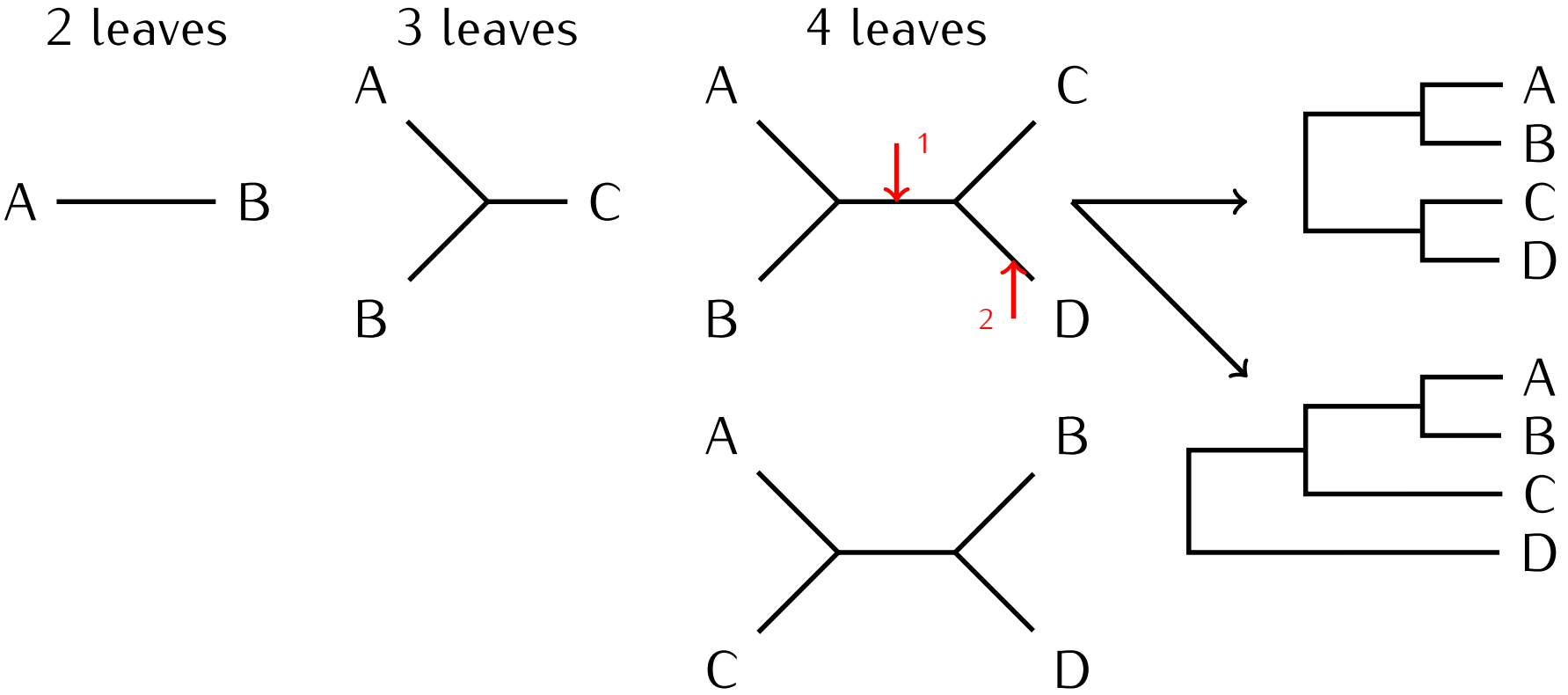


With branch lengths:

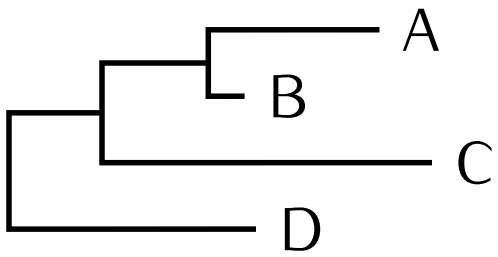


$$\frac{(2n-5)!}{2^{n-3}(n-2)!}$$
distinct unrooted trees

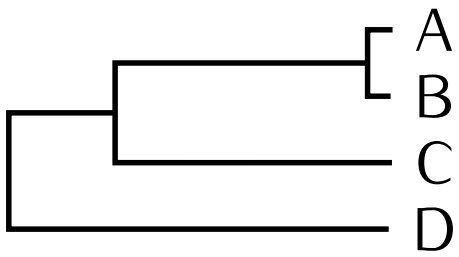
What is a tree?



With branch lengths:



With clock:



$$\frac{(2n-5)!}{2^{n-3}(n-2)!}$$
distinct unrooted trees

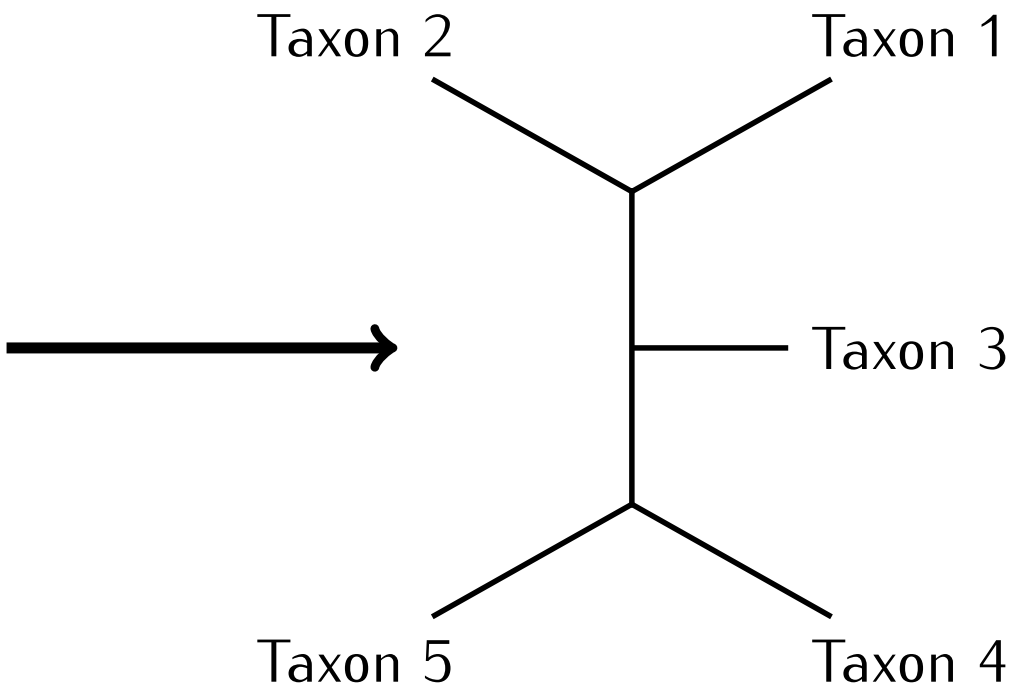
From sequences to trees

From sequences to trees

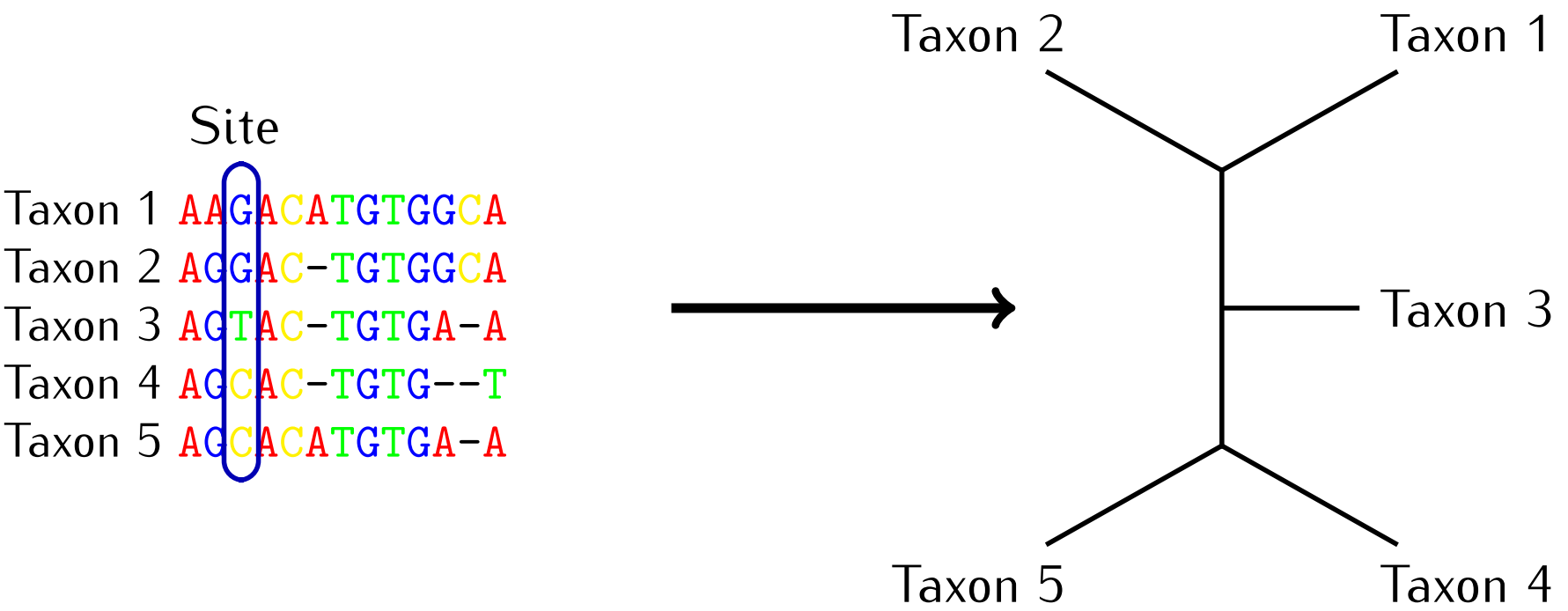
Taxon 1 AAGACATGTGGCA
Taxon 2 AGGAC-TGTGGCA
Taxon 3 AGTAC-TGTGA-A
Taxon 4 AGCAC-TGTG--T
Taxon 5 AGCACATGTGA-A

From sequences to trees

Taxon 1 AAGACATGTGGCA
Taxon 2 AGGAC-TGTGGCA
Taxon 3 AGTAC-TGTGA-A
Taxon 4 AGCAC-TGTG--T
Taxon 5 AGCACATGTGA-A



From sequences to trees



- Aligned homologous positions: sites
- Each site is a realization of a random variable

The phenetic approach

- Uses an overall similarity measure (commonly used in morphology)
- Consider that the similarity is (inversely) correlated to the evolutionary distance
- Build a tree from a pairwise distance matrix:

	T1	T2	T3	T4	T5
T1	0				
T2	2	0			
T3	3.5	4.5	0		
T4	8	10	8	0	
T5	6	8	9	3	0

The WPGMA clustering method

Weighted Pair Group Method using Average

1. Pick the smallest distance

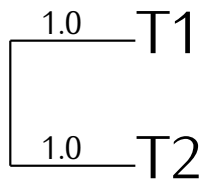
	T1	T2	T3	T4	T5
T1	0				
T2	2	0			
T3	3.5	4.5	0		
T4	8	10	8	0	
T5	6	8	9	3	0

The WPGMA clustering method

Weighted Pair Group Method using Average

2. Gather the corresponding groups

	T1	T2	T3	T4	T5
T1	0				
T2	2	0			
T3	3.5	4.5	0		
T4	8	10	8	0	
T5	6	8	9	3	0



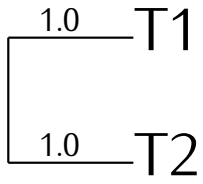
The WPGMA clustering method
Weighted Pair Group Method using Average

	T12	T3	T4	T5
T12	0			
T3	4	0		
T4	9	8	0	
T5	7	9	3	0

3. Recompute distances from the new group:

$$d_{ij,k} = \frac{d_{i,k} + d_{j,k}}{2}$$

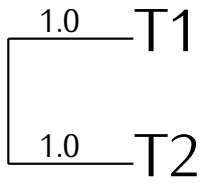
(Assumes that the rate of change is constant over time: molecular clock)



The WPGMA clustering method
Weighted Pair Group Method using Average

1. Pick the smallest distance

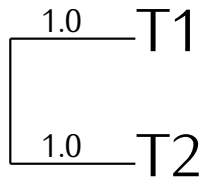
	T12	T3	**T4**	T5
T12	0			
T3	4	0		
T4	9	8	0	
T5	7	9	**3**	0

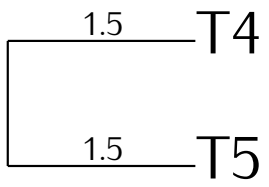


The WPGMA clustering method
Weighted Pair Group Method using Average

2. Gather the corresponding groups

	T12	T3	T4	T5
T12	0			
T3	4	0		
T4	9	8	0	
T5	7	9	3	0





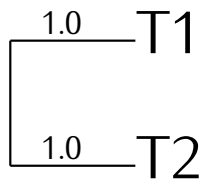
The WPGMA clustering method

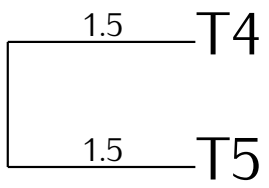
Weighted Pair Group Method using Average

	T12	T3	T45
T12	0		
T3	4	0	
T45	8	8.5	0

3. Recompute distances from the new group:

$$d_{ij,k} = \frac{d_{i,k} + d_{j,k}}{2}$$

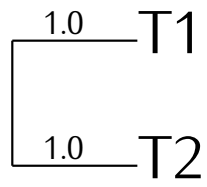


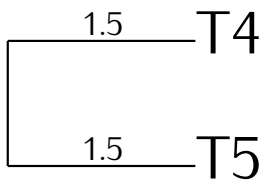


The WPGMA clustering method
Weighted Pair Group Method using Average

1. Pick the smallest distance

	T12	T3	T45
T12	0		
T3	4	0	
T45	8	8.5	0

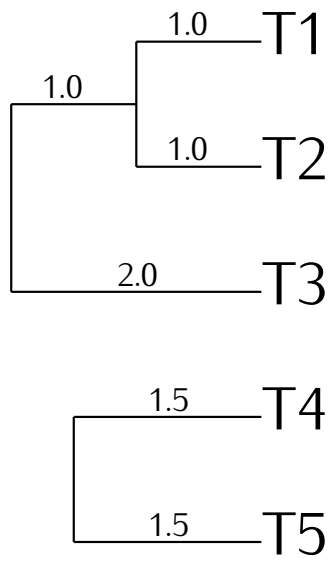




The WPGMA clustering method
Weighted Pair Group Method using Average

2. Gather the corresponding groups

	T12	T3	T45
T12	0		
T3	4	0	
T45	8	8.5	0

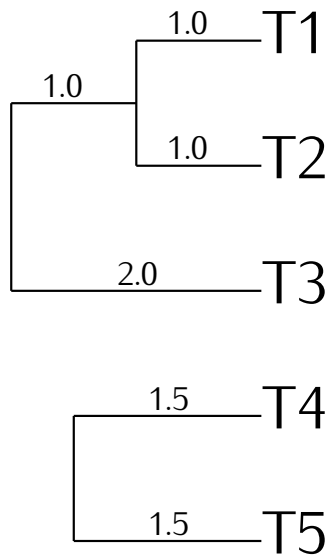


The WPGMA clustering method
Weighted Pair Group Method using Average

	T123	T45
T123	0	
T45	8.25	0

3. Recompute distances from the new group:

$$d_{ij,k} = \frac{d_{i,k} + d_{j,k}}{2}$$

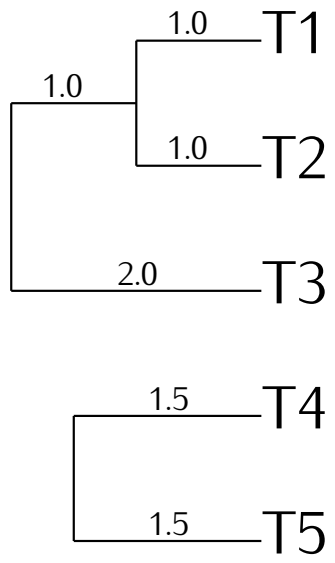


The WPGMA clustering method

Weighted Pair Group Method using Average

1. Pick the smallest distance

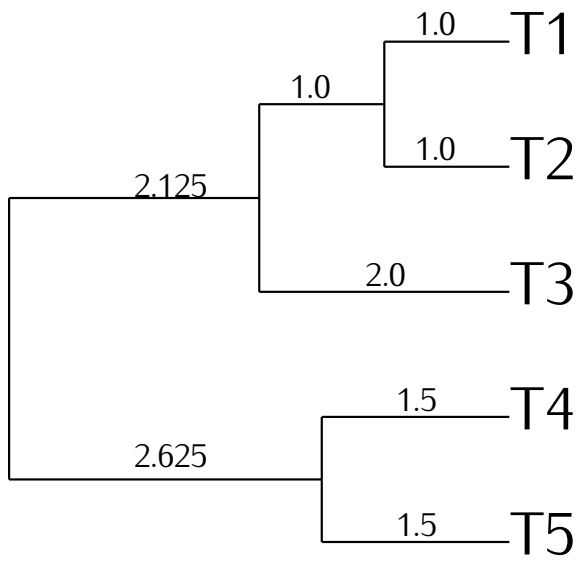
	T123	T45
T123	0	
T45	8.25	0



The WPGMA clustering method
Weighted Pair Group Method using Average

2. Gather the corresponding groups

	T123	T45
T123	0	
T45	8.25	0



Distance methods

- "Distance" methods are extensions of clustering techniques for the purpose of phylogenetics

Distance methods

- "Distance" methods are extensions of clustering techniques for the purpose of phylogenetics
- Two steps:
  1. getting a distance matrix
  2. building a tree from the matrix

Distance methods

- "Distance" methods are extensions of clustering techniques for the purpose of phylogenetics
- Two steps:
  1. getting a distance matrix
  2. building a tree from the matrix
- When using molecular sequences, the distance for a pair of species is the divergence estimated from the sequences. It can be
  - an edit distance (alignment score)
  - a proportion of mismatch
  - an estimated divergence from the observed matches (requires a model of evolution)

Additive distance matrix

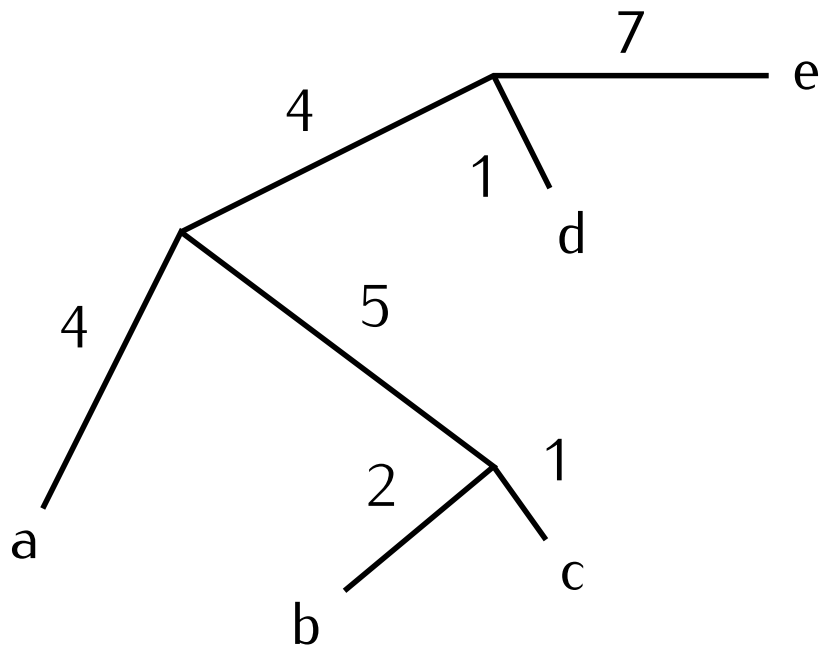
A distance matrix M is additive if it exists a corresponding phylogenetic tree T so that $d_{a,b} = d_{a,c} + d_{b,c}$, where c depicts the ancestral node for species a and b.

Additive distance matrix

A distance matrix M is additive if it exists a corresponding phylogenetic tree T so that $d_{a,b} = d_{a,c} + d_{b,c}$, where c depicts the ancestral node for species a and b.

M	a	b	c	d	e
a	0	11	10	9	15
b	11	0	3	12	18
c	10	3	0	11	17
d	9	12	11	0	8
e	15	18	17	8	0

$\Leftrightarrow$



Phylogenetic clustering
Biological distance matrices

- Distance matrices from real data are not additive

Phylogenetic clustering
Biological distance matrices

- Distance matrices from real data are not additive
- The goal is to find a tree for which the corresponding additive matrix is the closest as possible of the measured one (for instance based on the minimum least scares)

Phylogenetic clustering
Biological distance matrices

- Distance matrices from real data are not additive
- The goal is to find a tree for which the corresponding additive matrix is the closest as possible of the measured one (for instance based on the minimum least scares)
- This is a very difficult problem! Heuristics are needed

Phylogenetic clustering
Biological distance matrices

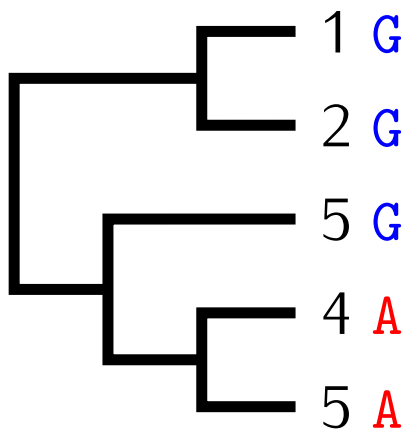
- Distance matrices from real data are not additive
- The goal is to find a tree for which the corresponding additive matrix is the closest as possible of the measured one (for instance based on the minimum least scares)
- This is a very difficult problem! Heuristics are needed
- The most famous heuristics:

Phylogenetic clustering
Biological distance matrices

- Distance matrices from real data are not additive
- The goal is to find a tree for which the corresponding additive matrix is the closest as possible of the measured one (for instance based on the minimum least scares)
- This is a very difficult problem! Heuristics are needed
- The most famous heuristics: Neighbor Joining • Saitou N and Nei M (1987). Molecular Biology and Evolution, 4:405-425 (one of the most cited biological paper!)
- Other heuristics: BioNJ, FastME

The cladistic approach

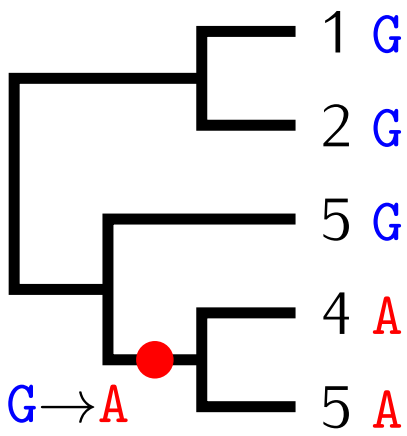
- Reconstruct the evolutionary history of the data



"good" site

The cladistic approach

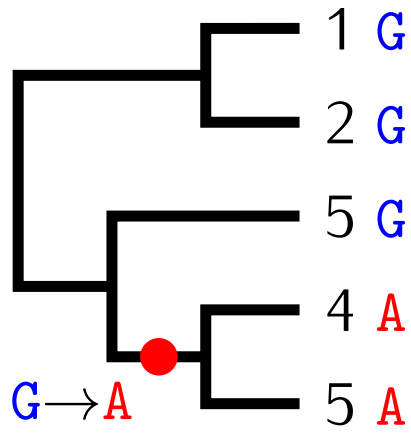
- Reconstruct the evolutionary history of the data
- A plethora of scenario, not all as likely



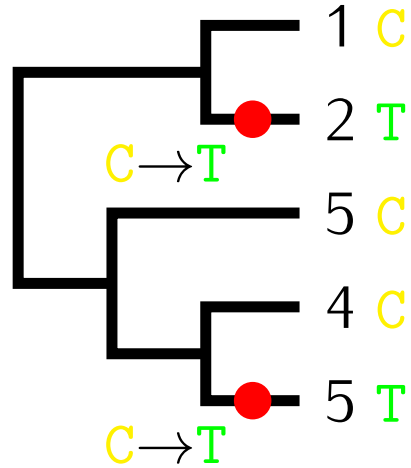
"good" site

The cladistic approach

- Reconstruct the evolutionary history of the data
- A plethora of scenario, not all as likely
- Assess the probability of a given scenario
  - for a site
  - for an alignment



"good" site



"bad" site

General hypotheses

1. All sites evolve essentially by substitutions (insertions, deletions, inversions are not accounted)
2. All sites evolve independently
3. All sites undergo the same process, notably
   - All sites evolve at the same rate
   - The substitution rate is constant over time

Maximum parsimony

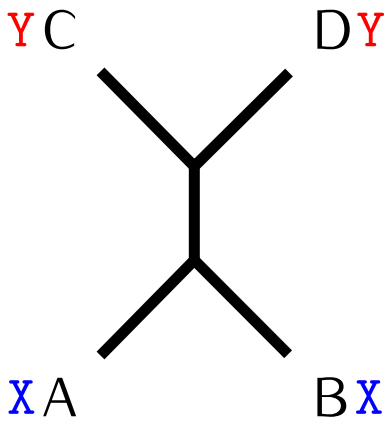
William of Ockham (1288–1346)

English medieval logician and Franciscan friar. Well known for its principle of parsimony, or Ockham's razor: the explanation of any phenomenon should make as few assumptions as possible, eliminating those that make no difference in the observable predictions of the explanatory hypothesis or theory.

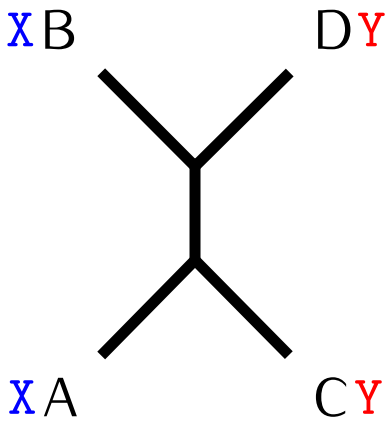
This is often paraphrased as "All other things being equal, the simplest solution is the best."

$\Rightarrow$ General statistic and scientific method.

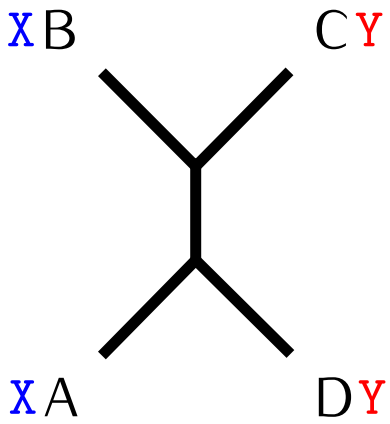
Maximum parsimony



Case 1



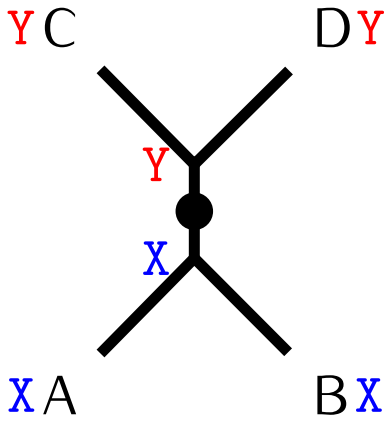
Case 2



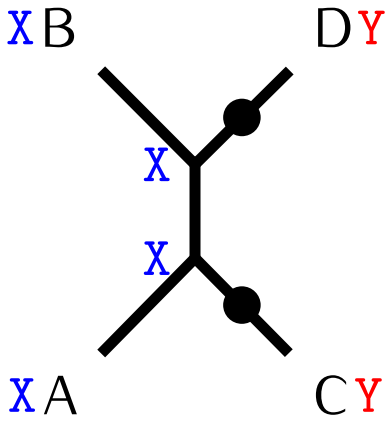
Case 3

- Three possible topologies

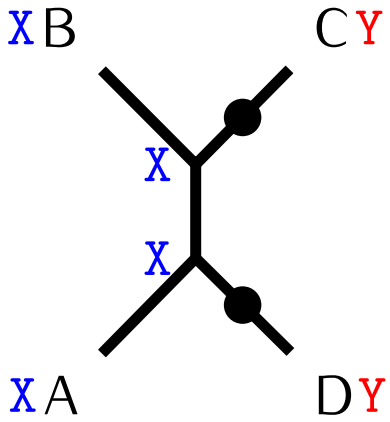
Maximum parsimony



Case 1



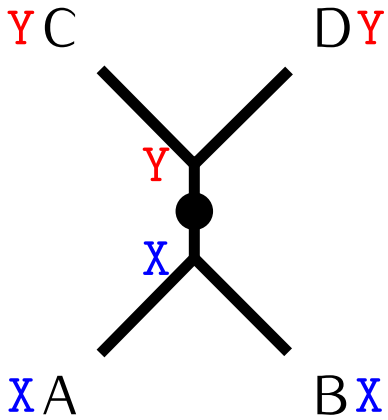
Case 2



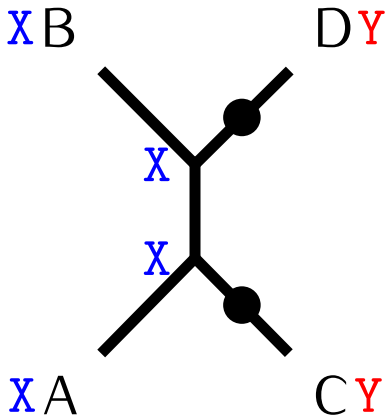
Case 3

- Three possible topologies

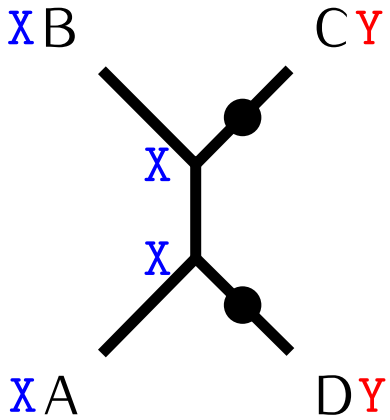
Maximum parsimony



Case 1



Case 2

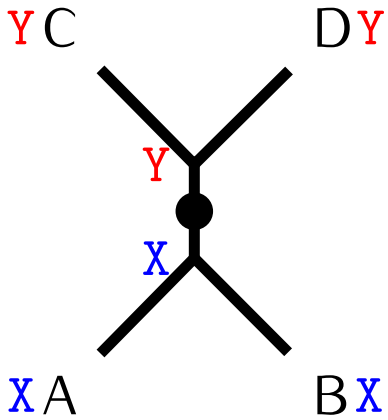


Case 3

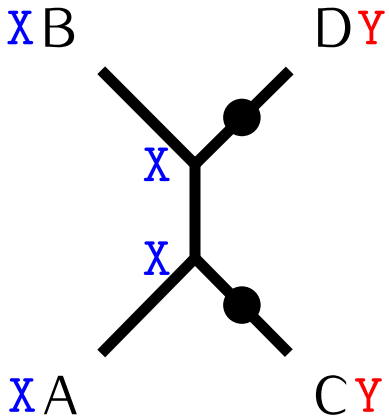
	Site 1	Site 2	Site 2		
A	X	X	X	X	X
B	X	Y	Y	X	X
C	Y	X	Y	X	Y
D	Y	Y	X	Y	Z
Case 1	1	2	2	1	2
Case 2	2	1	2	1	2
Case 3	2	2	1	1	2

- Three possible topologies
- Three types of informative sites + non-informative sites

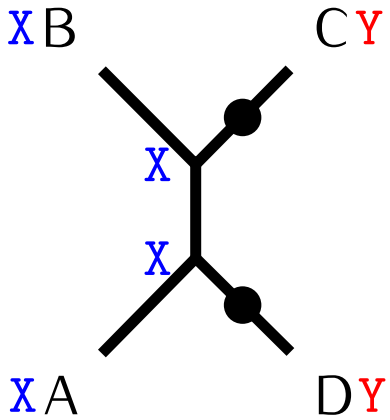
Maximum parsimony



Case 1



Case 2



Case 3

	Site 1	Site 2	Site 2		
A	X	X	X	X	X
B	X	Y	Y	X	X
C	Y	X	Y	X	Y
D	Y	Y	X	Y	Z
Case 1	1	2	2	1	2
Case 2	2	1	2	1	2
Case 3	2	2	1	1	2

- Three possible topologies
- Three types of informative sites + non-informative sites
- For one site, we take the most parsimonious scenario • For an alignment, we take the scenario in agreement with the majority of sites

How to find the best tree?

How to find the best tree?

1. Try all possible topologies (not possible if $n > 15$)

How to find the best tree?

1. Try all possible topologies (not possible if $n > 15$)
2. Stepwise addition:
   - Start with three random sequences
   - Pick a new sequence randomly, and find the best position (3 possibilities)
   - Pick a new sequence randomly, and find the best position (5 possibilities)
   - Pick a new sequence randomly, and find the best position (7 possibilities)
   - *etc.*

How to find the best tree?

1. Try all possible topologies (not possible if $n > 15$)
2. Stepwise addition:
   - Start with three random sequences
   - Pick a new sequence randomly, and find the best position (3 possibilities)
   - Pick a new sequence randomly, and find the best position (5 possibilities)
   - Pick a new sequence randomly, and find the best position (7 possibilities)
   - *etc.*
3. Start from an existing tree, and try to improve it

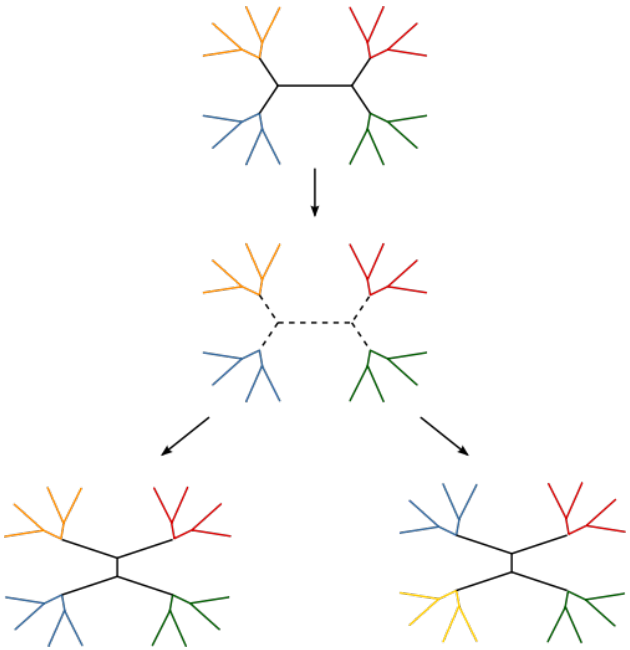
How to find the best tree?

1. Try all possible topologies (not possible if $n > 15$)
2. Stepwise addition:
   - Start with three random sequences
   - Pick a new sequence randomly, and find the best position (3 possibilities)
   - Pick a new sequence randomly, and find the best position (5 possibilities)
   - Pick a new sequence randomly, and find the best position (7 possibilities)
   - *etc.*
3. Start from an existing tree, and try to improve it
4. A combination of '2' and '3'

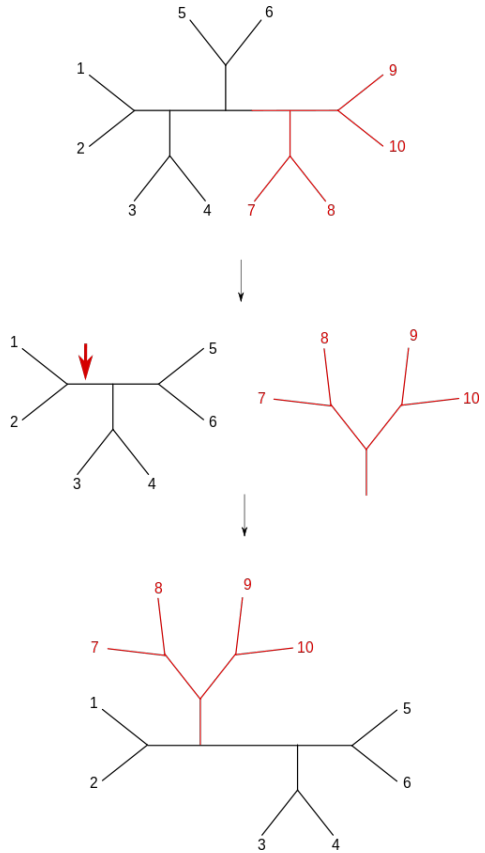
How to find the best tree?

1. Try all possible topologies (not possible if $n > 15$)
2. Stepwise addition:
   - Start with three random sequences
   - Pick a new sequence randomly, and find the best position (3 possibilities)
   - Pick a new sequence randomly, and find the best position (5 possibilities)
   - Pick a new sequence randomly, and find the best position (7 possibilities)
   - *etc.*
3. Start from an existing tree, and try to improve it
4. A combination of '2' and '3'
5. A combination of a distance method and '3'

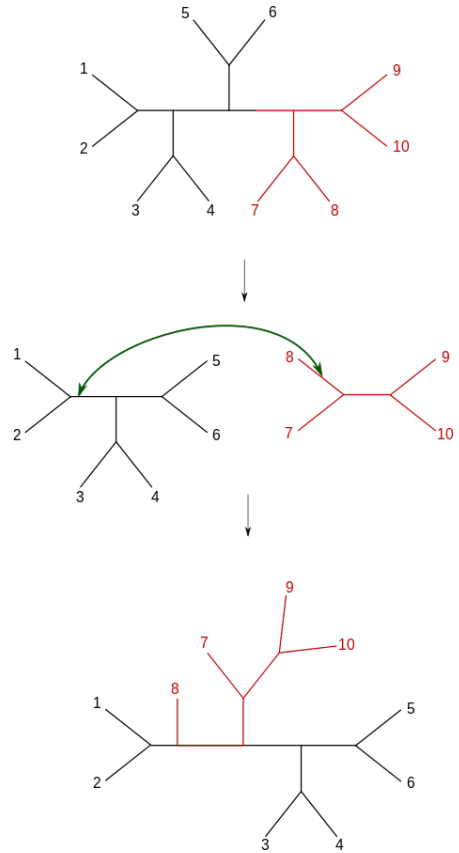
Topology ‘movements’



Nearest Neighbor Interchange (NNI)

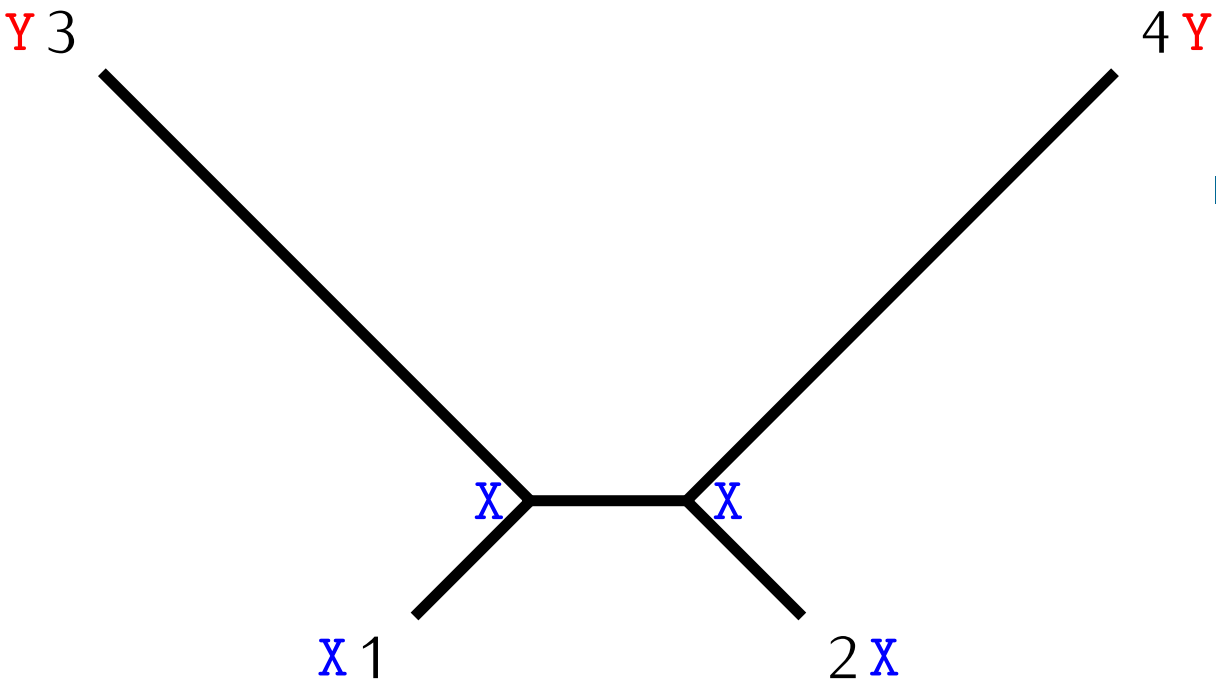


Subtree Pruning Regrafting (SPR)



Tree Bisection Reconnection (TBR)

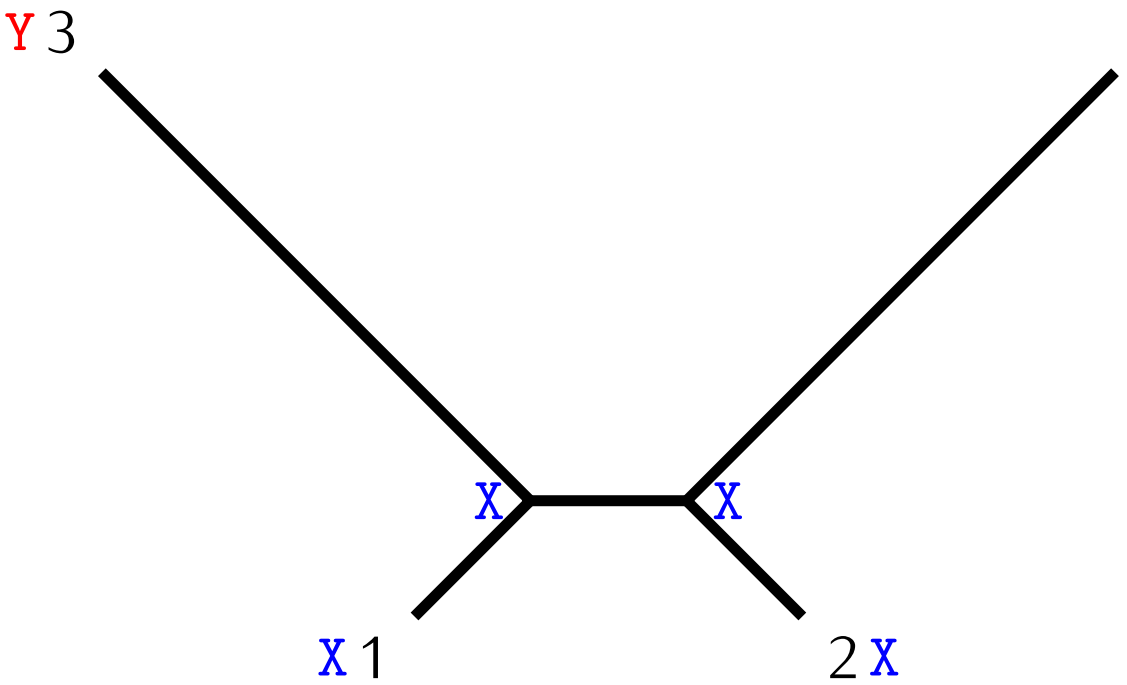
Problems with the parsimony approach



	a	b	c
1	X	X	X
2	X	Y	Y
3	Y	X	Y
4	Y	Y	X

- The 'right' site is $b : ((1,3),(2,4))$, but sites of type a are more frequent

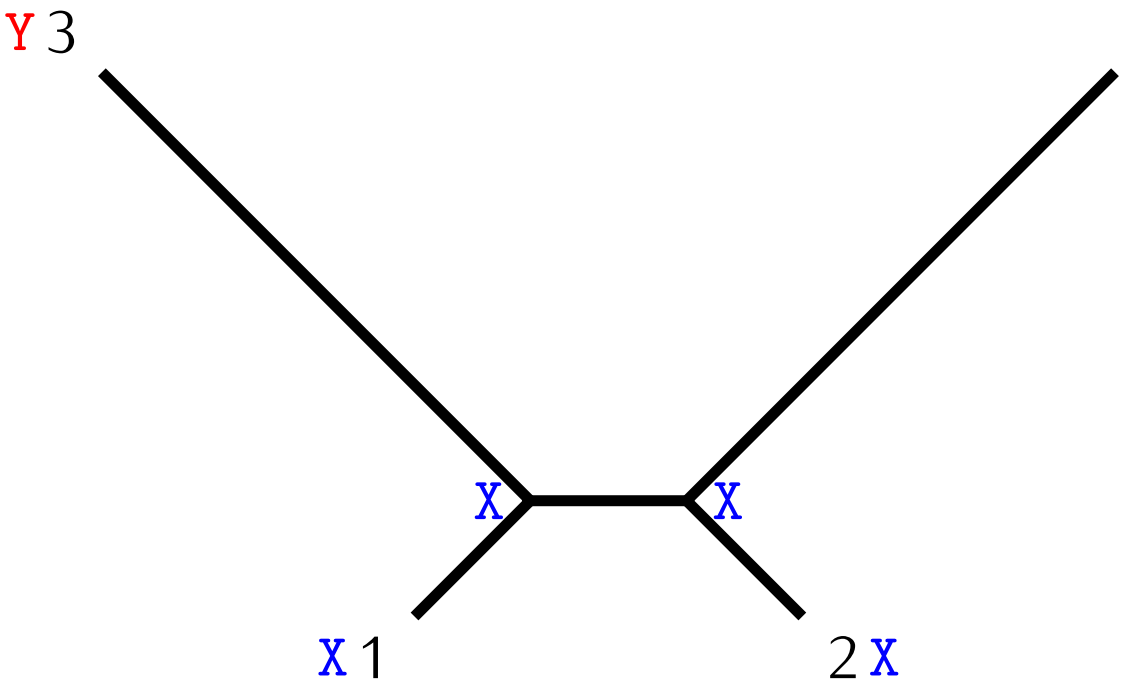
Problems with the parsimony approach



	a	b	c
1	X	X	X
2	X	Y	Y
3	Y	X	Y
4	Y	Y	X

- The 'right' site is $b : ((1,3),(2,4))$, but sites of type a are more frequent
- According to the maximum parsimony criterion, the a tree is the correct one

Problems with the parsimony approach



	a	b	c	d	e	f
1	X	X	X	X	X	X
2	X	Y	Y	X	X	X
3	Y	X	Y	Y	X	Y
4	Y	Y	X	Z	Y	X

- The 'right' site is $b : ((1,3),(2,4))$, but sites of type a are more frequent
- According to the maximum parsimony criterion, the a tree is the correct one
- Using non-informative sites can help resolving the issue

Long-branch attraction

Is the guinea-pig a rodent?

Dan Graur*†, Winston A. Hide†‡ & Wen-Hsiung Li†§

* Department of Zoology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Ramat Aviv 69978, Israel
† Center for Demographic and Population Genetics, University of Texas, PO Box 20334, Houston, Texas 77225, USA
‡ Department of Cell Biology, Baylor College of Medicine, Houston, Texas 77030, USA

THE guinea-pig (*Cavia porcellus*), traditionally classified as a New World hystricomorph rodent, often shows anomalous morphological and molecular features in comparison with other eutherian mammals[1–14]. For example, its insulin differs from that of other mammals in anabolic and growth-promoting activities and in its capability to form hexamers[5,6]. Indeed, the literature about the molecular evolution of guinea-pigs abounds in references to 'convergent evolution', 'extremely rapid rates of substitution', and 'unique evolutionary mechanisms'. These claims are based on the assumption that the guinea-pig is a rodent. Our phylogenetic analyses of amino-acid sequence data, however, imply that the guinea-pig diverged before the separation of the primates and the artiodactyls from the myomorph rodents (rats and mice). If true, then the myomorphs and the caviomorphs do not constitute a natural clade, and the Caviomorpha (or the Histricomorpha) should be elevated in taxonomical rank and regarded as a separate mammalian order distinct from the Rodentia. If, as suggested by recent data[15,16], the myomorphs branched off before the divergence among the carnivores, lagomorphs, artiodactyls and primates, then the new order would represent an early divergence in eutherian

Markov model

- We model the evolution of each position (site) of a sequence independently

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\Pr(X(t) = A) \quad = \quad \Pr(X(t_0) = A) \times \Pr(A \to A)$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned} \Pr(X(t) = A) &= \Pr(X(t_0) = A) \times \Pr(A \to A) \\ &+ \Pr(X(t_0) = C) \times \Pr(C \to A) \end{aligned}$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned}\Pr(X(t) = A) &= \Pr(X(t_0) = A) \times \Pr(A \to A) \\ &+ \Pr(X(t_0) = C) \times \Pr(C \to A) \\ &+ \Pr(X(t_0) = G) \times \Pr(G \to A)\end{aligned}$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned} \Pr(X(t) = A) &= \Pr(X(t_0) = A) \times \Pr(A \to A) \\ &+ \Pr(X(t_0) = C) \times \Pr(C \to A) \\ &+ \Pr(X(t_0) = G) \times \Pr(G \to A) \\ &+ \Pr(X(t_0) = T) \times \Pr(T \to A) \end{aligned} \quad (1)$$

Markov model

- We model the evolution of each position (site) of a sequence independently
- The state $X(t)$ of a site at time t depends only on the current state:

$$\begin{aligned} \Pr(X(t) = A) &= \Pr(X(t_0) = A) \times \Pr(A \to A) \\ &+ \Pr(X(t_0) = C) \times \Pr(C \to A) \\ &+ \Pr(X(t_0) = G) \times \Pr(G \to A) \\ &+ \Pr(X(t_0) = T) \times \Pr(T \to A) \end{aligned} \quad (1)$$

- Similar equations are writable for $\Pr(X(t) = C)$, $\Pr(X(t) = G)$ and $\Pr(X(t) = T)$.

Matrix notation

- We can gather all equations in a more compact form. We note

$$x(t) = \begin{pmatrix} X(t) = A & X(t) = C & X(t) = G & X(t) = T \end{pmatrix}$$

Matrix notation

- We can gather all equations in a more compact form. We note

$$x(t) = \begin{pmatrix} X(t) = A & X(t) = C & X(t) = G & X(t) = T \end{pmatrix}$$

- And we can write

$$x(t) = x(t_0) \times \underbrace{\begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}}_{P}$$

where $p_{ij} = \Pr(i \to j)$.

Matrix notation

- We can gather all equations in a more compact form. We note

$$x(t) = \begin{pmatrix} X(t) = A & X(t) = C & X(t) = G & X(t) = T \end{pmatrix}$$

- And we can write

$$x(t) = x(t_0) \times \underbrace{\begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}}_{P}$$

  where $p_{ij} = \Pr(i \to j)$.

- 'P' defines the substitution process.

A few more considerations

- We have

$$\forall i, \sum_j p_{i,j} = 1$$

that is

$$\Pr(A \to A) + \Pr(A \to C) + \Pr(A \to G) + \Pr(A \to T) = 1$$

- If we assume that all types of mutations are equi-probable (Jukes and Cantor, 1969), we can simplify:

$$P_{(JC69)} = \begin{pmatrix} 1-3r & r & r & r \\ r & 1-3r & r & r \\ r & r & 1-3r & r \\ r & r & r & 1-3r \end{pmatrix}$$

Continuous time

We assume that the process does not change over time, so we can write the equations for any time t:

$$t = t_0 + dt_0, \quad r = \alpha \cdot dt_0$$

$$x(t_0 + dt_0) = x(t_0) \times \begin{pmatrix} 1 - 3\alpha dt_0 & \alpha dt_0 & \alpha dt_0 & \alpha dt_0 \\ \alpha dt_0 & 1 - 3\alpha dt_0 & \alpha dt_0 & \alpha dt_0 \\ \alpha dt_0 & \alpha dt_0 & 1 - 3\alpha dt_0 & \alpha dt_0 \\ \alpha dt_0 & \alpha dt_0 & \alpha dt & 1 - 3\alpha dt_0 \end{pmatrix}$$

$$x(t_0 + dt_0) = x(t_0) + x(t_0) \cdot Q dt_0$$

$$\frac{x(t_0 + dt_0) - x(t_0)}{dt_0} = x(t_0) \cdot Q$$

Continuous time

- We obtain a differential equation by having $dt_0 \to 0$:

$$\frac{\partial x(t)}{\partial t} = Q \cdot x(t)$$

Continuous time

- We obtain a differential equation by having $dt_0 \to 0$:

$$\frac{\partial x(t)}{\partial t} = Q \cdot x(t)$$

- Q is called the generator of the substitution process, and we have

$$Q_{(JC69)} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

with

$$\forall i, \sum_j q_{i,j} = 0$$

Continuous time

- We obtain a differential equation by having $dt_0 \to 0$:

$$\frac{\partial x(t)}{\partial t} = Q \cdot x(t)$$

- Q is called the generator of the substitution process, and we have

$$Q_{(JC69)} = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}$$

with

$$\forall i, \sum_j q_{i,j} = 0$$

- This resolves into

$$x(t) = x(t_0) \cdot \exp(Q \cdot t)$$

Conclusion

We can compute the probability that a certain sequence ($x(t_0)$) transforms into another given sequence ($x(t)$) after a known time (t) and given a certain substitution process specified by its generator (Q).

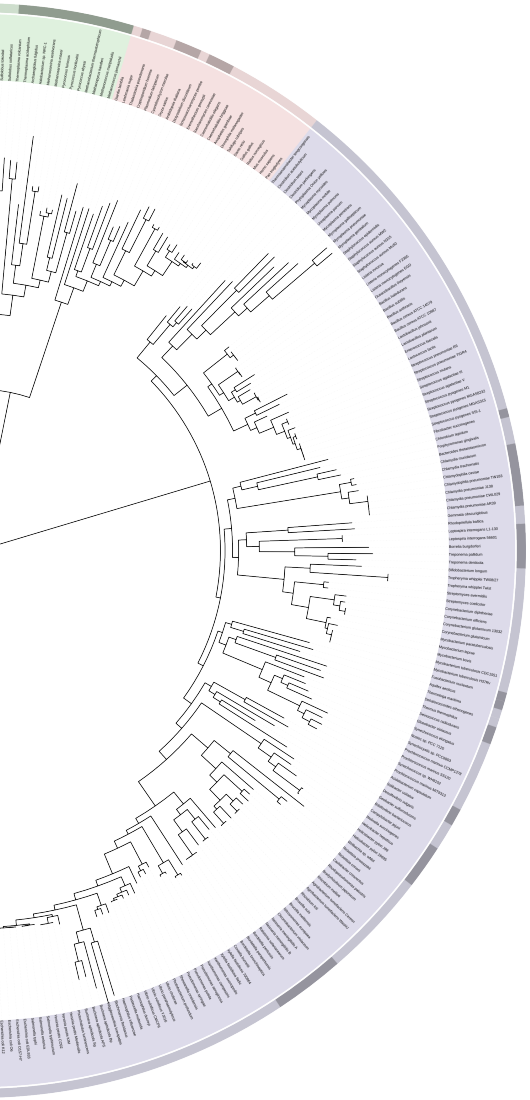
Conclusion

We can compute the probability that a certain sequence ($x(t_0)$) transforms into another given sequence ($x(t)$) after a known time (t) and given a certain substitution process specified by its generator (Q).

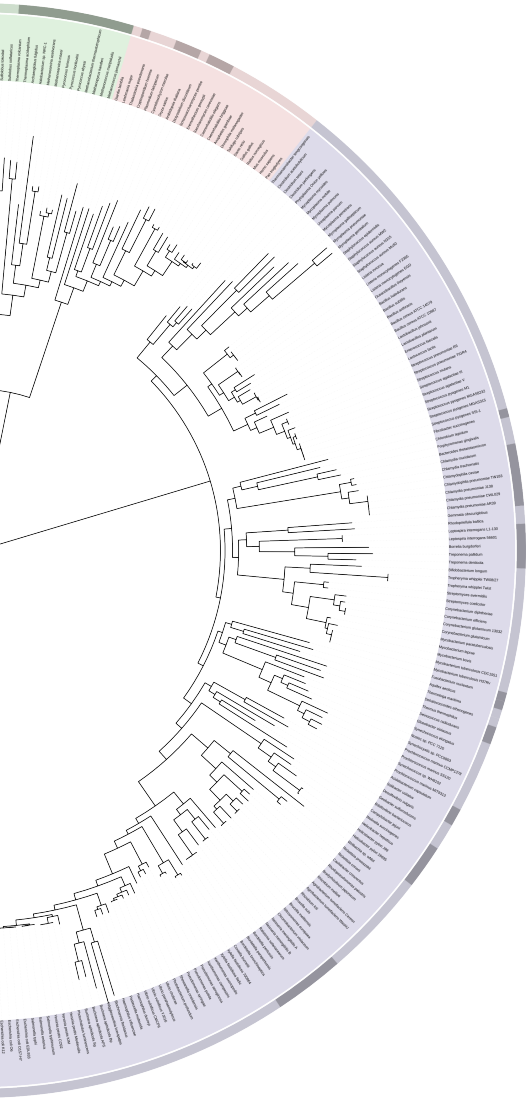
So what???

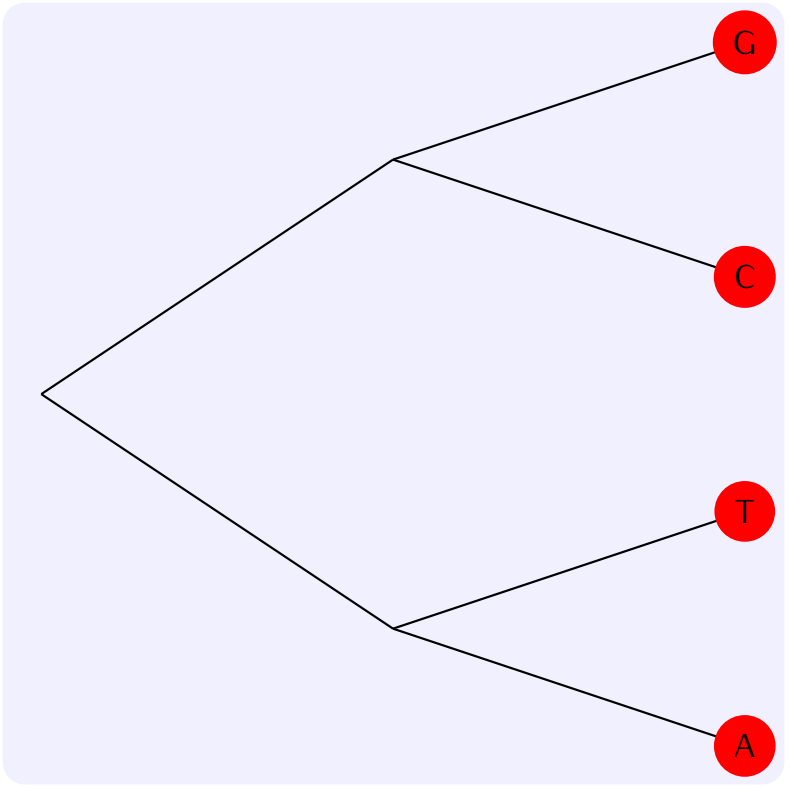
If we have two sequences and Q, we can compute t which maximizes this probability $\rightarrow$ unbiased estimate of the divergence between the two sequences!

Evolution along a tree

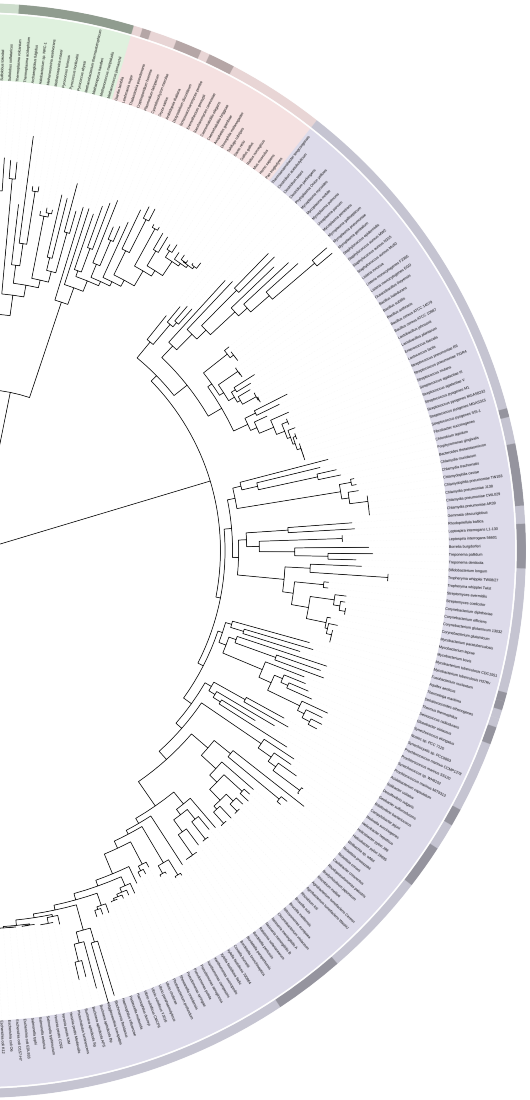


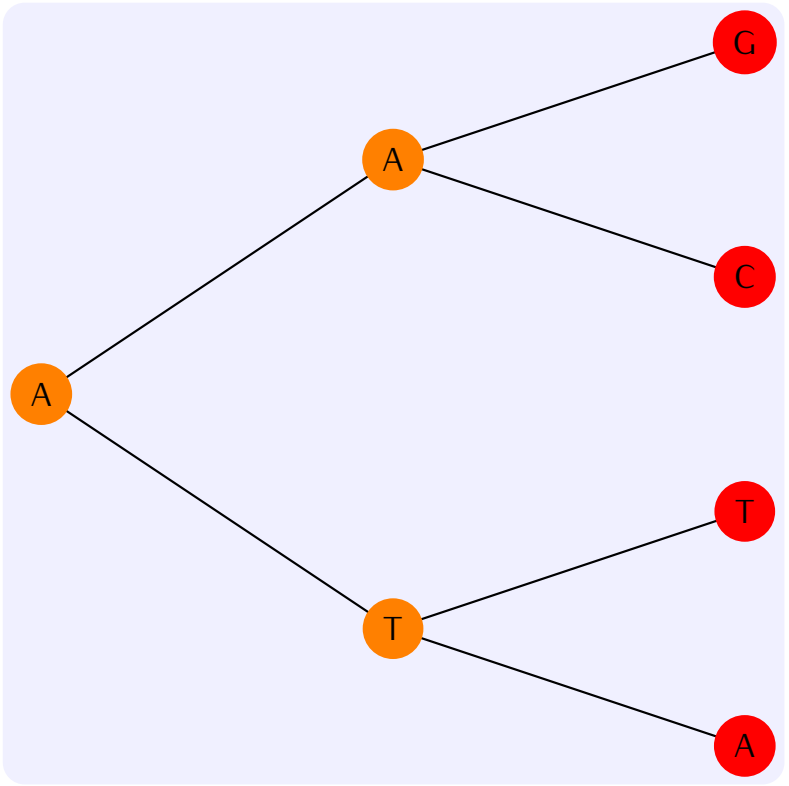
Evolution along a tree



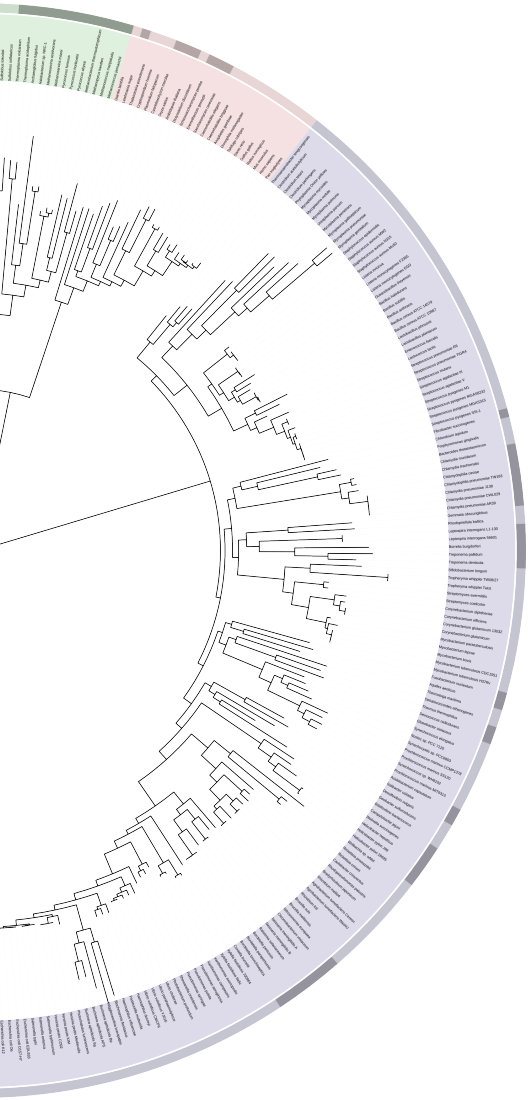


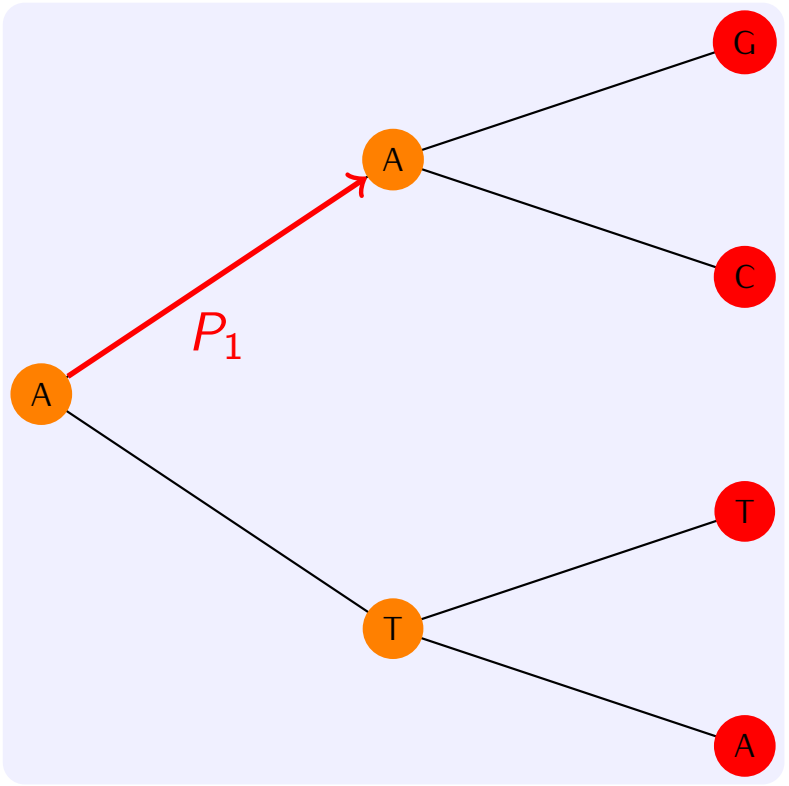
Evolution along a tree



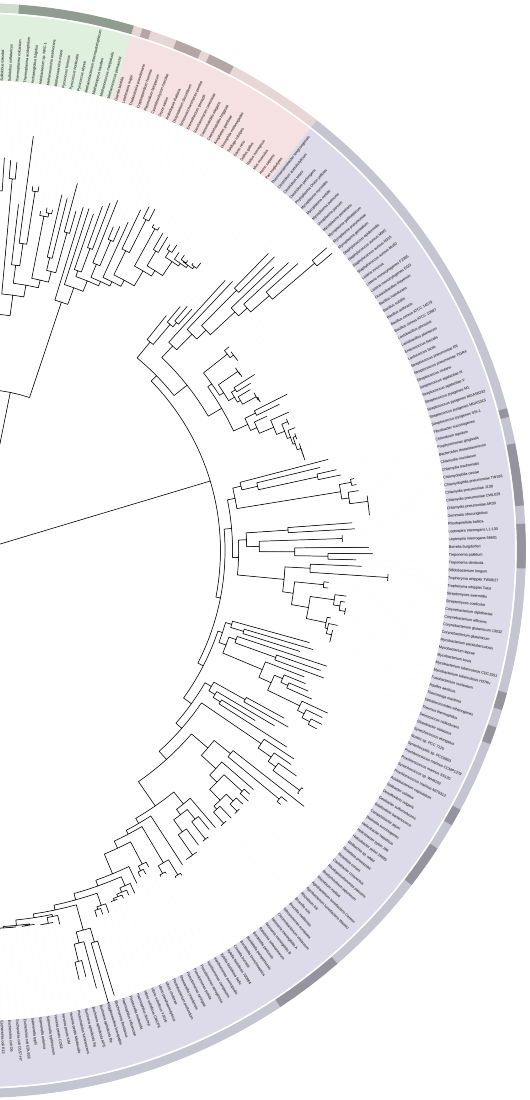


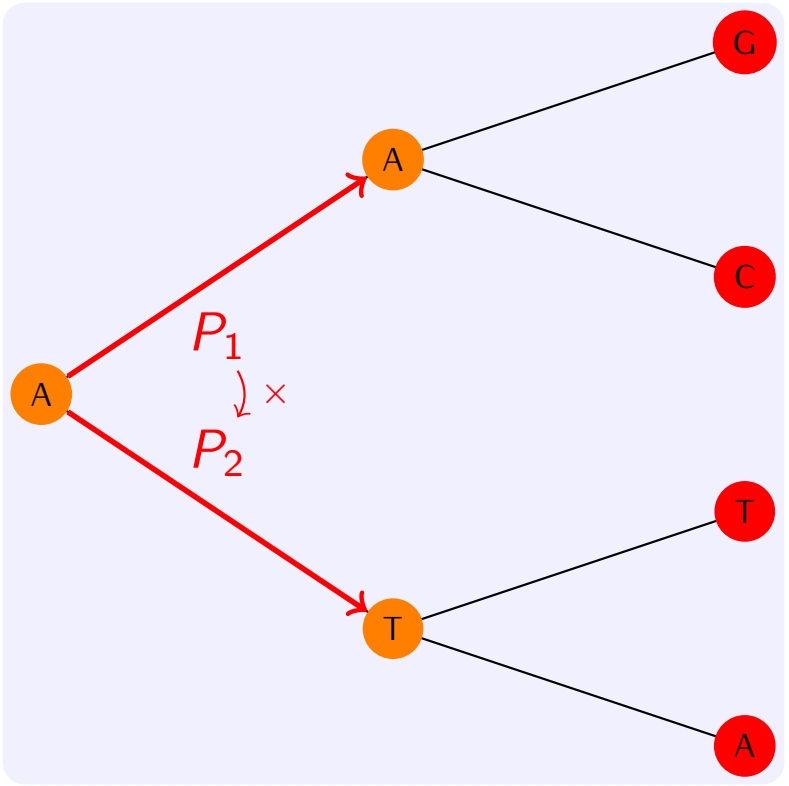
Evolution along a tree



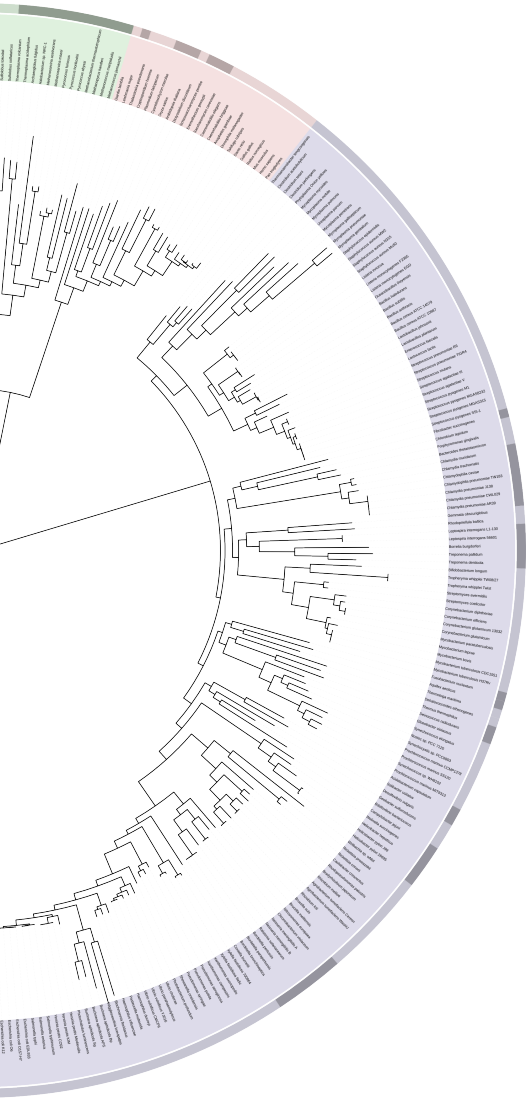


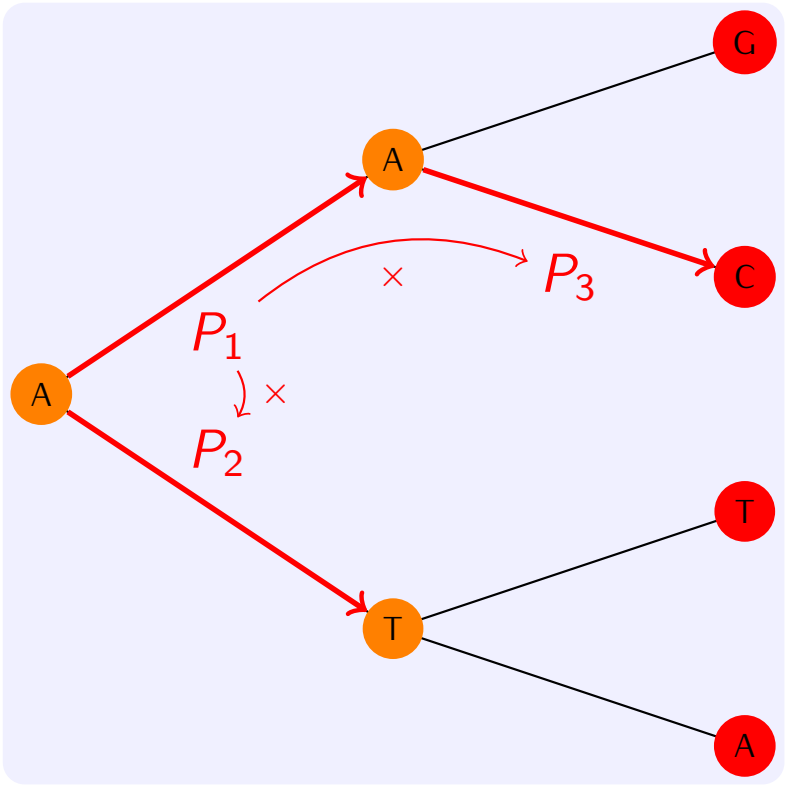
Evolution along a tree



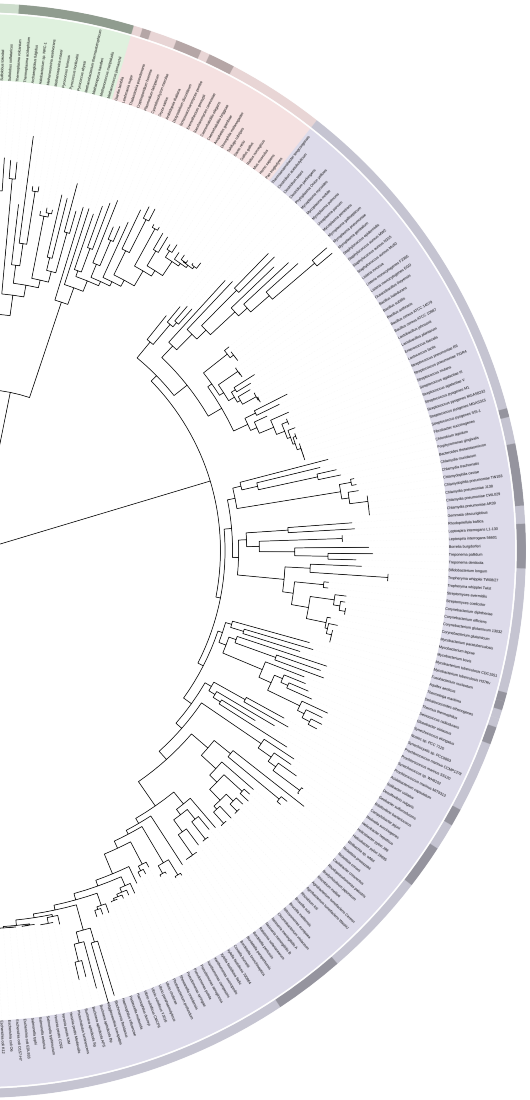


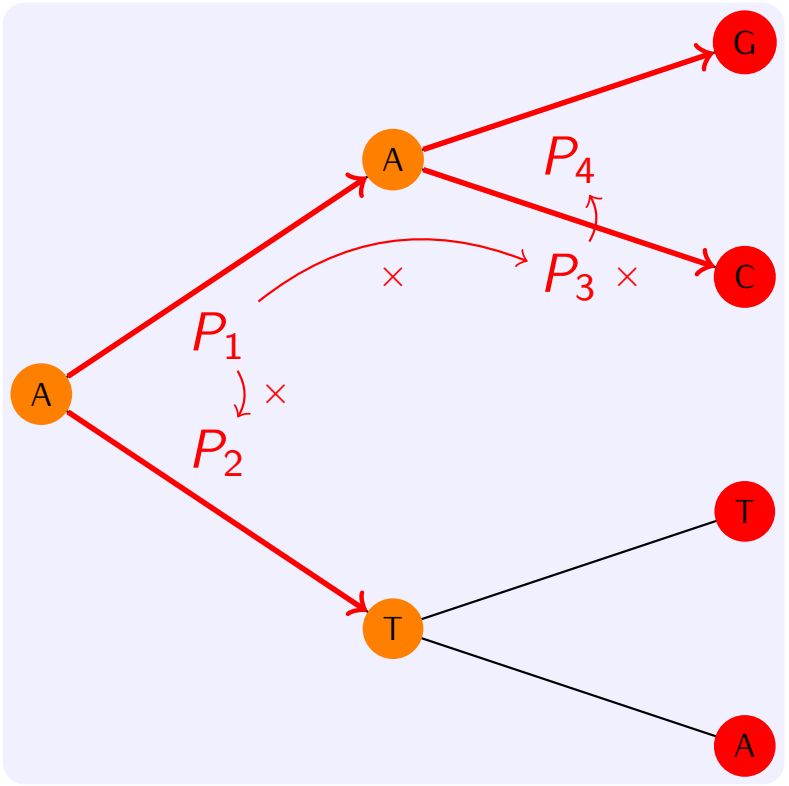
Evolution along a tree



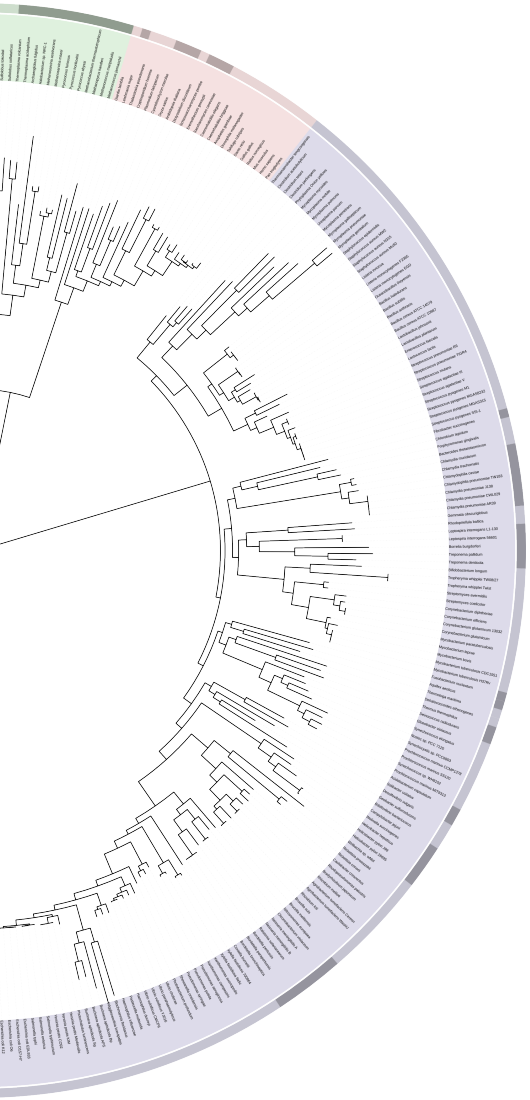


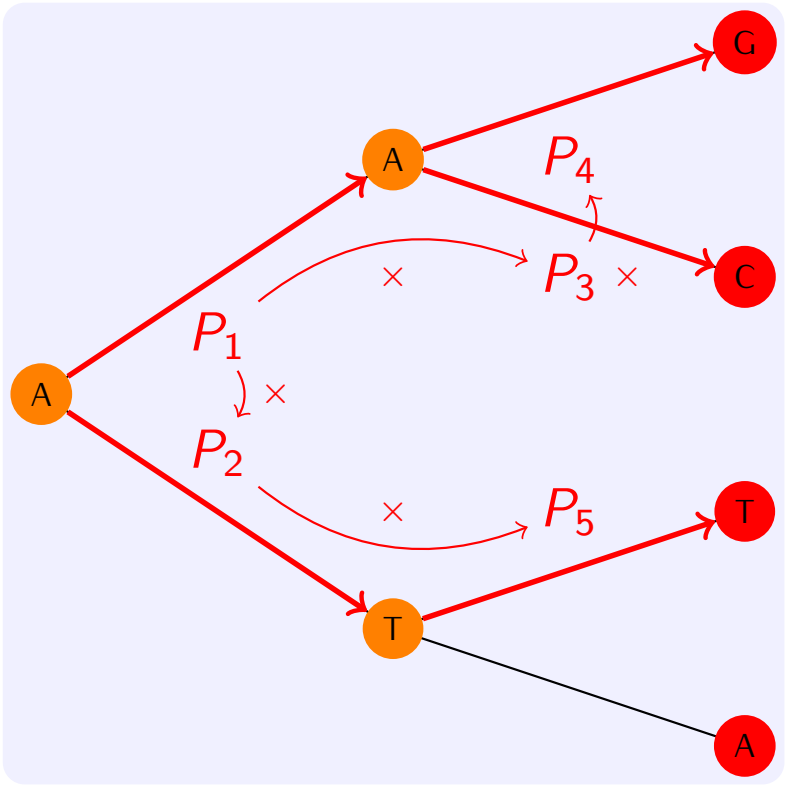
Evolution along a tree



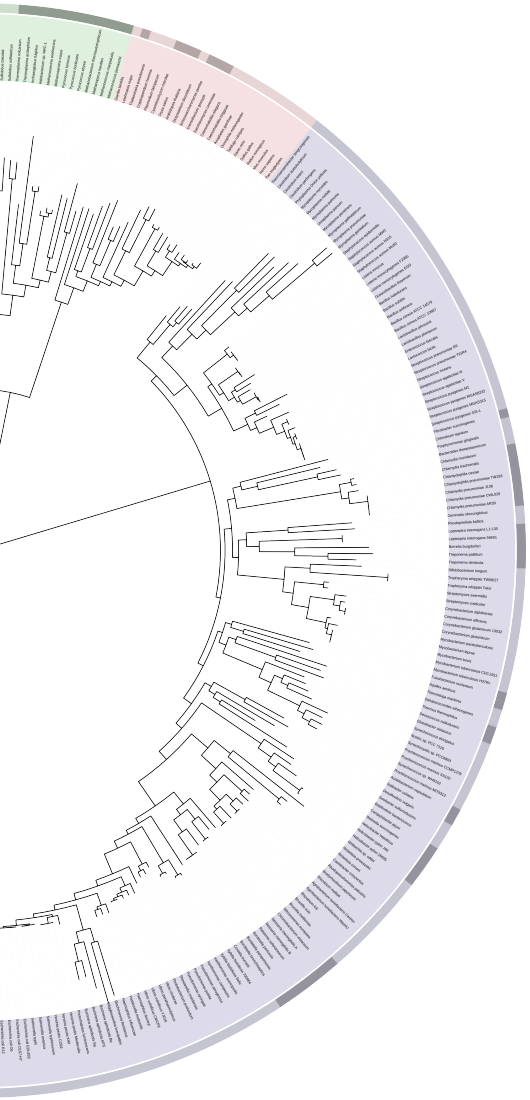


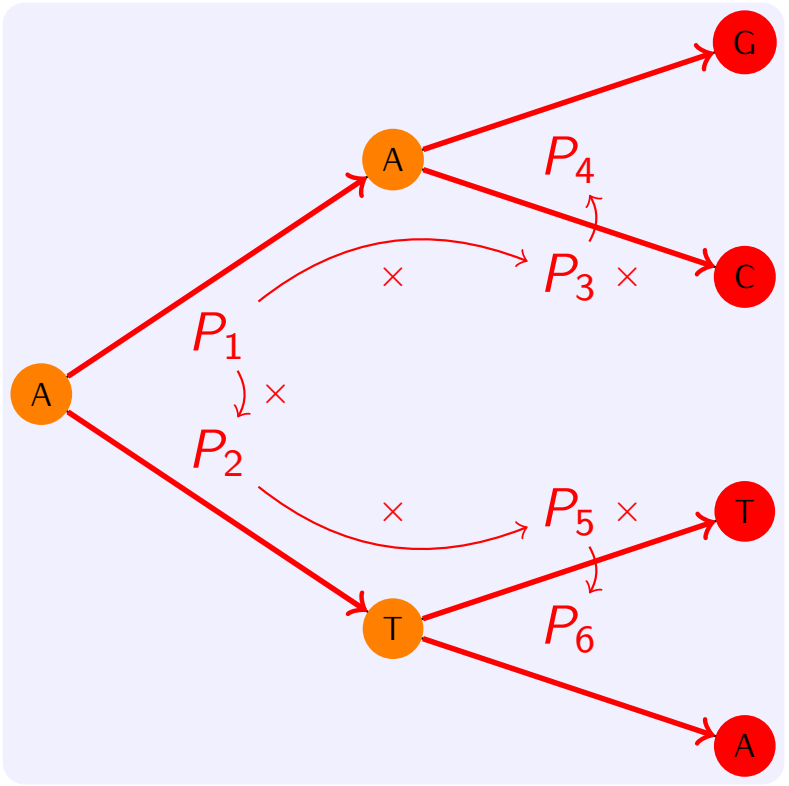
Evolution along a tree



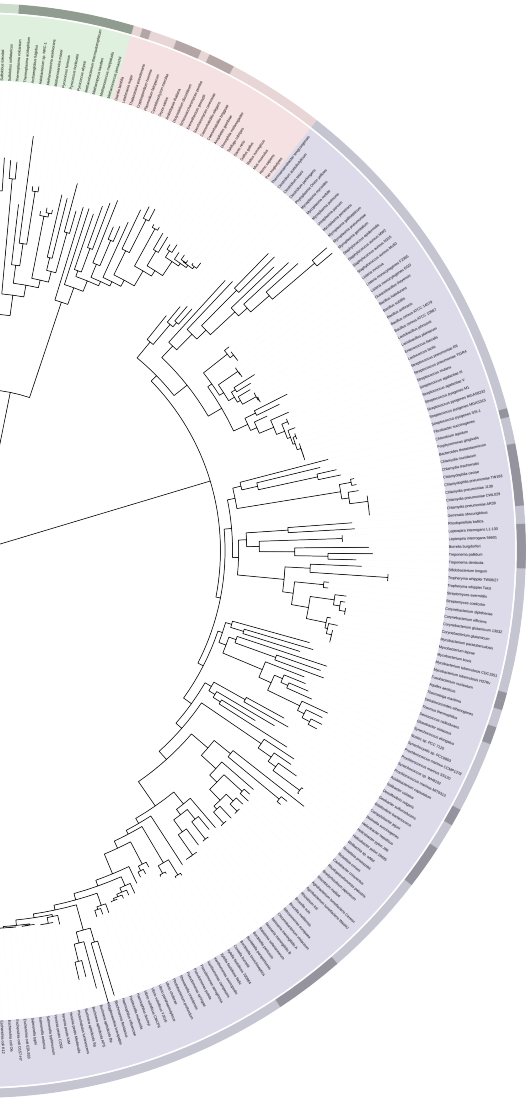


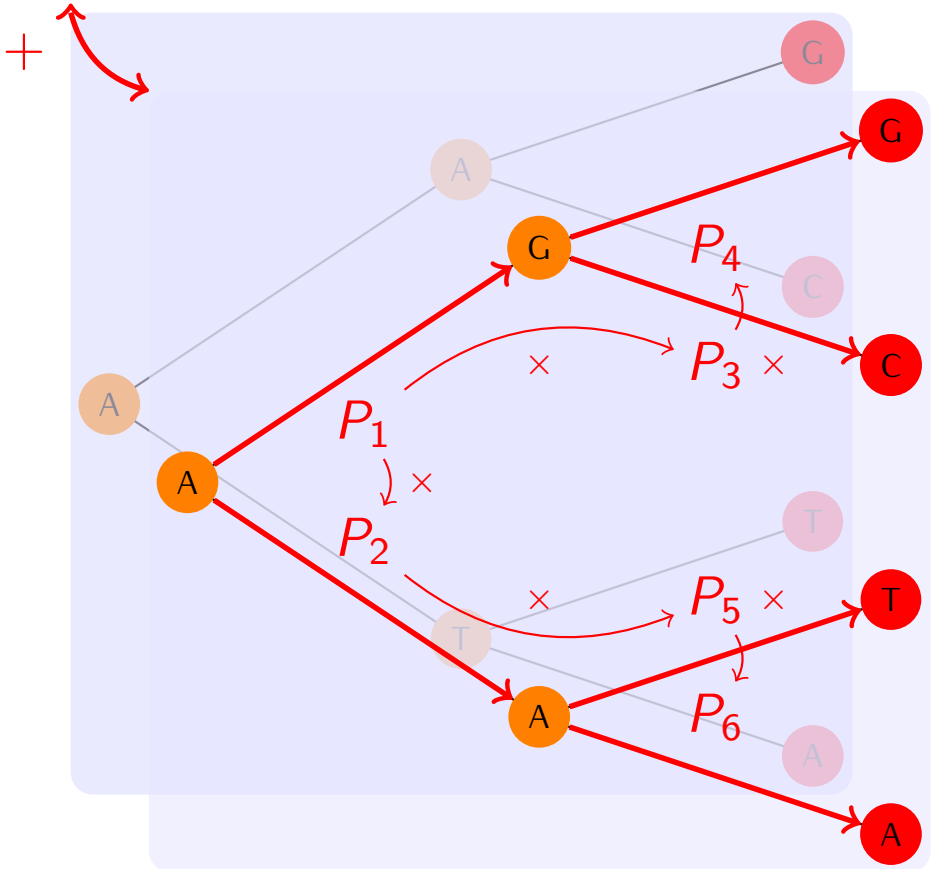
Evolution along a tree



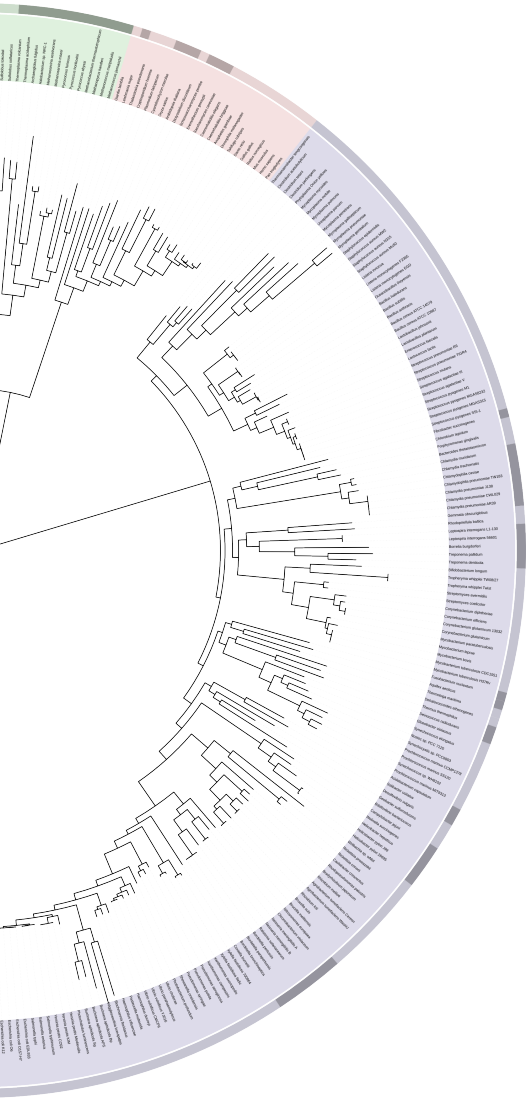


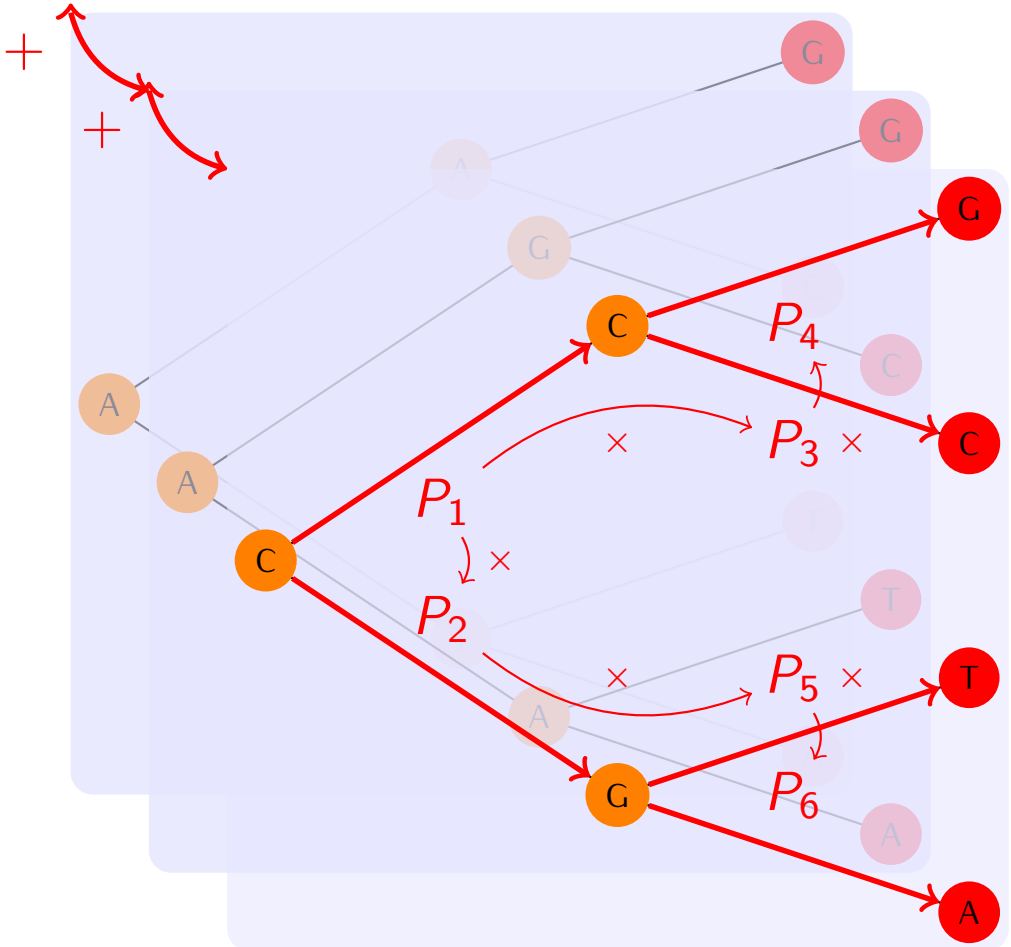
Evolution along a tree



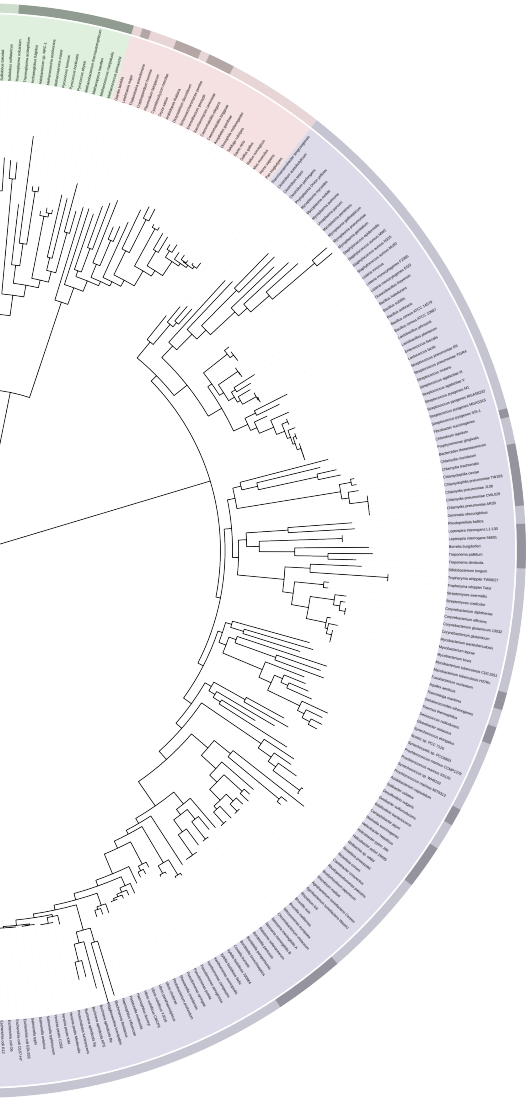


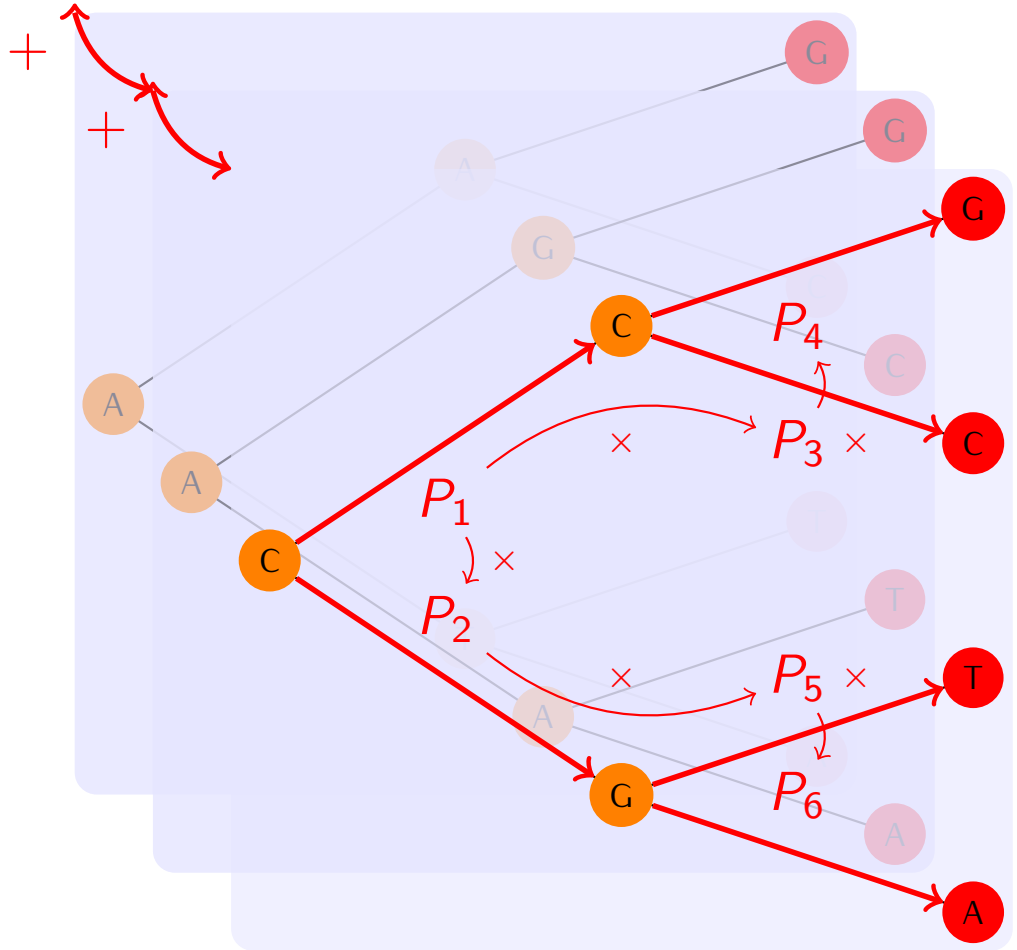
Evolution along a tree





Evolution along a tree





$$L = \sum_{\text{Ancestors}} P_1 \times P_2 \times P_3 \times P_4 \times P_5 \times P_6$$

Maximum likelihood

Maximum-likelihood estimation (MLE)

MLE is a method of estimating the parameters of a statistical model. For a given dataset and underlying statistical model, the maximum likelihood estimator corresponds to the set of values of the model parameters that maximizes the likelihood function. (The method was initially proposed by statistician Ronald Aylmer Fisher in 1922.)

Maximum likelihood

Maximum-likelihood estimation (MLE)

MLE is a method of estimating the parameters of a statistical model. For a given dataset and underlying statistical model, the maximum likelihood estimator corresponds to the set of values of the model parameters that maximizes the likelihood function. (The method was initially proposed by statistician Ronald Aylmer Fisher in 1922.)

- General statistical framework
- Allows to perform model comparisons
- Allows to get confidence intervals of estimates

[However...]

...is tree topology (really) a parameter?

The bootstrap procedure

1. Sample sites from the original alignment

The bootstrap procedure

1. Sample sites from the original alignment
2. Build a new phylogeny from the *neo* data set

The bootstrap procedure

1. Sample sites from the original alignment
2. Build a new phylogeny from the *neo* data set
3. Compare the new phylogeny to the one inferred from the data: which internal branches are present?

The bootstrap procedure

1. Sample sites from the original alignment
2. Build a new phylogeny from the *neo* data set
3. Compare the new phylogeny to the one inferred from the data: which internal branches are present?
4. GOTO [1] many times

The bootstrap procedure

1. Sample sites from the original alignment
2. Build a new phylogeny from the *neo* data set
3. Compare the new phylogeny to the one inferred from the data: which internal branches are present?
4. GOTO [1] many times

$\Rightarrow$ Allows to compute to which extent each internal branch is supported by the data

What can bias the reconstruction?

- Long branch attraction

What can bias the reconstruction?

- Long branch attraction
- Model misspecification

What can bias the reconstruction?

- Long branch attraction
- Model misspecification
- Inconsistency between the history of sequences and the history of species:
  - Stochasticity, incomplete lineage sorting
  - Introgression, horizontal gene transfer
  - Selection, convergence