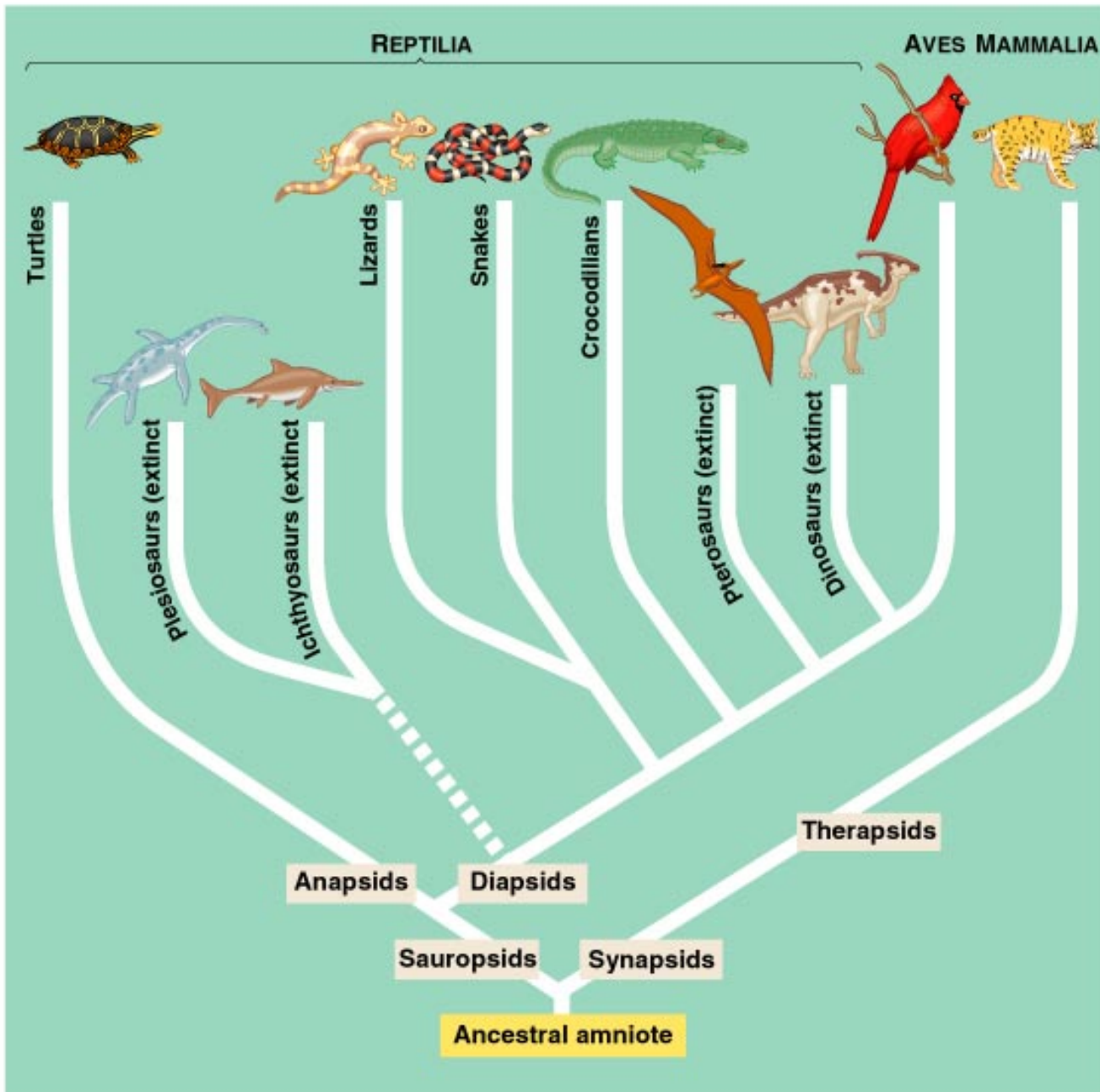

Introduction to Phylogenetics

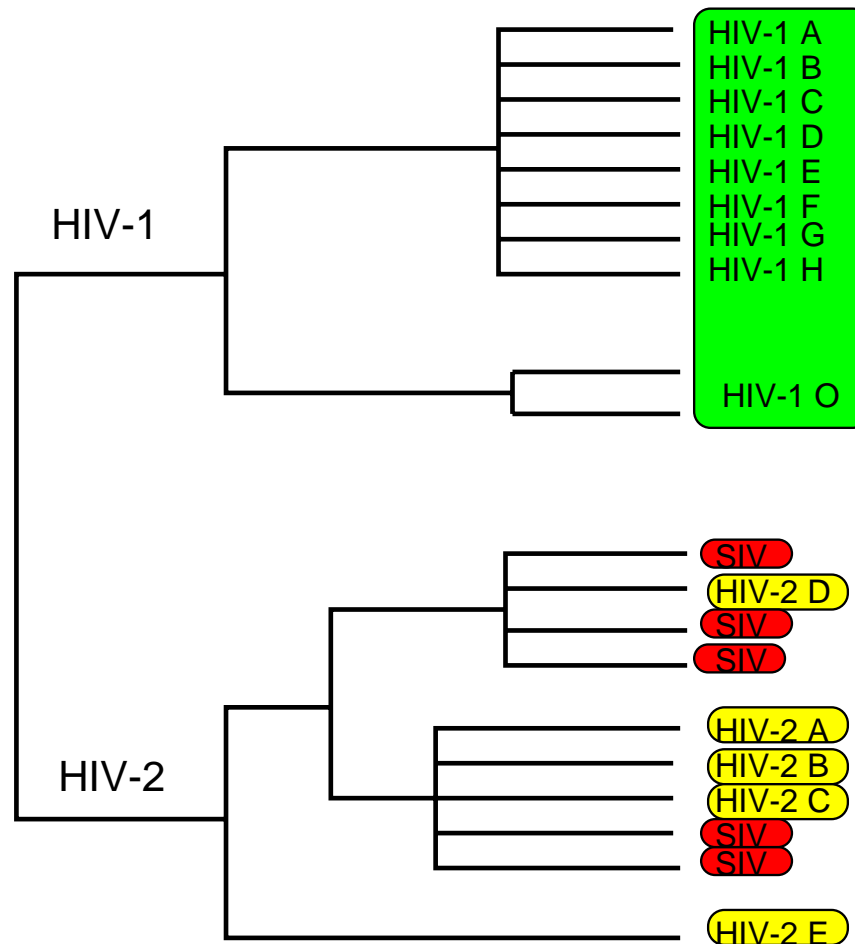
Dirk Husmeier

Biomathematics and Statistics Scotland
at the Scottish Crop Research Institute
Invergowrie, Dundee DD2 5DA, UK

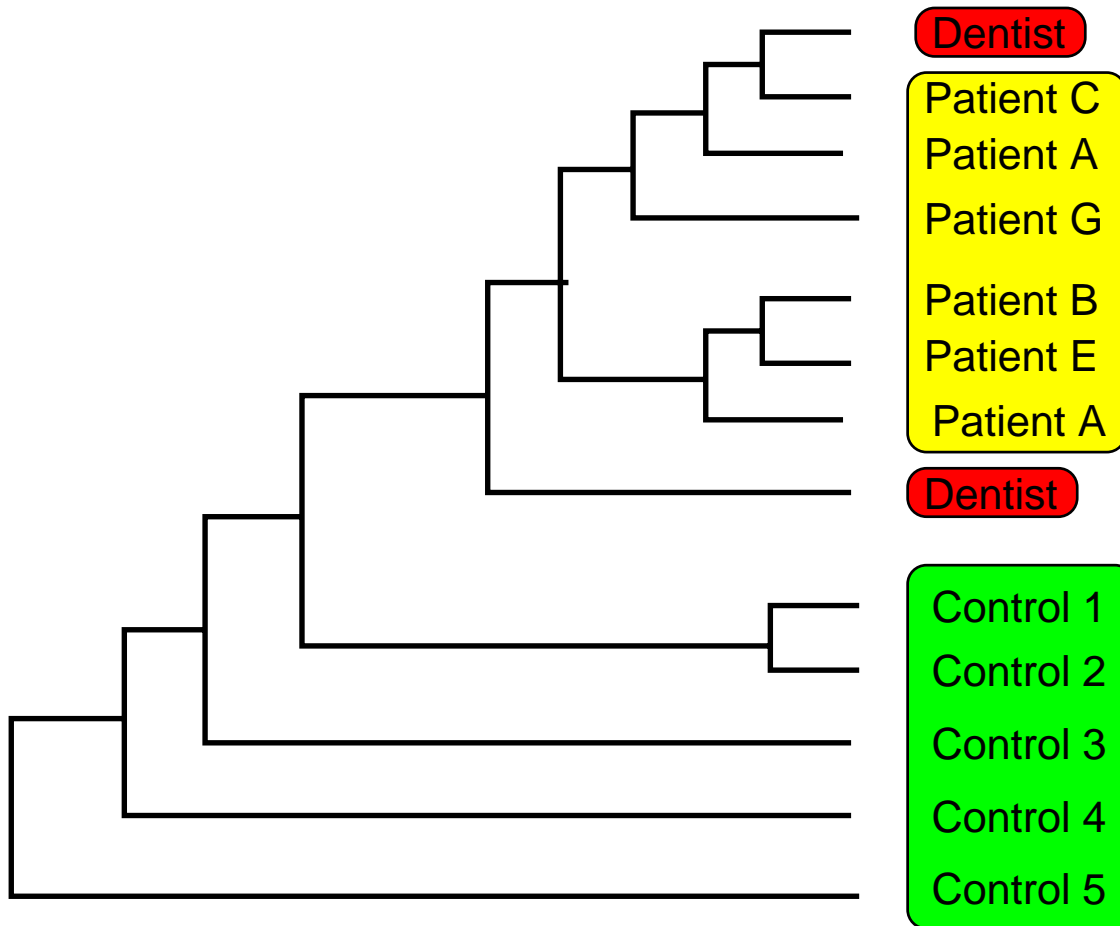
Email: dirk@bioss.ac.uk

<http://www.bioss.ac.uk/~dirk>





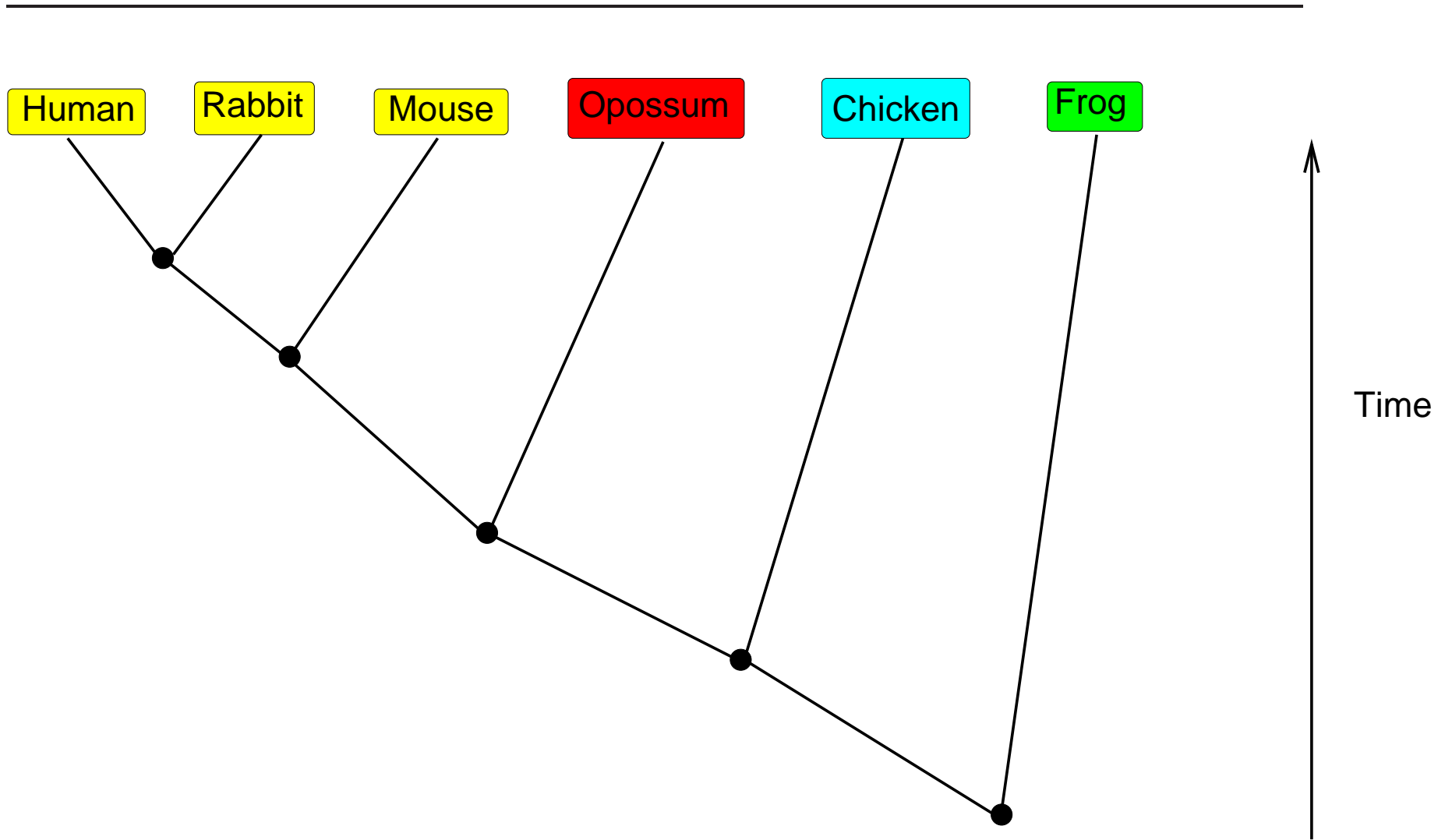
Adapted from Holmes (1998): Evolution in Health and Disease, Oxford University Press



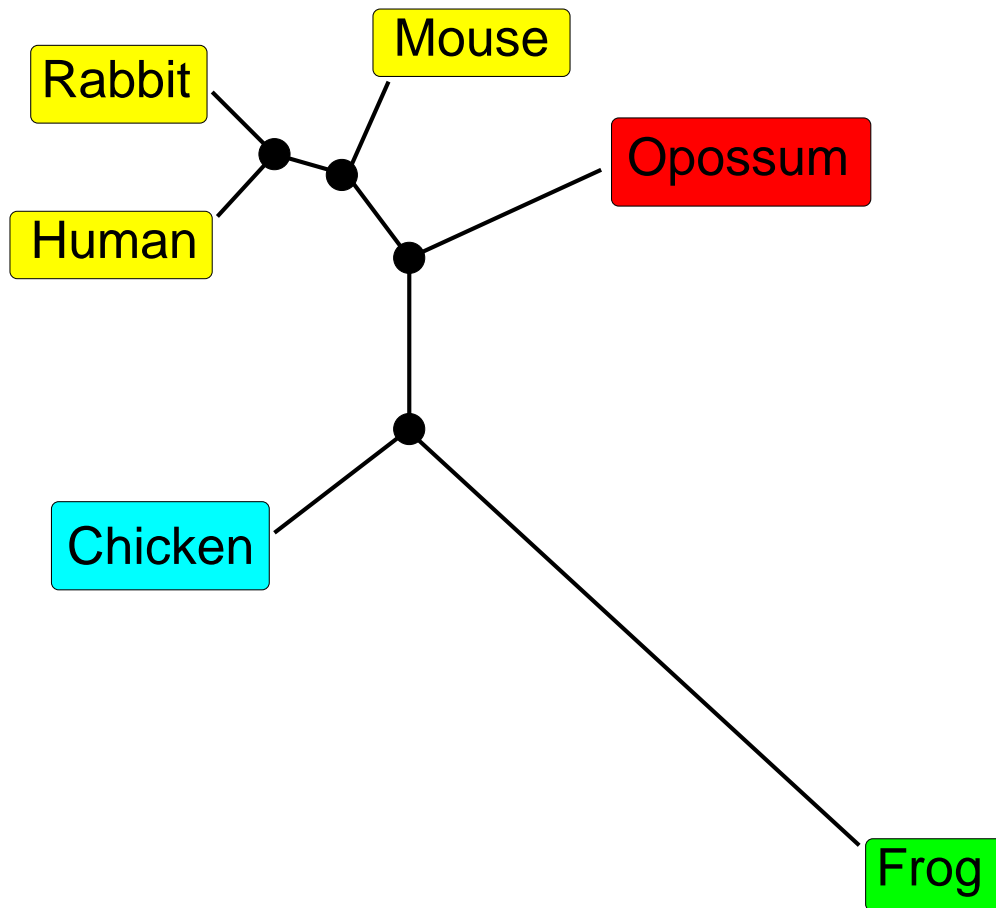
Data from Ou et al. (1992): Science 256, 1165-1171

Tree adapted from Page & Holmes (1998), Blackwell Science

Rooted Phylogenetic Tree



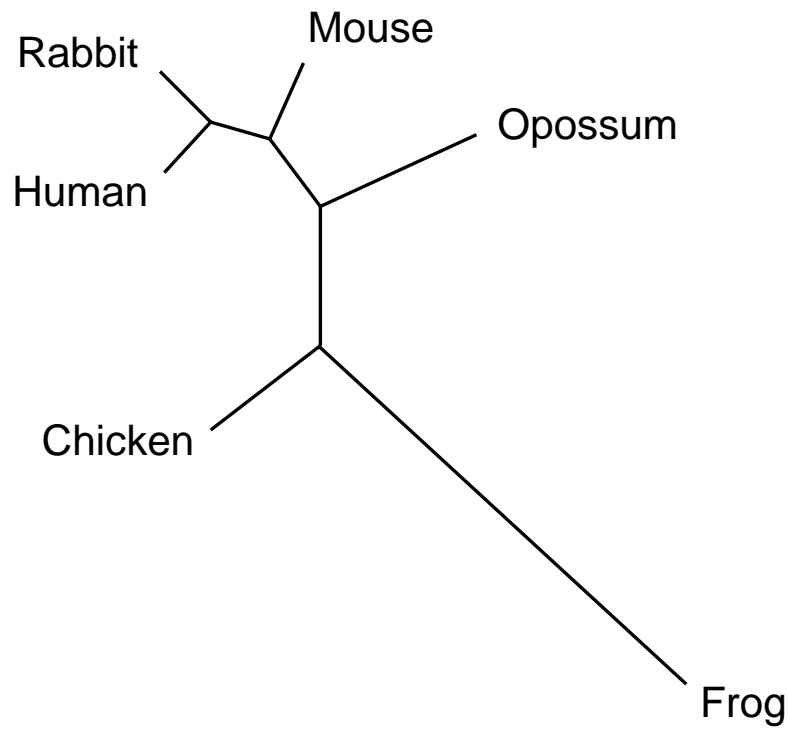
Unrooted Phylogenetic Tree



--> Topology

--> Branch lengths

Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T



--> Topology
 --> Branch lengths

Methods of phylogenetic inference

- Clustering
- Parsimony
- Likelihood

Inferring Phylogeny from Pairwise Distances

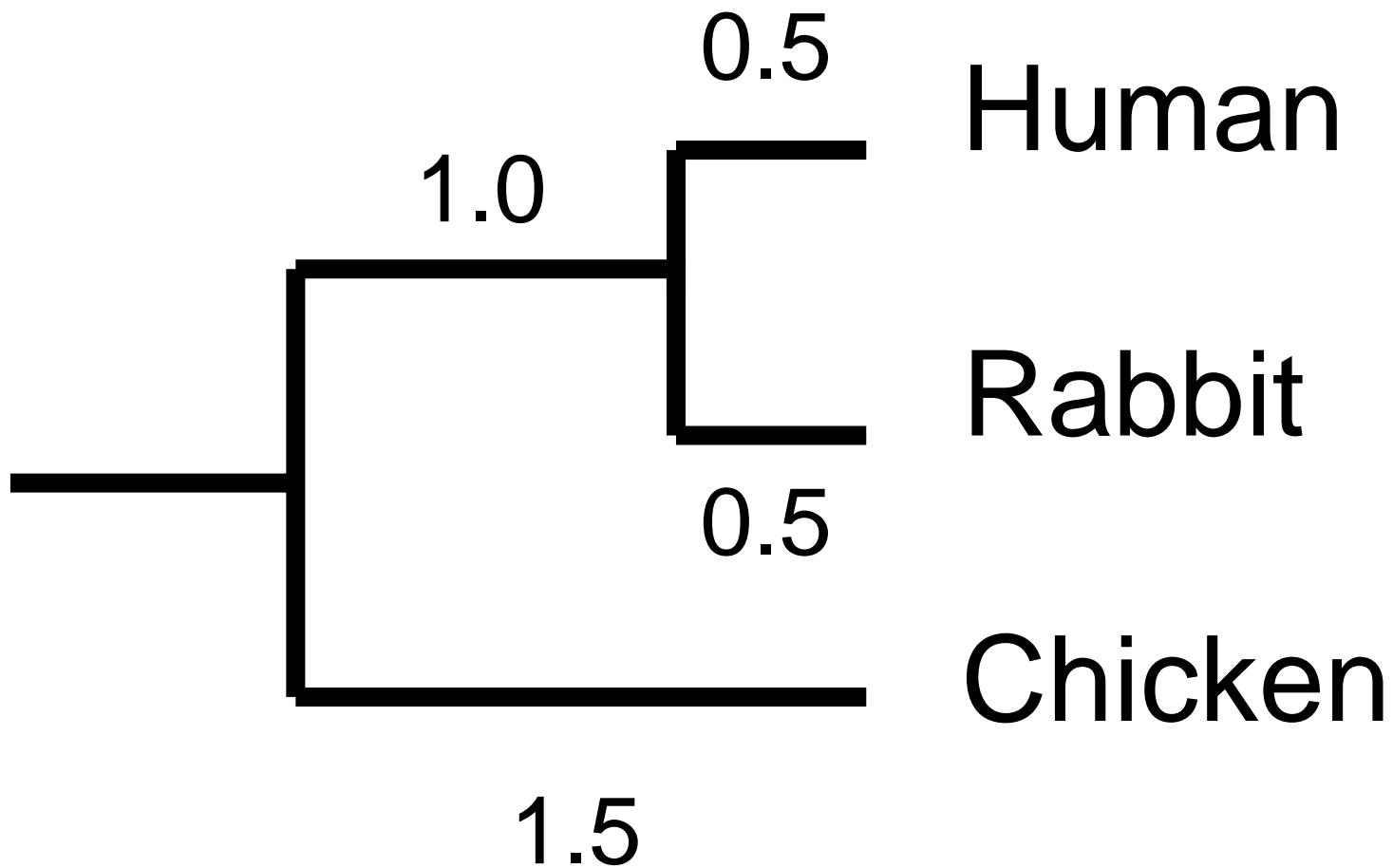
Human ... T G T A T C G C T C ...
Rabbit ... T G T G T C G C T C ...

Human ... T G T A T C G C T C ...
Chicken ... A G T C T C G T T C ...

Rabbit ... T G T G T C G C T C ...
Chicken ... A G T C T C G T T C ...

	Rabbit	Chicken
Human	1	3
Rabbit		3

Inferring Phylogeny from Pairwise Distances

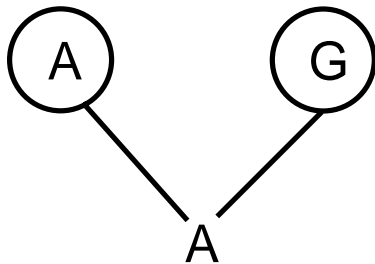


Genetic Distance

- Naive distance measure: **Hamming distance** $d_0 =$
Proportion of sites at which the two sequences differ.
- Poor measure of the actual number of evolutionary changes, as a site can undergo **repeated substitutions**.

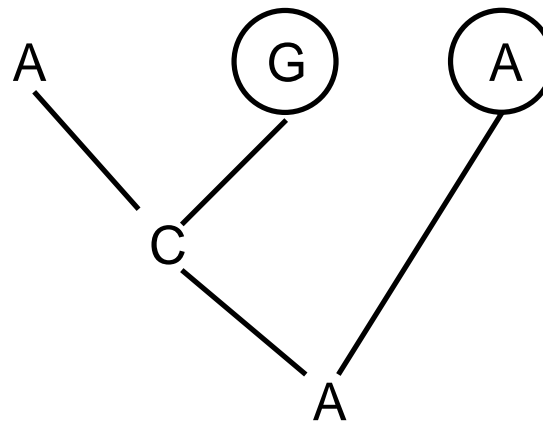
Single substitution

1 change, 1 difference



Multiple substitution

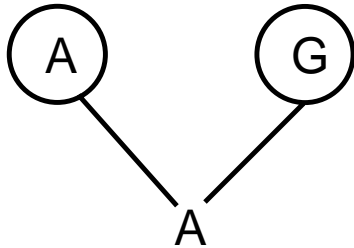
2 changes, 1 difference



Multiple Hits and Reversals

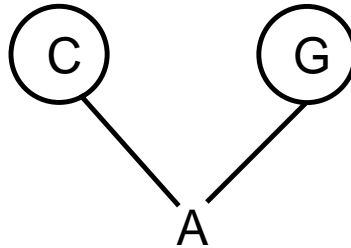
Single substitution

1 change, 1 difference



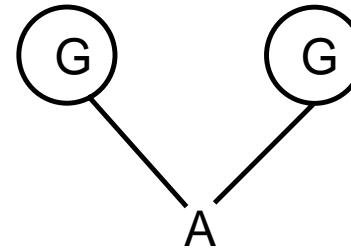
Coincidental substitution

2 changes, 1 difference



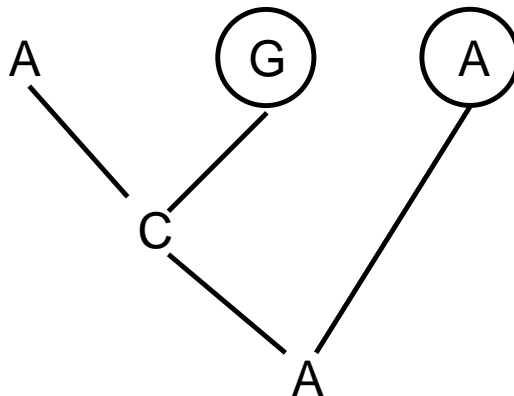
Parallel substitution

2 changes, no difference



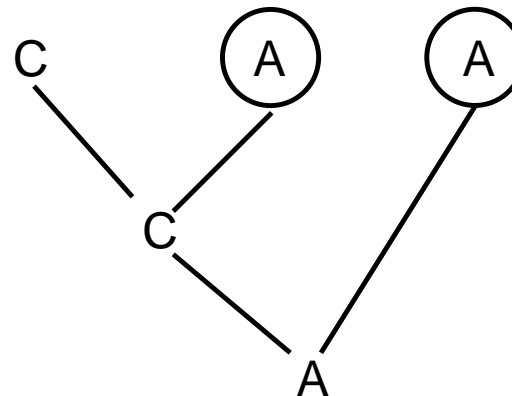
Multiple substitution

2 changes, 1 difference



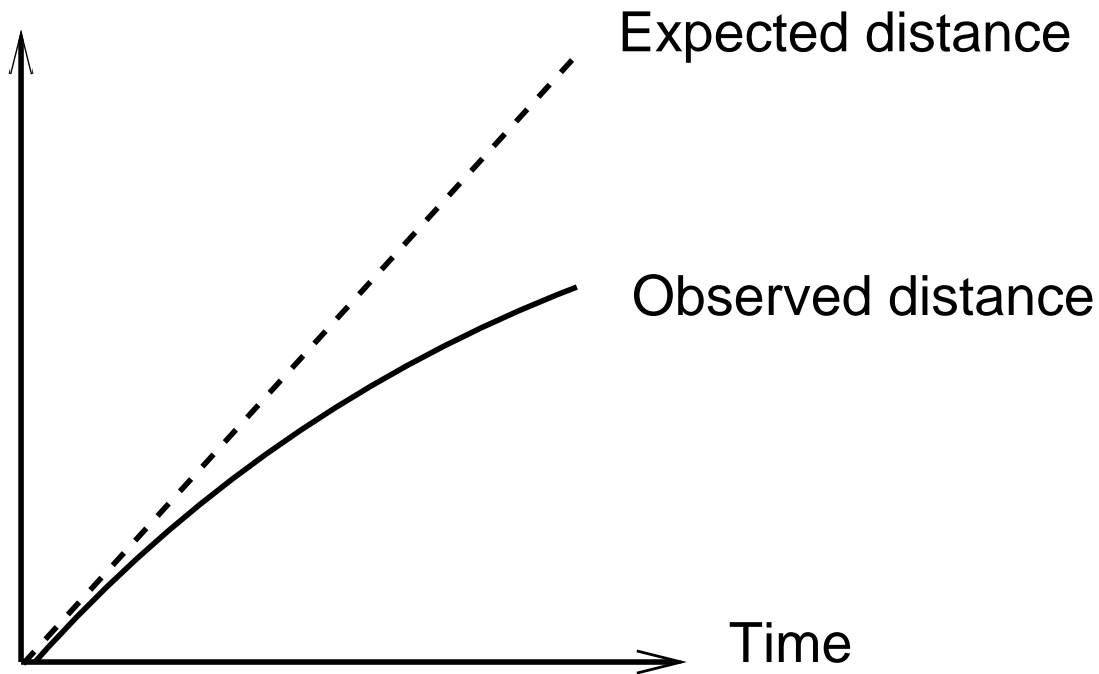
Back substitution

2 changes, no difference



Observed and Expected Genetic Distances

Sequence distance



Observed distance: $d_0(t \rightarrow \infty) = 3/4$

Corrected distance: $d = -\frac{3}{4} \log(1 - \frac{4}{3}d_0)$

Inferring Phylogeny by Clustering: UPGMA

Definition

Distance d_{AB} between clusters A, B from individual distances d_{ab} :

$$d_{AB} = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d_{ab}$$

Algorithm

Initialization

- Assign each sequence i to its own cluster C_i . Define one leaf for each sequence, and place at height zero.

Iteration

- Determine the two clusters i, j for which d_{ij} is minimal.
- Define a new cluster $C_k = C_i \cup C_j$
- Define a new node k with daughter nodes i and j , and place it at height $d_{ij}/2$.
- Add k to the current clusters and remove i and j .

Termination

- When only two clusters i, j remain, place the root at height $d_{ij}/2$.



1



2



3



5



4

1

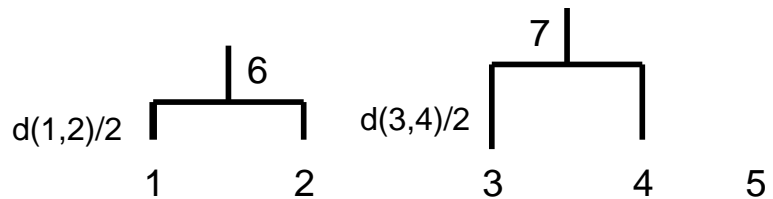
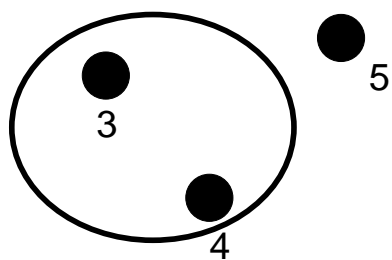
2

3

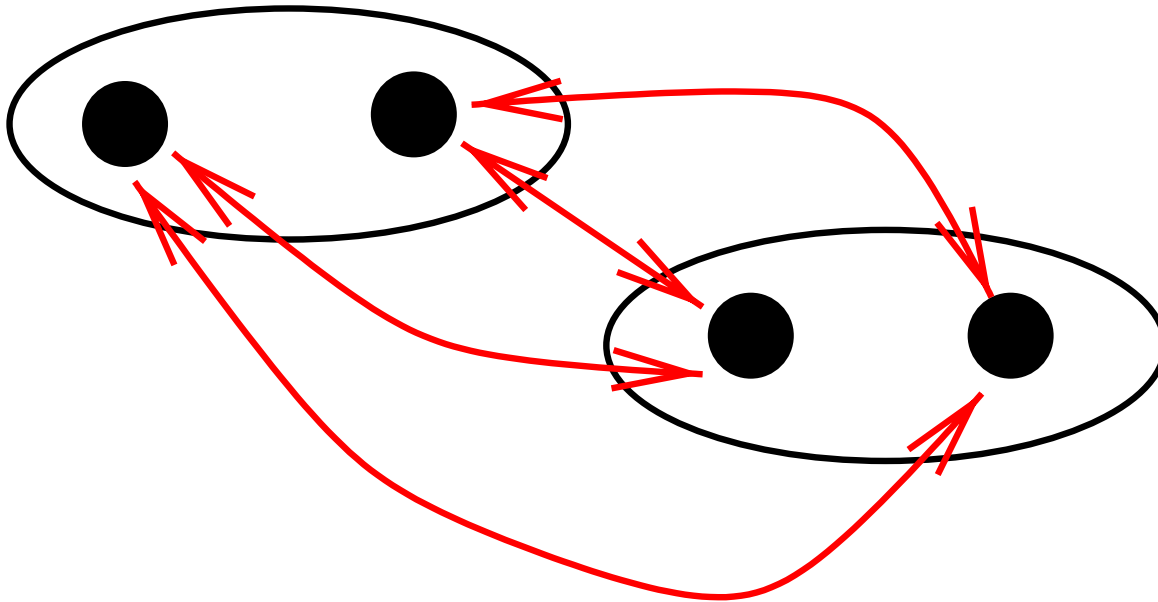
4

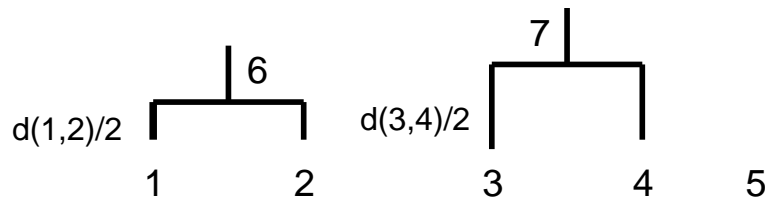
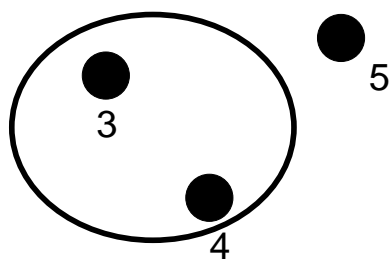
5

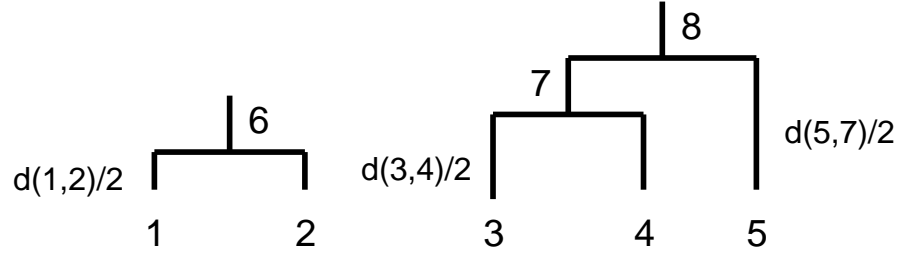
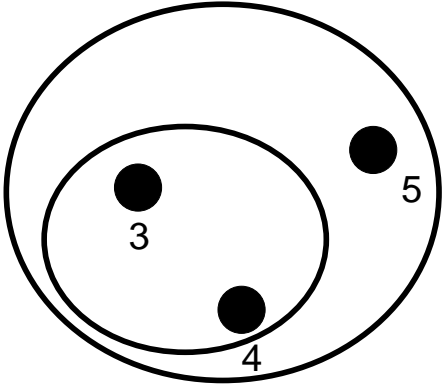


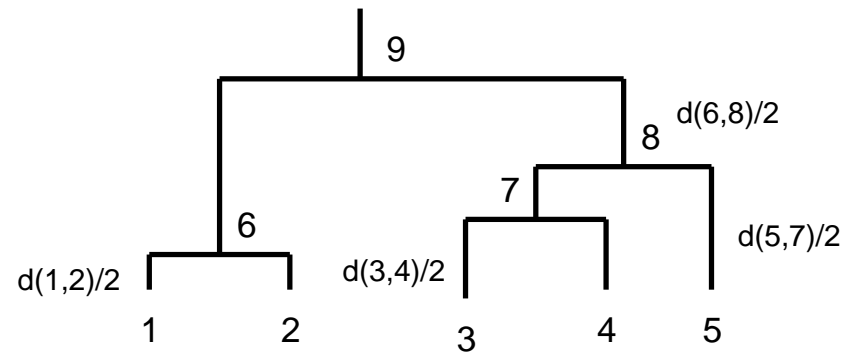
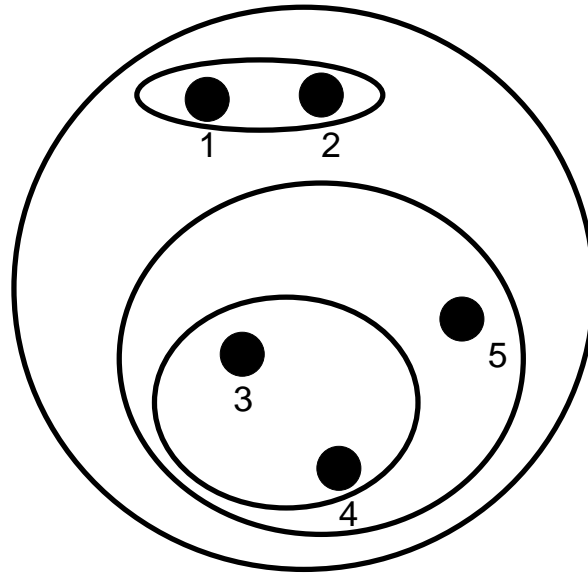


Distance between clusters

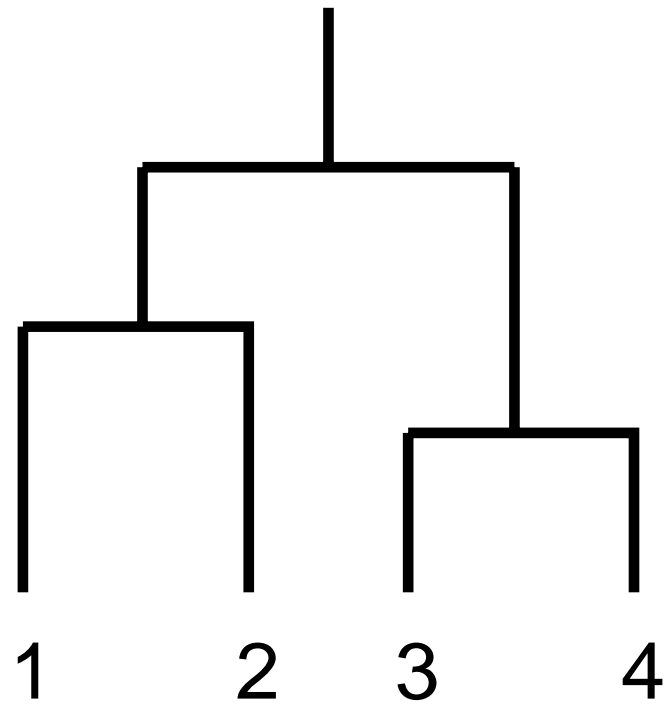
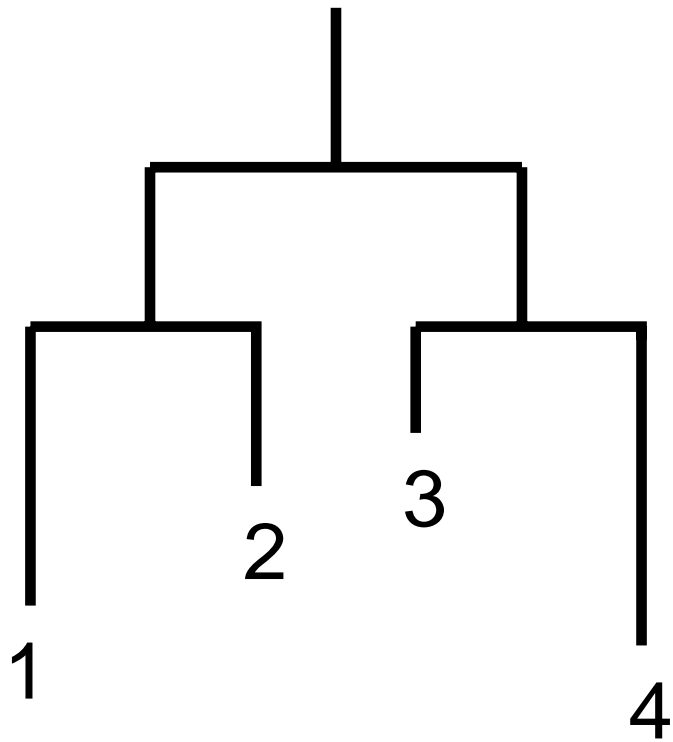




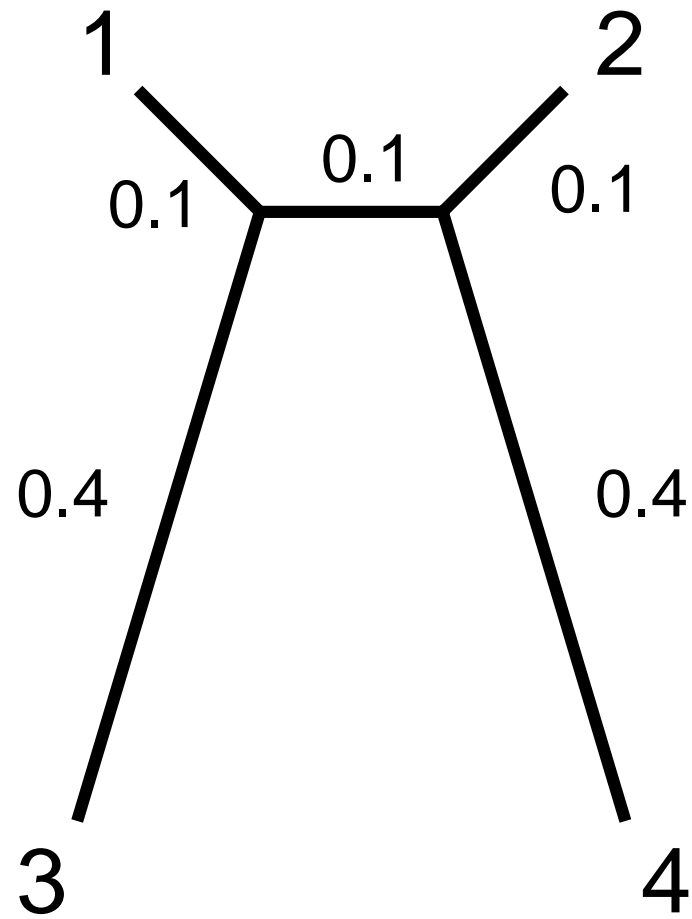




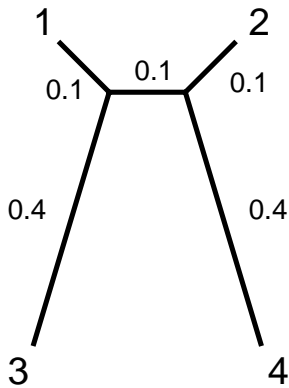
Limitation of UPGMA: Ultrametric Trees



Failure of UPGMA



Neighbour Joining: Corrected Distances



Definition of **corrected 'distance'**: $D_{ij} = d_{ij} - \bar{d}_i - \bar{d}_j$
Average distance to all other leaves: $\bar{d}_i = \frac{1}{|L|-2} \sum_{k \in L} d_{ik}$

$$\bar{d}_1 = \frac{1}{2}(0.3 + 0.6 + 0.5) = 0.7 = \bar{d}_2$$

$$\bar{d}_3 = \frac{1}{2}(0.5 + 0.6 + 0.9) = 1.0 = \bar{d}_4$$

$$D_{12} = d_{12} - \bar{d}_1 - \bar{d}_2 = 0.3 - 0.7 - 0.7 = -1.1$$

$$D_{13} = d_{13} - \bar{d}_1 - \bar{d}_3 = 0.5 - 0.7 - 1.0 = -1.2 < D_{12}$$

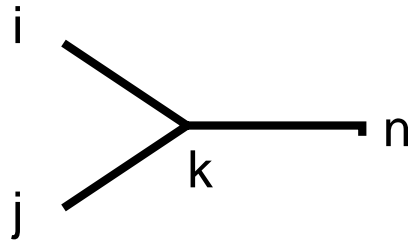
Tree Metric

Non-negativity: $d_{ab} \geq 0$

Symmetry: $d_{ab} = d_{ba}$

Distinctness: $d_{ab} = 0$ if and only if $a = b$.

Triangle Inequality: $d_{ij} \leq d_{ik} + d_{kj} \longrightarrow d_{ij} = d_{ik} + d_{kj}$



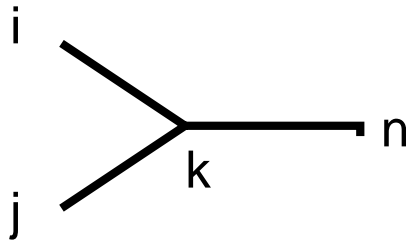
Tree Metric

Non-negativity: $d_{ab} \geq 0$

Symmetry: $d_{ab} = d_{ba}$

Distinctness: $d_{ab} = 0$ if and only if $a = b$.

Triangle Inequality: $d_{ij} \leq d_{ik} + d_{kj} \longrightarrow d_{ij} = d_{ik} + d_{kj}$



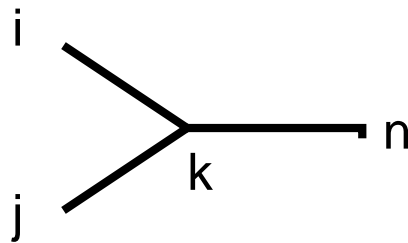
$$d_{in} = d_{ik} + d_{kn}$$

$$d_{jn} = d_{jk} + d_{kn}$$

$$\Rightarrow 2d_{kn} = d_{in} + d_{jn} - d_{ik} - d_{kj}$$

$$\Rightarrow d_{kn} = \frac{1}{2}(d_{in} + d_{jn} - d_{ij})$$

Inferring Phylogeny by Clustering: Neighbour Joining



$$d_{in} = d_{ik} + d_{kn}$$

$$d_{jn} = d_{jk} + d_{kn}$$

$$\Rightarrow 2d_{kn} = d_{in} + d_{jn} - d_{ik} - d_{kj}$$

$$\Rightarrow d_{kn} = \frac{1}{2}(d_{in} + d_{jn} - d_{ij})$$

Iteration

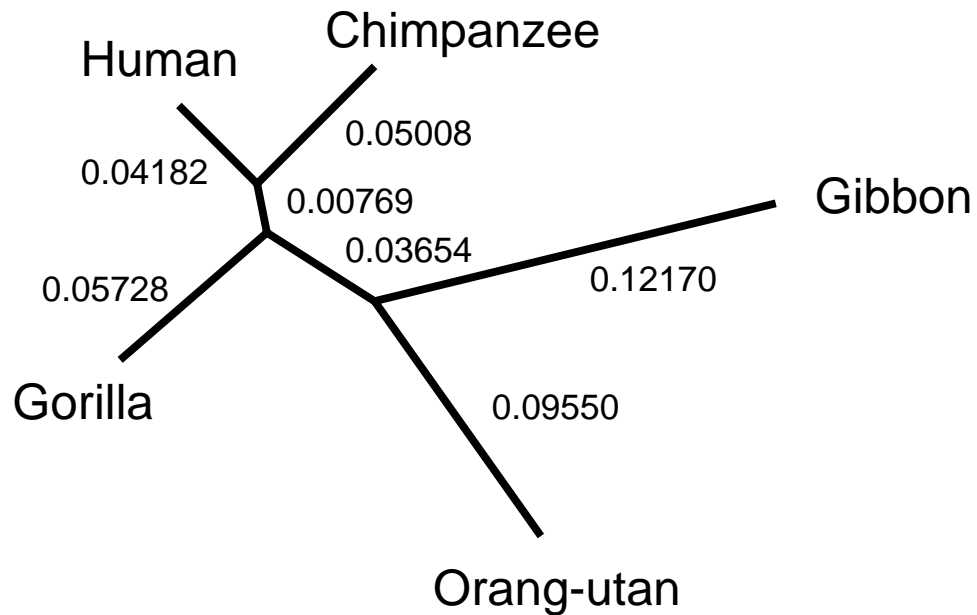
- Find pair of node (i, j) that minimize D_{ij} .
- Replace (i, j) by new node k with new distances:

$$d_{kn} = \frac{1}{2}(d_{in} + d_{jn} - d_{ij})$$

Application of Neighbour Joining

	Human	Chimp	Gorilla	Orang-utan	Gibbon
Human	-	0.0919	0.1083	0.1790	0.2057
Chimp	0.0919	-	0.1134	0.1940	0.2168
Gorilla	0.1068	0.1151	-	0.1882	0.2170
Orang-utan	0.1816	0.1898	0.1893	-	0.2172
Gibbon	0.2078	0.2160	0.2155	0.2172	-

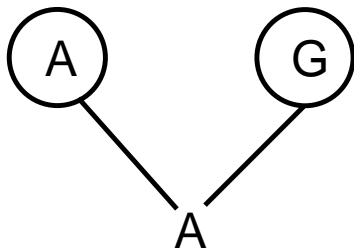
Right: Observed distances. Left: Distances estimated from tree.



Observed distances < Actual distances

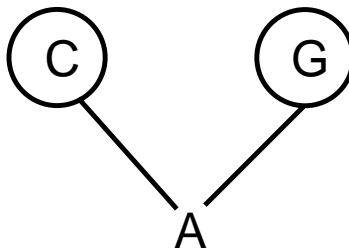
Single substitution

1 change, 1 difference



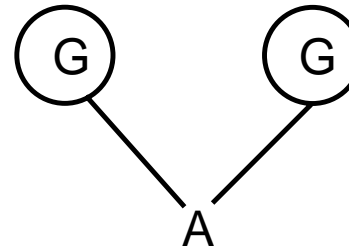
Coincidental substitution

2 changes, 1 difference



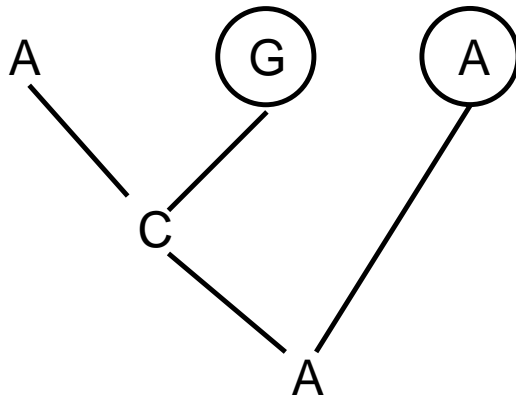
Parallel substitution

2 changes, no difference



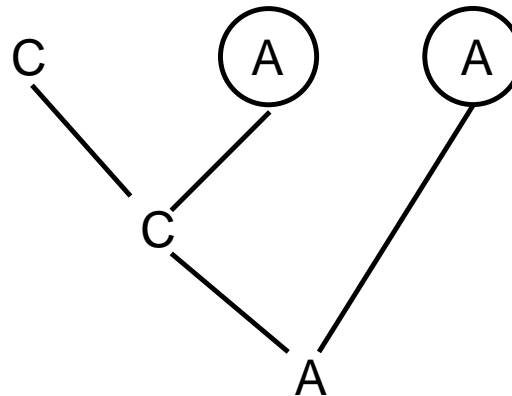
Multiple substitution

2 changes, 1 difference



Back substitution

2 changes, no difference



Shortcomings of Distance and Clustering Methods

- Loss of Information

	Sequences			Distances
1	T T A T T A A C G	→	2	3
2	A A T T T A A C G		3	5 4
3	A A A A A T A C G		4	5 4 2
4	A A A A A A T C G			1 2 3

- Uninterpretable branch lengths

- $d_{ij}^{tree} < d_{ij}^{obs}$ biologically impossible
- Occasionally even $d_{ij}^{tree} < 0$

- The method does not optimize an objective function

Clustering methods merely produce a tree, but do not allow us

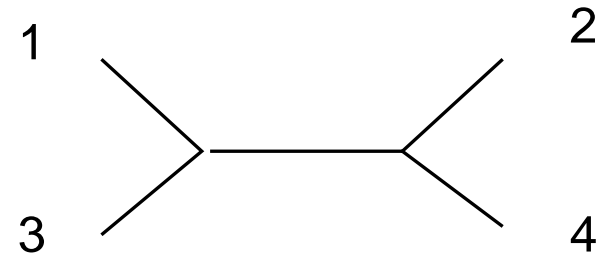
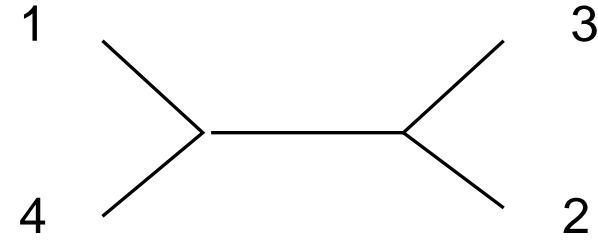
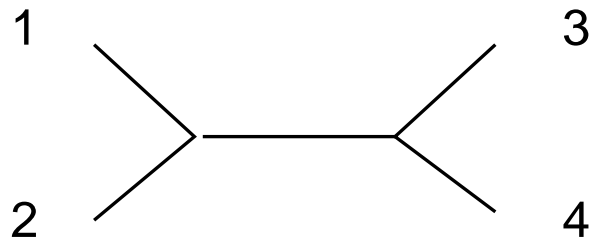
- to evaluate the quality of the tree
- to evaluate competing hypotheses

Methods of phylogenetic inference

- Clustering
- Parsimony
- Likelihood

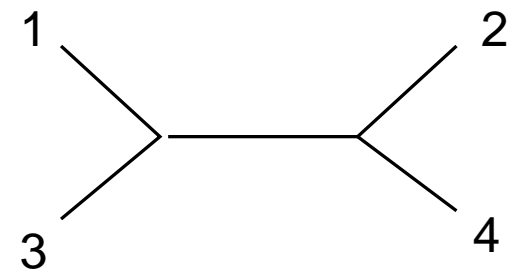
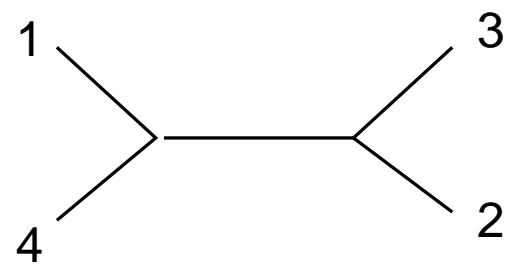
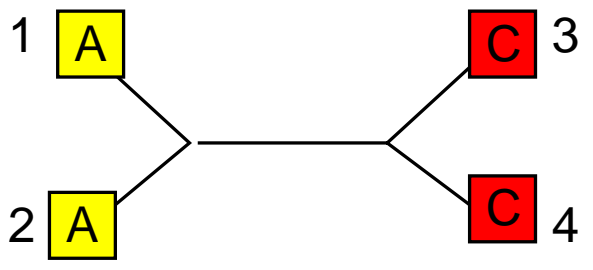
1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

■ ■ ■



∇

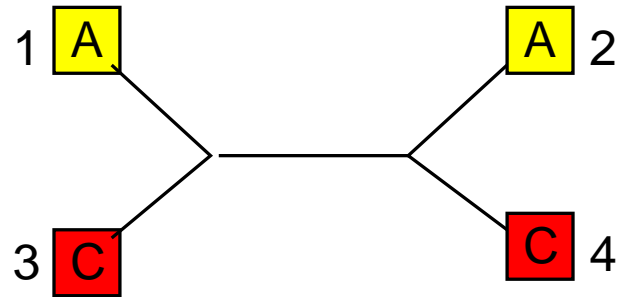
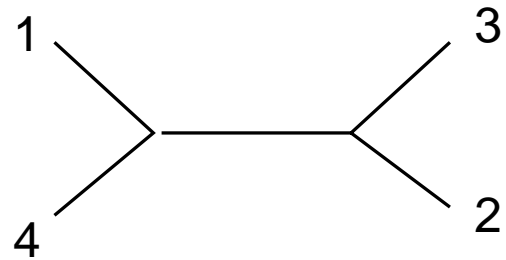
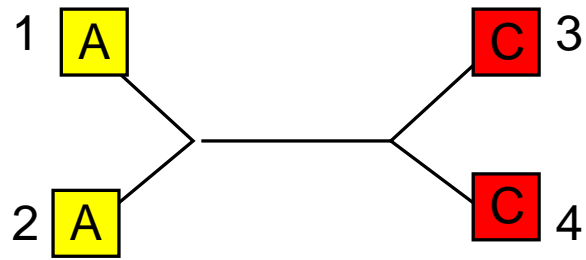
1	A	C	A	C	G			
2	A	T	T	C	G			
3	C	G	A	G	G	▪	▪	▪
4	C	G	G	C	G			



∨

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

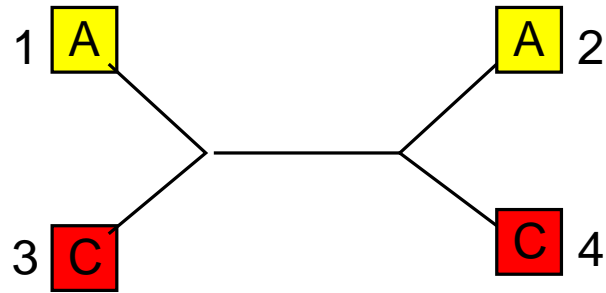
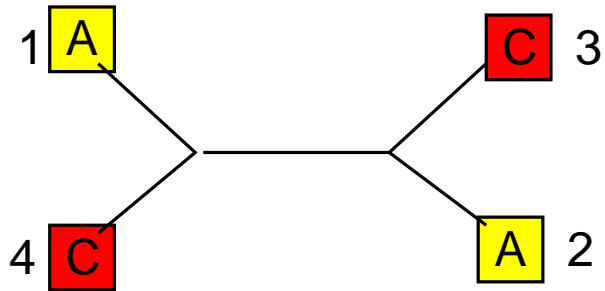
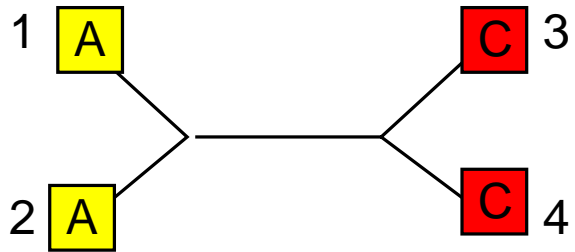
▪ ▪ ▪



∨

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

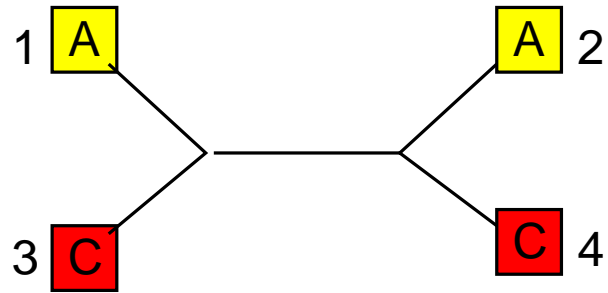
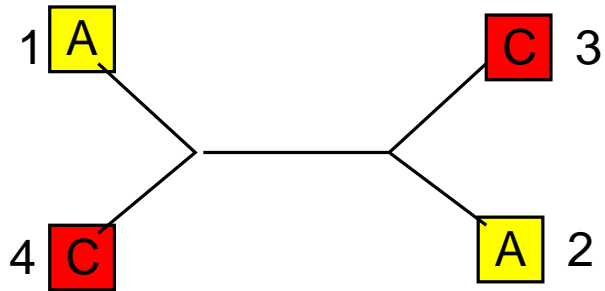
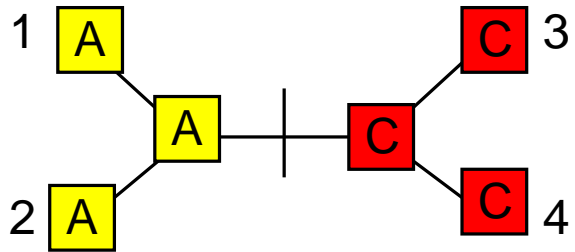
. . .



∨

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

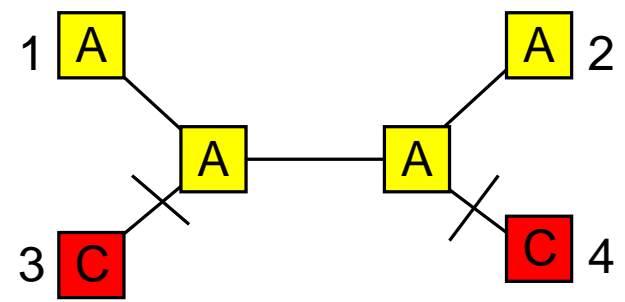
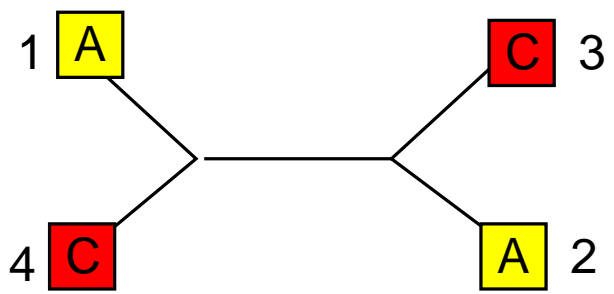
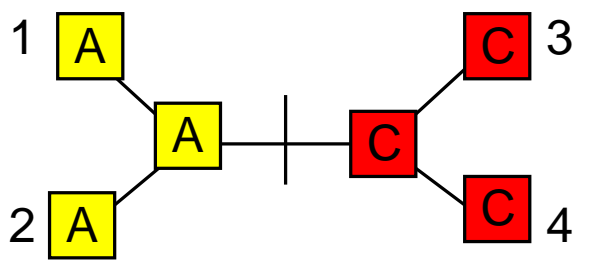
. . .



↓

1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

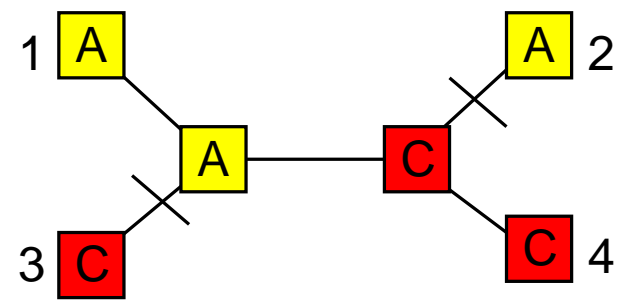
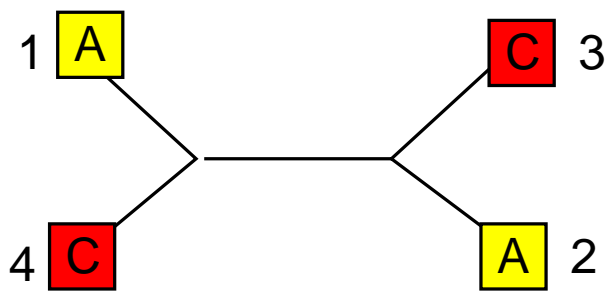
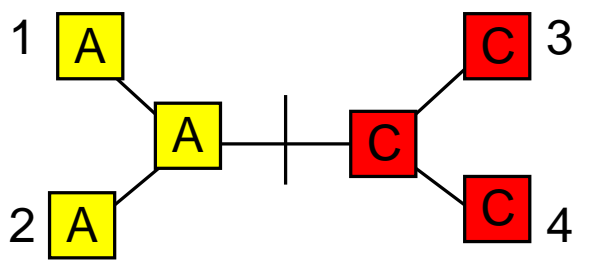
. . .



∇

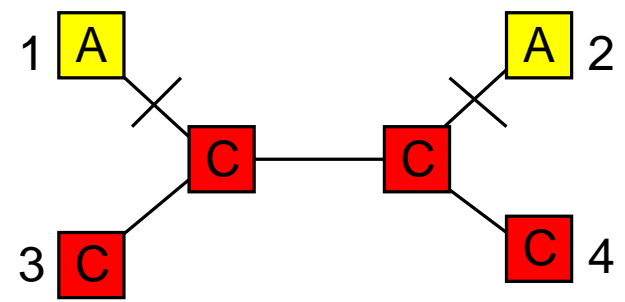
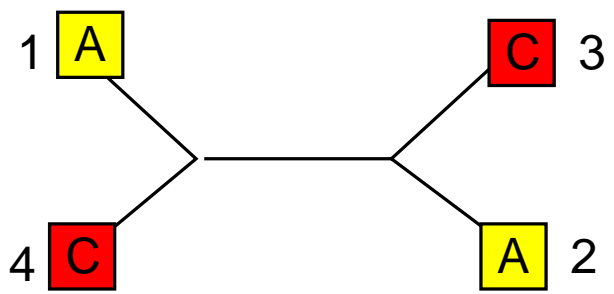
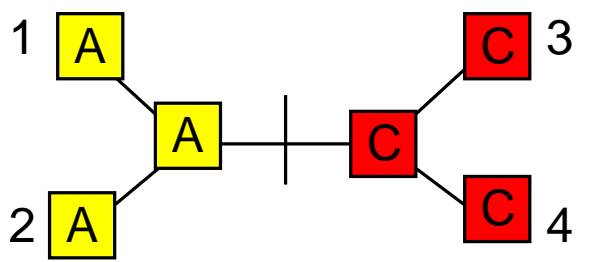
1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

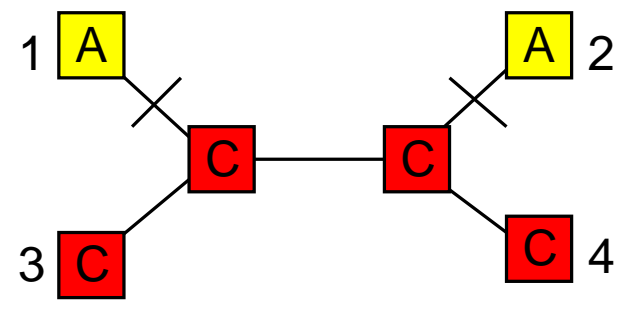
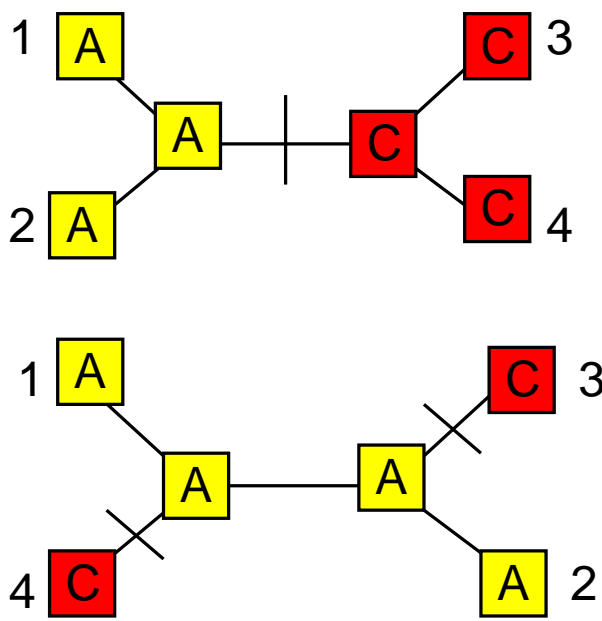
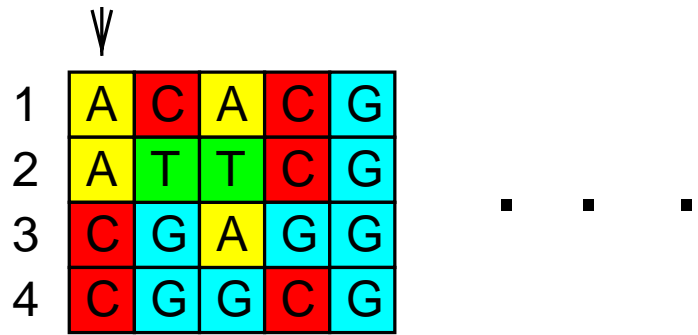
. . .



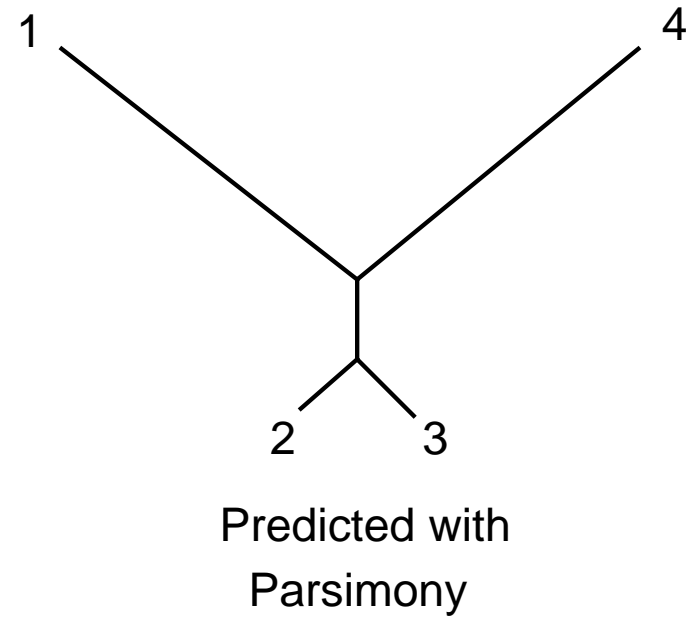
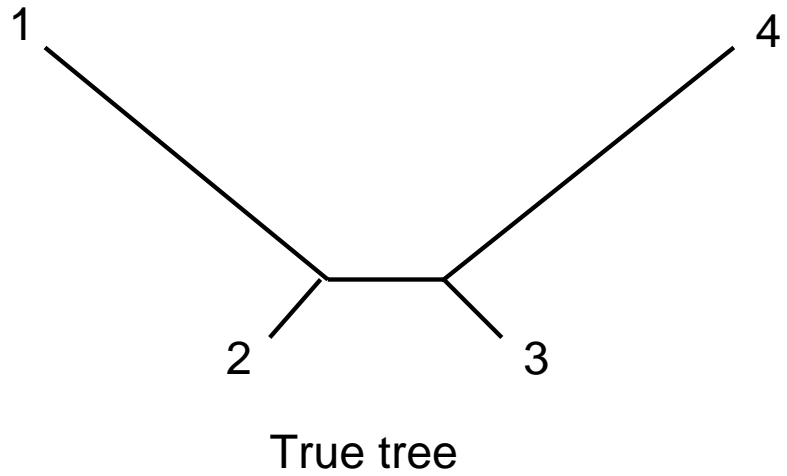
	∇				
1	A	C	A	C	G
2	A	T	T	C	G
3	C	G	A	G	G
4	C	G	G	C	G

. . .





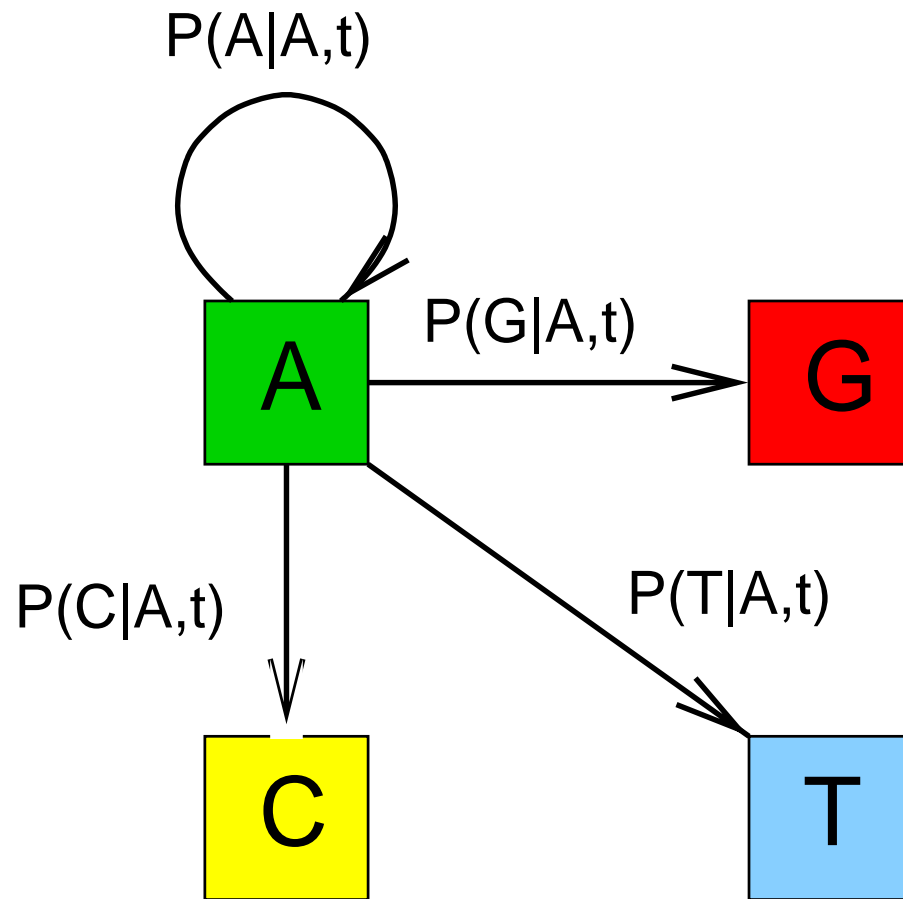
Failure of parsimony



Methods of phylogenetic inference

- Clustering
- Parsimony
- Likelihood

Mutation probabilities



Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

- The Markov process is **homogenous**:

$$P[y(t + t_0)|y(t_0)] = P[y(t)|y(0)]$$

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

- The Markov process is **homogenous**:

$$P[y(t + t_0)|y(t_0)] = P[y(t)|y(0)]$$

- The Markov process is the **same for all positions**

Markov model of evolution

$$\mathbf{P}(t) = \begin{bmatrix} P(y(t) = A|y(0) = A) & P(y(t) = A|y(0) = G) & \dots \\ P(y(t) = G|y(0) = A) & P(y(t) = G|y(0) = G) & \dots \\ P(y(t) = C|y(0) = A) & P(y(t) = C|y(0) = G) & \dots \\ P(y(t) = T|y(0) = A) & P(y(t) = T|y(0) = G) & \dots \end{bmatrix}$$

- Process is **Markov**:

$$P[y(t + \Delta t)|y(t), y(t - \Delta t), \dots] = P[y(t + \Delta t)|y(t)]$$

- The Markov process is **homogenous**:

$$P[y(t + t_0)|y(t_0)] = P[y(t)|y(0)]$$

- The Markov process is the **same for all positions**
- Substitutions at different positions are **independent** of each other:

$$P[(y_1(t), \dots, y_N(t)|y_1(0), \dots, y_N(0))] = \prod_{i=1}^N P[y_i(t)|y_i(0)]$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t)$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t)$$

$$\frac{d\mathbf{P}}{dt} = \mathbf{R}\mathbf{P} \implies \mathbf{P}(t) = e^{\mathbf{R}t}$$

Transition Rates

$$\mathbf{P}(0) = \mathbf{I}$$

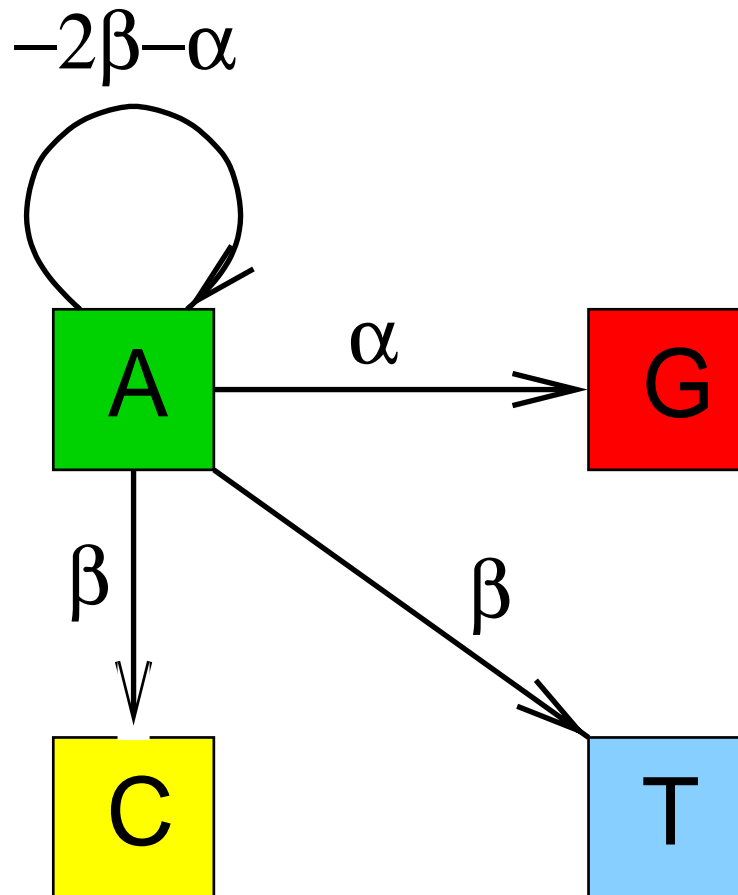
$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

$$\mathbf{P}(t + dt) = \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t)$$

$$\frac{d\mathbf{P}}{dt} = \mathbf{R}\mathbf{P} \implies \mathbf{P}(t) = e^{\mathbf{R}t}$$

$$\mathbf{R} = \begin{bmatrix} -2\beta - \alpha & \beta & \alpha & \beta \\ \beta & -2\beta - \alpha & \beta & \alpha \\ \alpha & \beta & -2\beta - \alpha & \beta \\ \beta & \alpha & \beta & -2\beta - \alpha \end{bmatrix}$$

Transition Rates



Transition Probabilities

$$\mathbf{P}(t) = e^{\mathbf{R}t} = \begin{bmatrix} d(t) & f(t) & g(t) & f(t) \\ f(t) & d(t) & f(t) & g(t) \\ g(t) & f(t) & d(t) & f(t) \\ f(t) & g(t) & f(t) & d(t) \end{bmatrix}$$

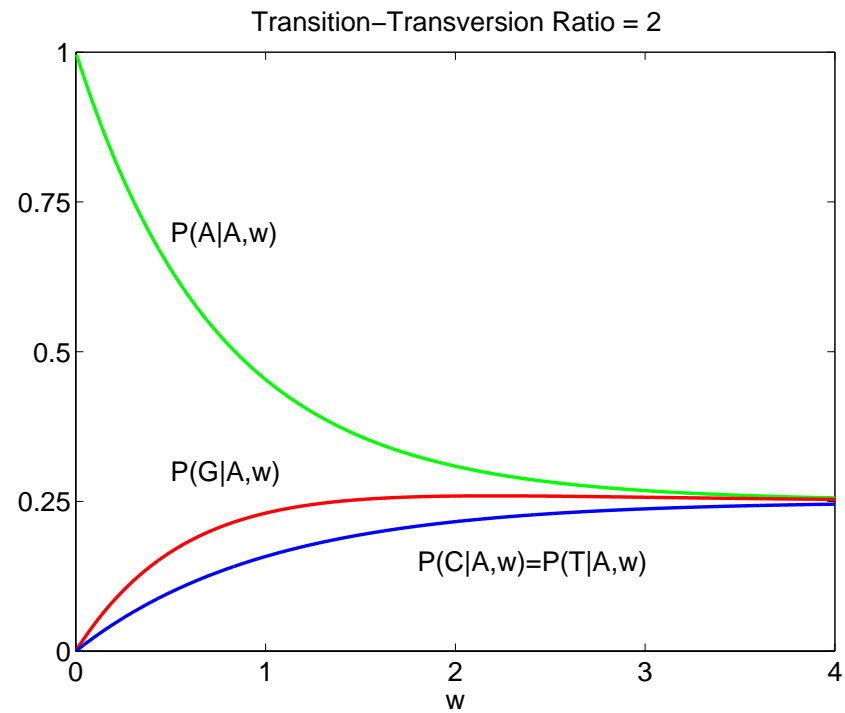
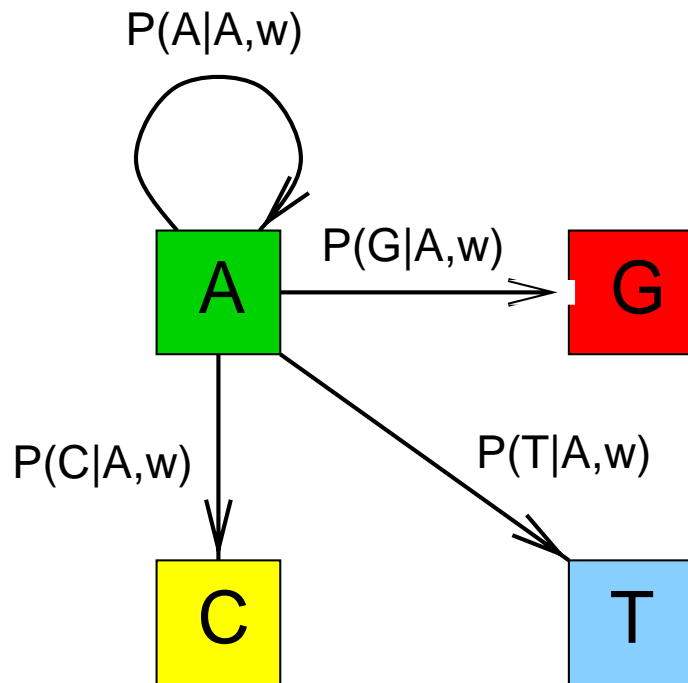
$$f(t) = \frac{1}{4}(1 - e^{-4\beta t}) \quad g(t) = \frac{1}{4}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}) \quad d(t) = 1 - 2f(t) - g(t)$$

Molecular time: $w = 4\beta t$

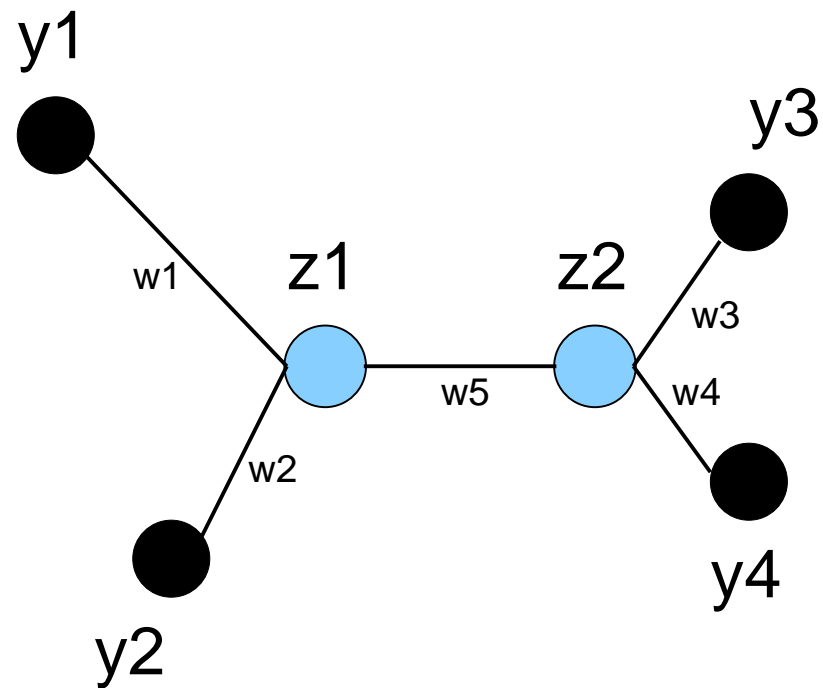
$$\begin{aligned} f(w) &= \frac{1}{4}(1 - e^{-w}) \\ g(w) &= \frac{1}{4}(1 + e^{-w} - 2e^{-\frac{\tau+1}{2}w}) \\ d(w) &= 1 - 2f(w) - g(w) \end{aligned}$$

Transition-transversion ratio: $\tau = \frac{\alpha}{\beta}$

Transition probabilities

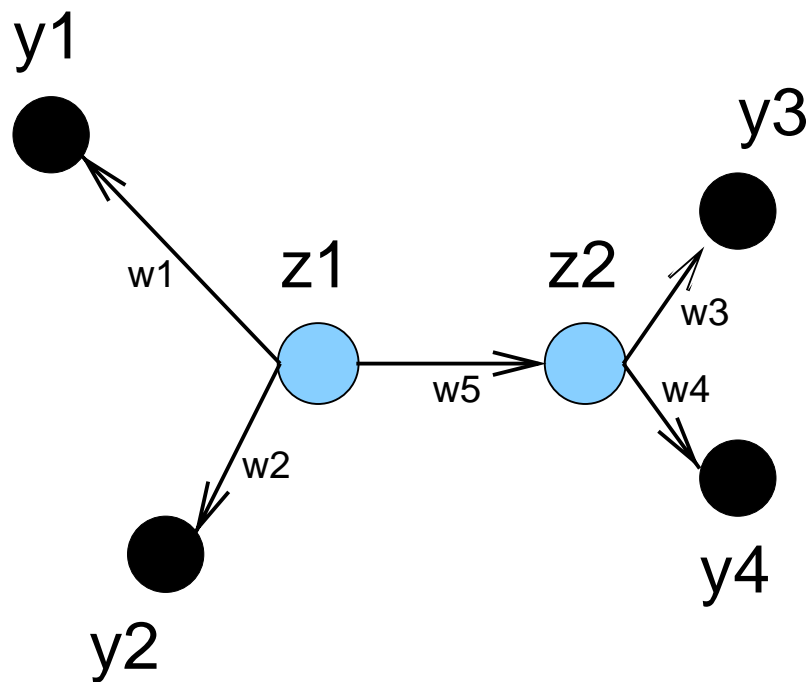


Phylogenetic tree as an undirected graph



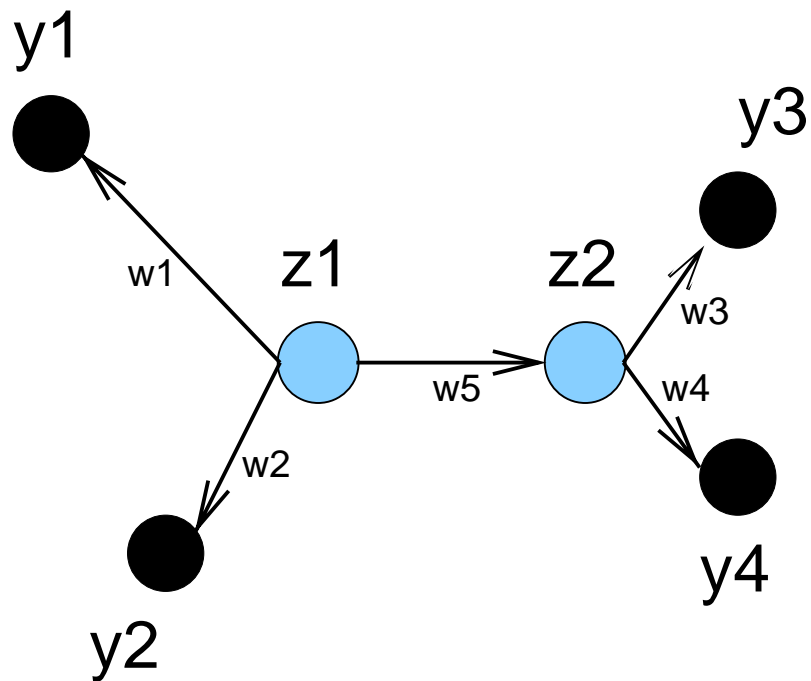
$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

Phylogenetic tree as a directed graph



$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

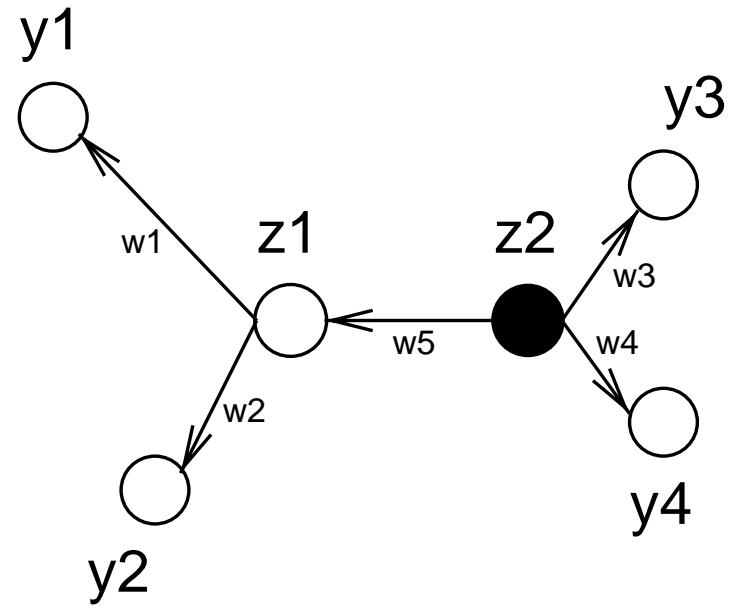
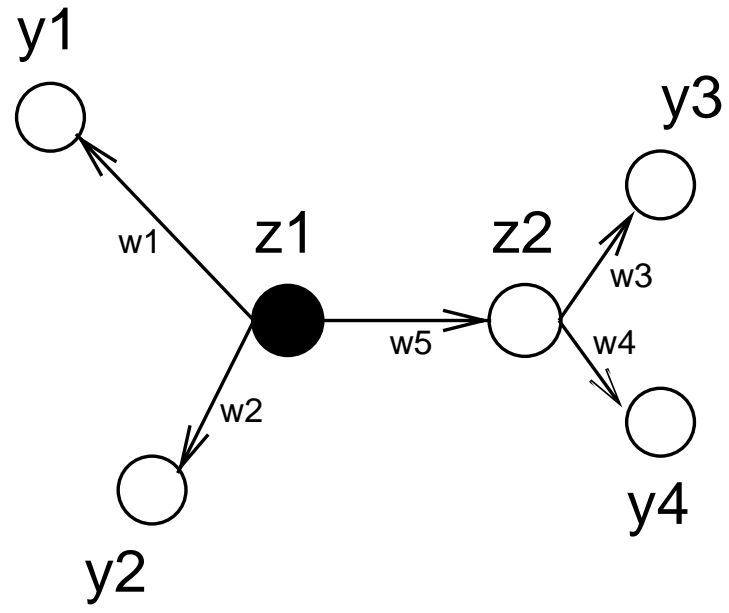
Phylogenetic tree as a directed graph



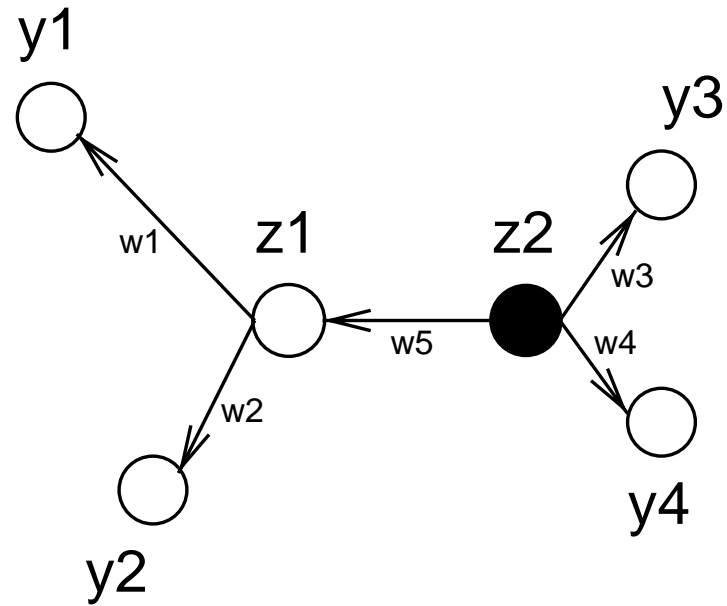
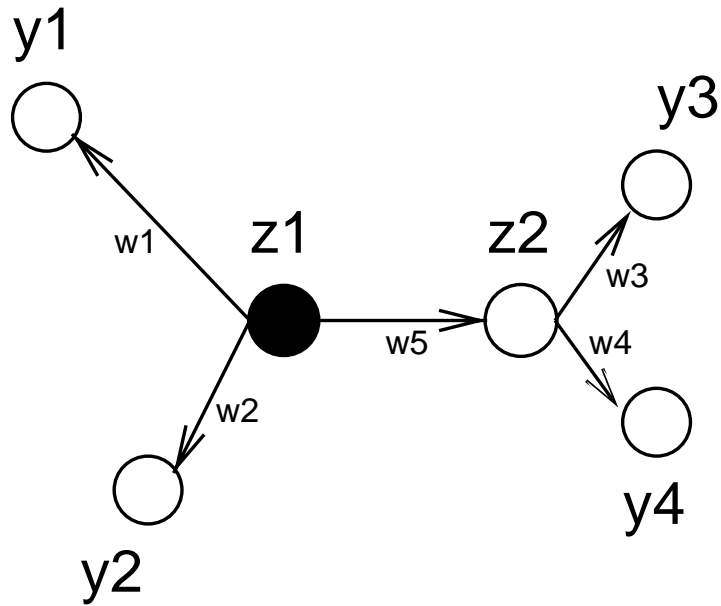
$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1)$$

Different directed graphs

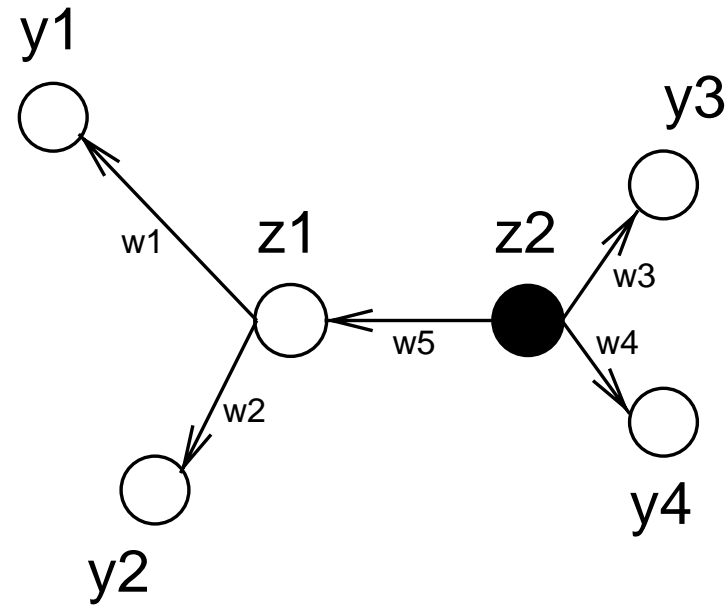
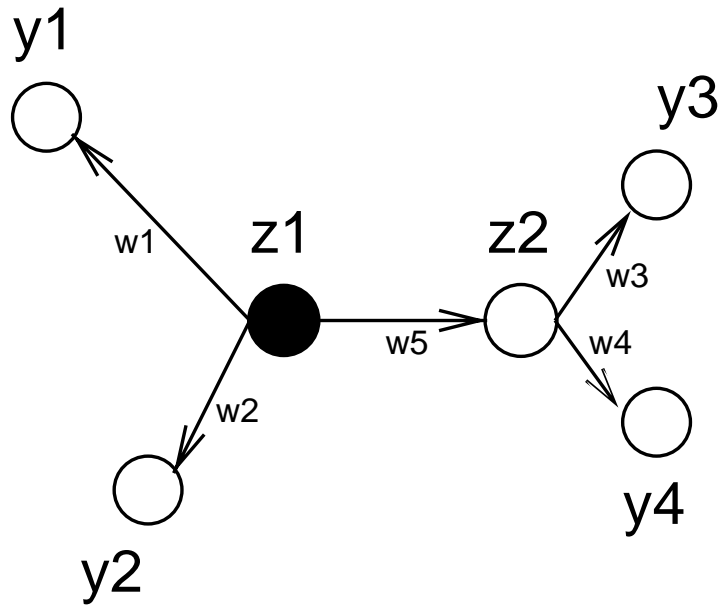


Different directed graphs



Left : $P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) =$
 $P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_2 | z_1, w_5) P(z_1)$

Different directed graphs



Left : $P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) =$
 $P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_2 | z_1, w_5) P(z_1)$

Right : $P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) =$
 $P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1 | z_2, w_5) P(z_2)$

Reversibility

We can *not* decide on the direction of evolutionary processes.

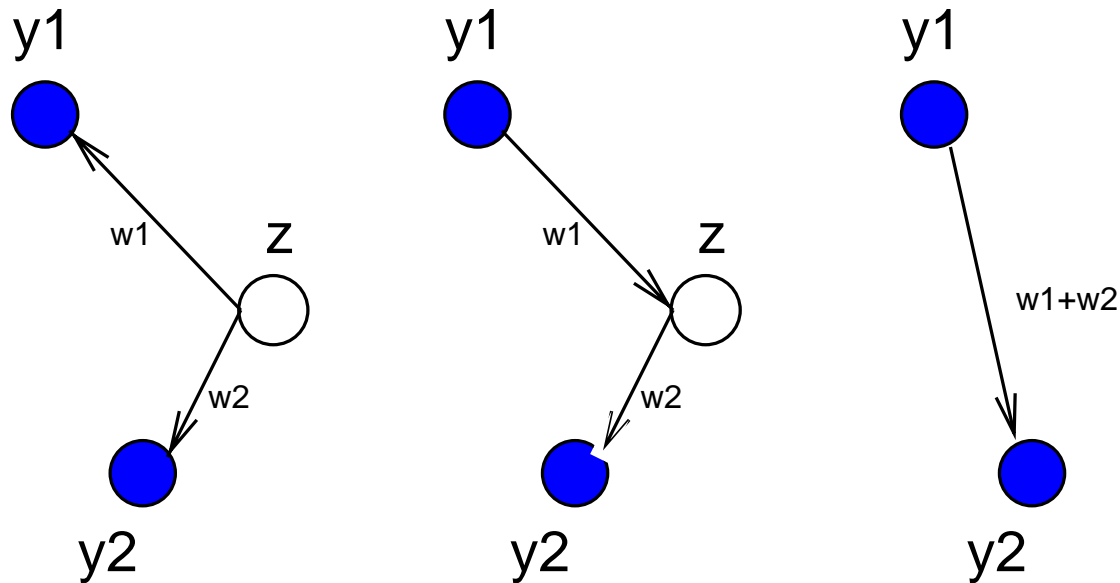
$$P(z_1|z_2, w_5)P(z_2) = P(z_2|z_1, w_5)P(z_1)$$

- Changing the position of the root and the direction of the arcs does not affect the probability.
- All directed graphs are in the same equivalence class.

Root elimination

Homogeneous Markov chain \implies Multiplicativity of the substitution matrices:

$$\mathbf{P}(w_1)\mathbf{P}(w_2) = \mathbf{P}(w_1 + w_2)$$

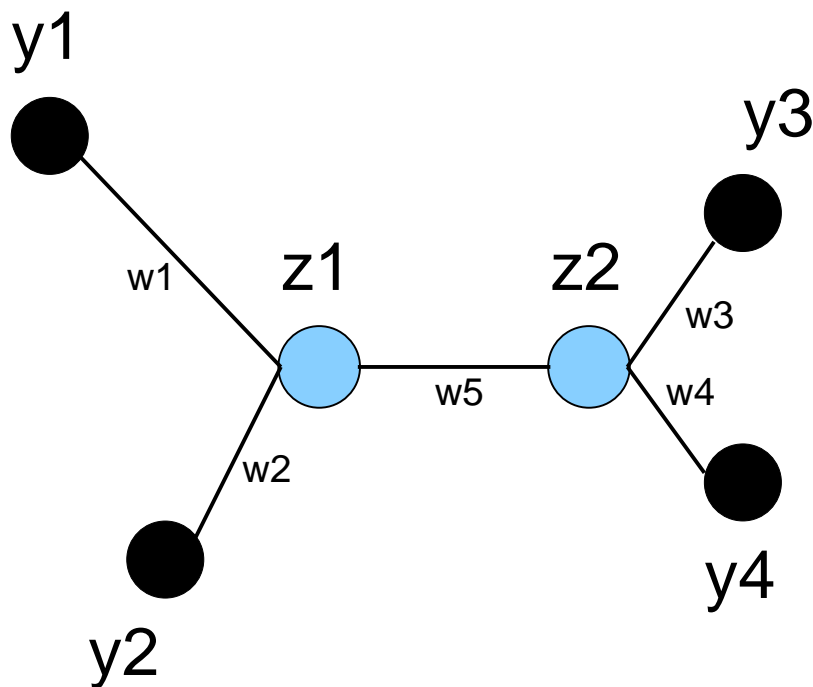


$$\sum_z P(y_2|z, w_2)P(y_1|z, w_1)P(z) \quad \sum_z P(y_2|z, w_2)P(z|y_1, w_1)P(y_1) \quad P(y_2|y_1, w_1+w_2)P(y_1)$$

Reversibility \implies left = middle

Multiplicativity \implies middle = right

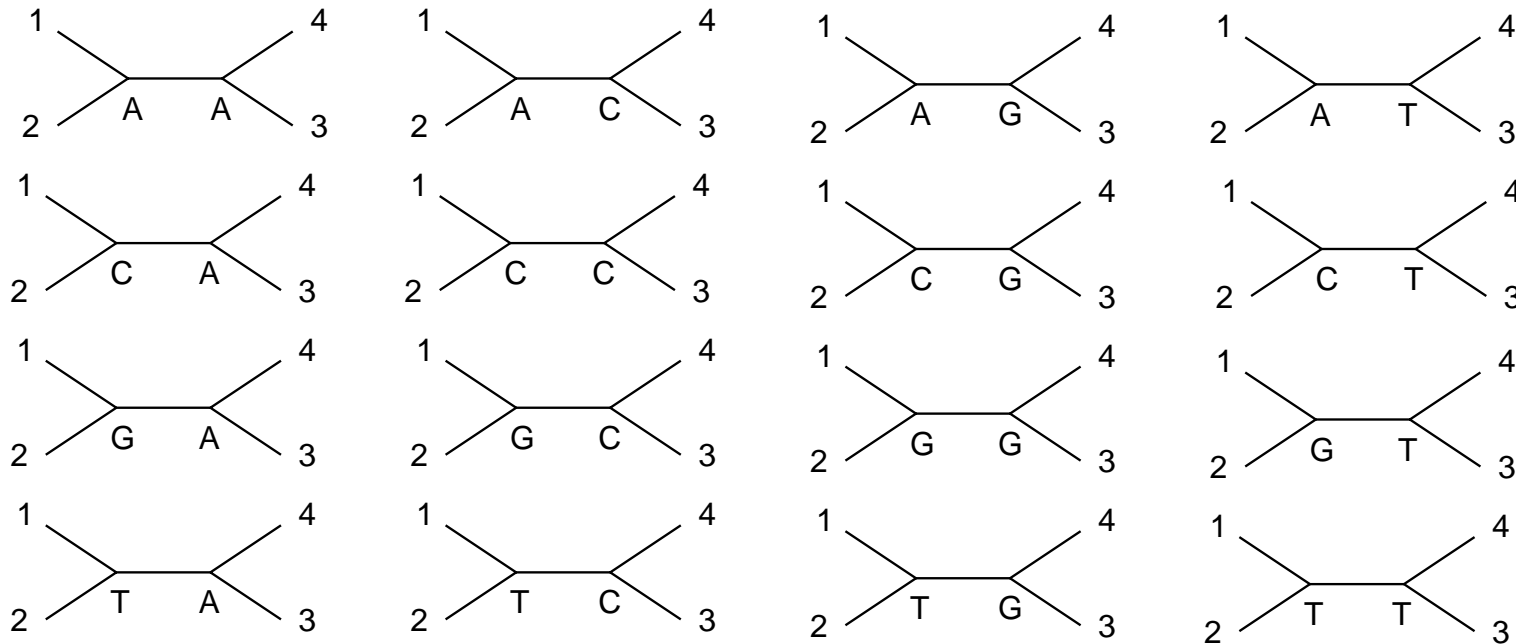
Expansion of the joint probability



$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

$$= P(y_1 | z_1, w_1) P(y_2 | z_1, w_2) P(z_2 | z_1, w_5) P(y_3 | z_2, w_3) P(y_4 | z_2, w_4) P(z_1)$$

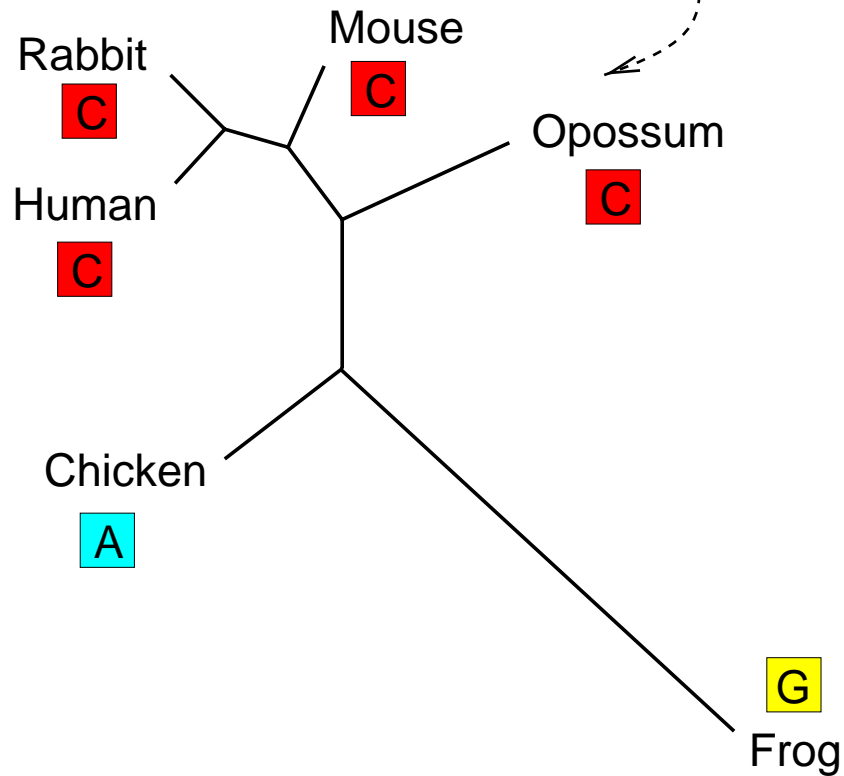
Marginalisation



$$P(y_1, y_2, y_3, y_4 | \mathbf{w}) = \sum_{z_1} \sum_{z_2} P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w})$$

↓

Frog	G	C	T	T	G	A	C	T	T	C	T	G	A	G	G	T	T
Chicken	G	C	G	T	A	A	C	T	T	C	A	C	A	T	G	A	T
Human	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Rabbit	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T
Mouse	G	C	G	T	C	A	C	T	T	G	A	C	A	G	G	C	T
Opossum	G	C	G	T	C	A	C	T	T	G	A	G	A	C	G	C	T



--> Likelihood

Topology
Branch lengths

Maximum likelihood



Heads

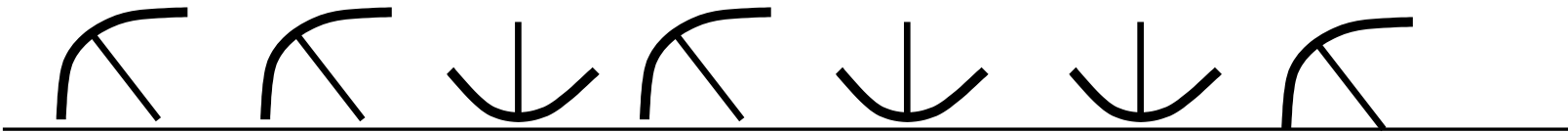
Tails

Probability

θ

$1-\theta$

Data



N tosses, k observations of "heads"

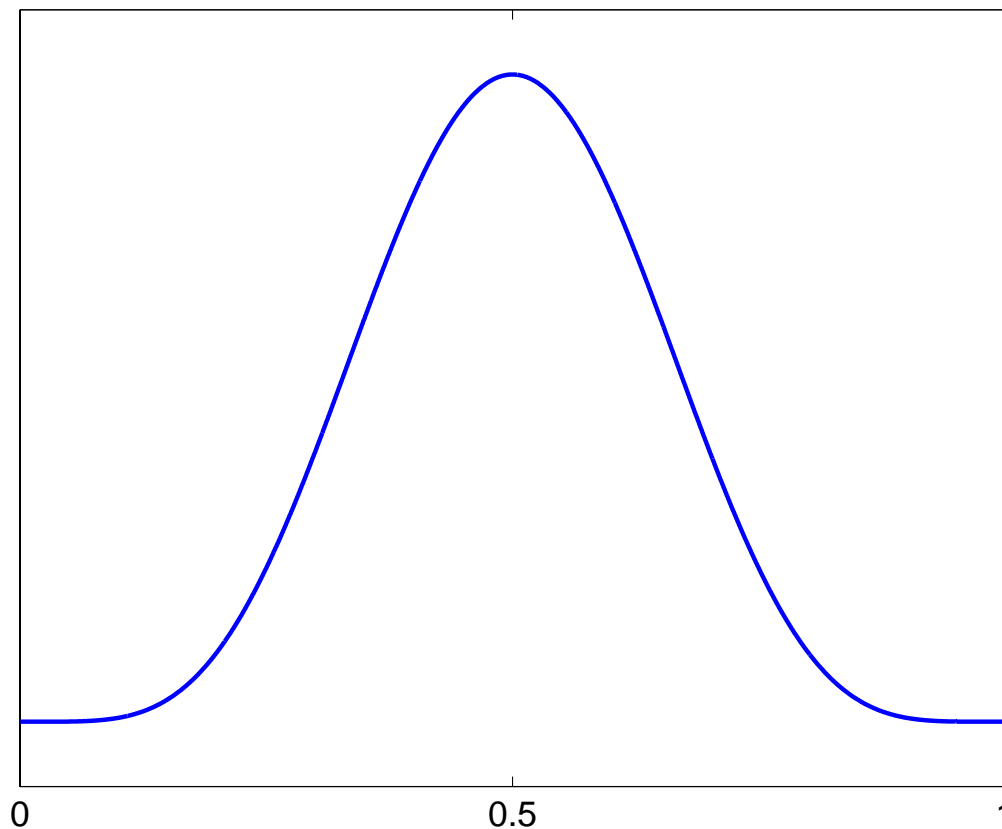
$$P(D|\theta) = \theta^k (1 - \theta)^{N-k} \binom{N}{k}$$

$$\log P(D|\theta) = k \log \theta + (N - k) \log(1 - \theta) + C$$

$$P(D|\theta) = \theta^k (1 - \theta)^{N-k} \binom{N}{k}$$

$$\log P(D|\theta) = k \log \theta + (N - k) \log(1 - \theta) + C$$

Example: $\log P(D|\theta)$ for equal numbers of heads and tails and $N = 10$



$$\log P(D|\theta) = k \log \theta + (N - k) \log(1 - \theta) + C$$

$$\frac{d}{d\theta} \log P(D|\theta) = \frac{k}{\theta} - \frac{N-k}{1-\theta} = 0$$

$$k(1 - \theta) = (N - k)\theta$$

$$k = N\theta$$

$$\theta = \frac{k}{N}$$

$$\log P(D|\theta) = k \log \theta + (N - k) \log(1 - \theta) + C$$

$$\frac{d}{d\theta} \log P(D|\theta) = \frac{k}{\theta} - \frac{N-k}{1-\theta} = 0$$

$$k(1 - \theta) = (N - k)\theta$$

$$k = N\theta$$

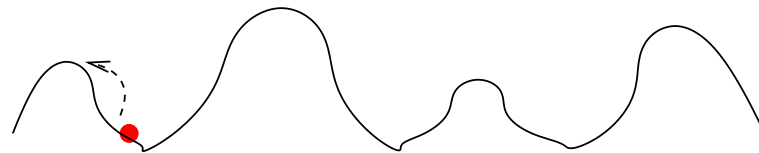
$$\theta = \frac{k}{N}$$

Apply the same idea in phylogenetics

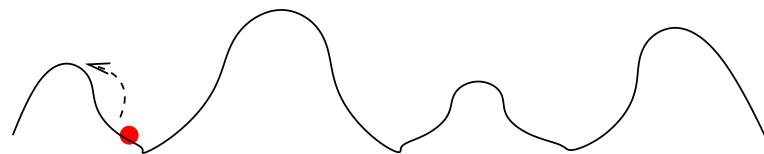
- Find tree topology S and vector of branch lengths \mathbf{w} that maximize likelihood $\log P(D|S, \mathbf{w})$.

- Find tree topology S and vector of branch lengths \mathbf{w} that maximize likelihood $\log P(D|S, \mathbf{w})$.
- No analytic solution.
- Find maximum in a high-dimensional space with a heuristic hill climbing method.

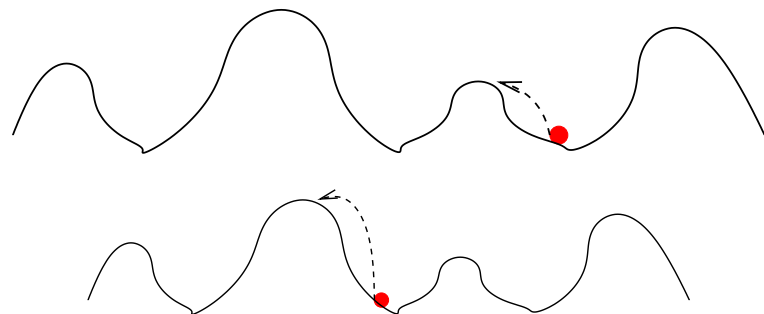
- Find tree topology S and vector of branch lengths \mathbf{w} that maximize likelihood $\log P(D|S, \mathbf{w})$.
- No analytic solution.
- Find maximum in a high-dimensional space with a heuristic hill climbing method.
- Given topology S , optimise branch lengths \mathbf{w} by gradient ascent: $\Delta \mathbf{w} \propto \nabla \log P(D|S, \mathbf{w})$



- Find tree topology S and vector of branch lengths \mathbf{w} that maximize likelihood $\log P(D|S, \mathbf{w})$.
- No analytic solution.
- Find maximum in a high-dimensional space with a heuristic hill climbing method.
- Given topology S , optimise branch lengths \mathbf{w} by gradient ascent: $\Delta \mathbf{w} \propto \nabla \log P(D|S, \mathbf{w})$



- Repeat for different tree topologies S .



NP hard problem

- For M taxa, there are $(2M - 5)!!$ unrooted trees.
- $M = 4 \longrightarrow 3!! = 3$
- $M = 6 \longrightarrow 7!! = 7 \times 5 \times 3 = 105$
- $M = 10 \longrightarrow \approx 2 \times 10^6$
- $M = 20 \longrightarrow \approx 2 \times 10^{20}$

NP hard problem

- For M taxa, there are $(2M - 5)!!$ unrooted trees.
- $M = 4 \longrightarrow 3!! = 3$
- $M = 6 \longrightarrow 7!! = 7 \times 5 \times 3 = 105$
- $M = 10 \longrightarrow \approx 2 \times 10^6$
- $M = 20 \longrightarrow \approx 2 \times 10^{20}$
- M large \longrightarrow Exhaustive search impossible
- Heuristic search methods

Heuristic search methods

- DNAML (Felsenstein)
- Tree-Puzzle (Strimmer et al.)

DNAML

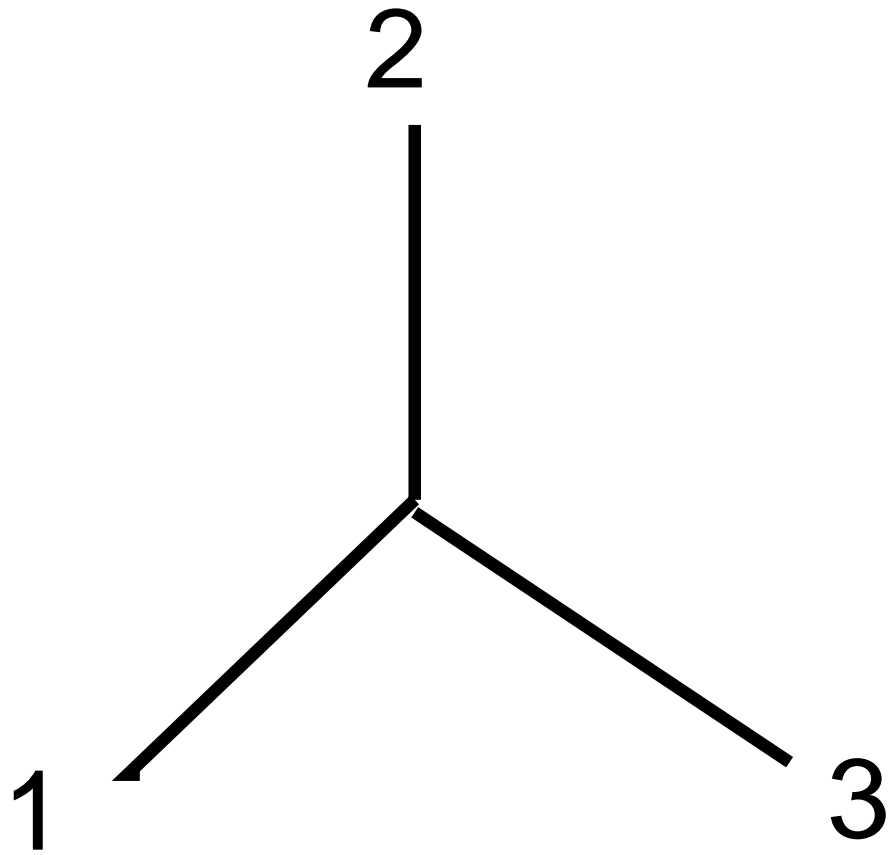
-
- J. Felsenstein
 - Implemented in the software package **PHYLIP**
 - Available from <http://evolution.genetics.washington.edu/phylip.html>

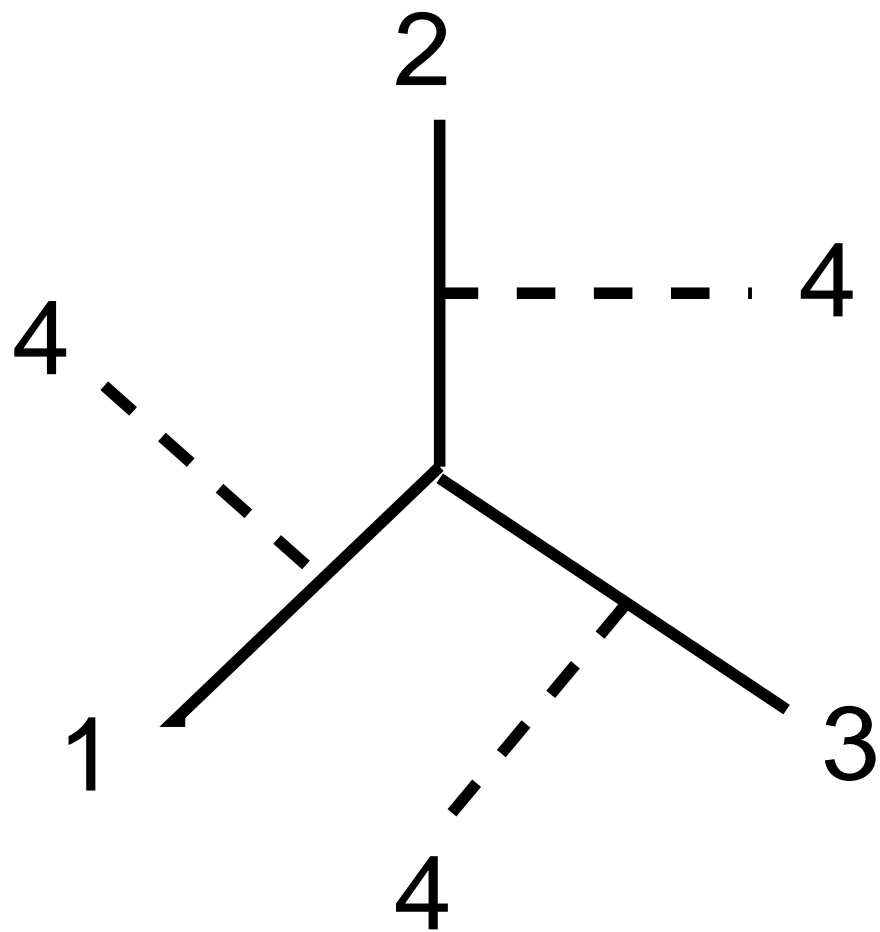
DNAML

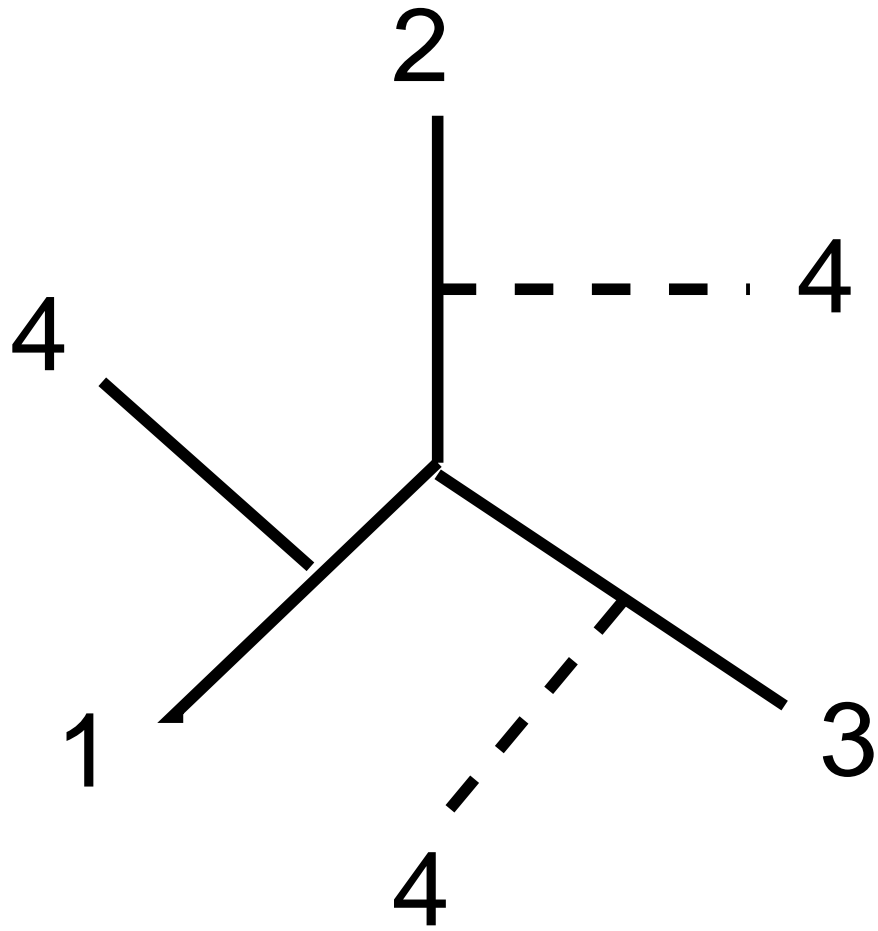
- J. Felsenstein
- Implemented in the software package **PHYLIP**
- Available from <http://evolution.genetics.washington.edu/phylip.html>

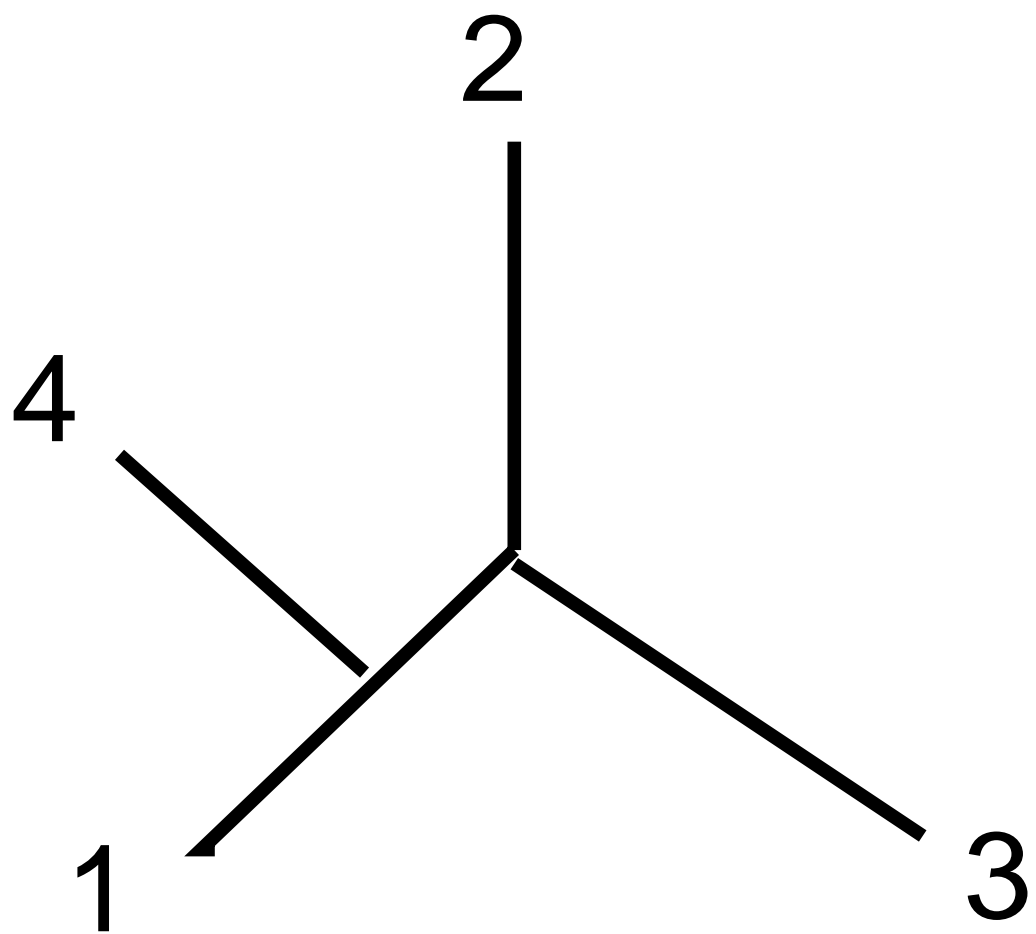
Idea:

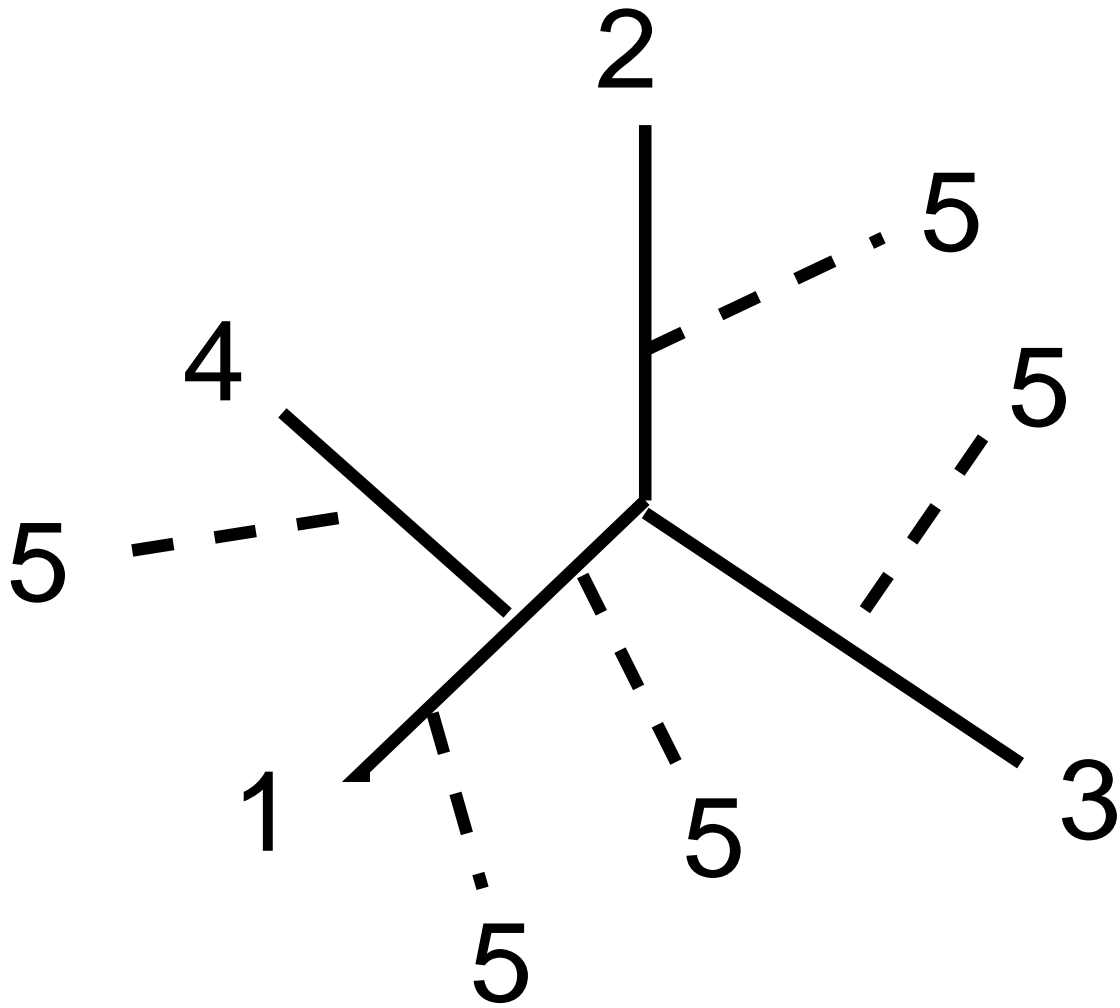
- Stepwise addition of taxa.
- Branch swapping.
- Repeat for different sequence orders.

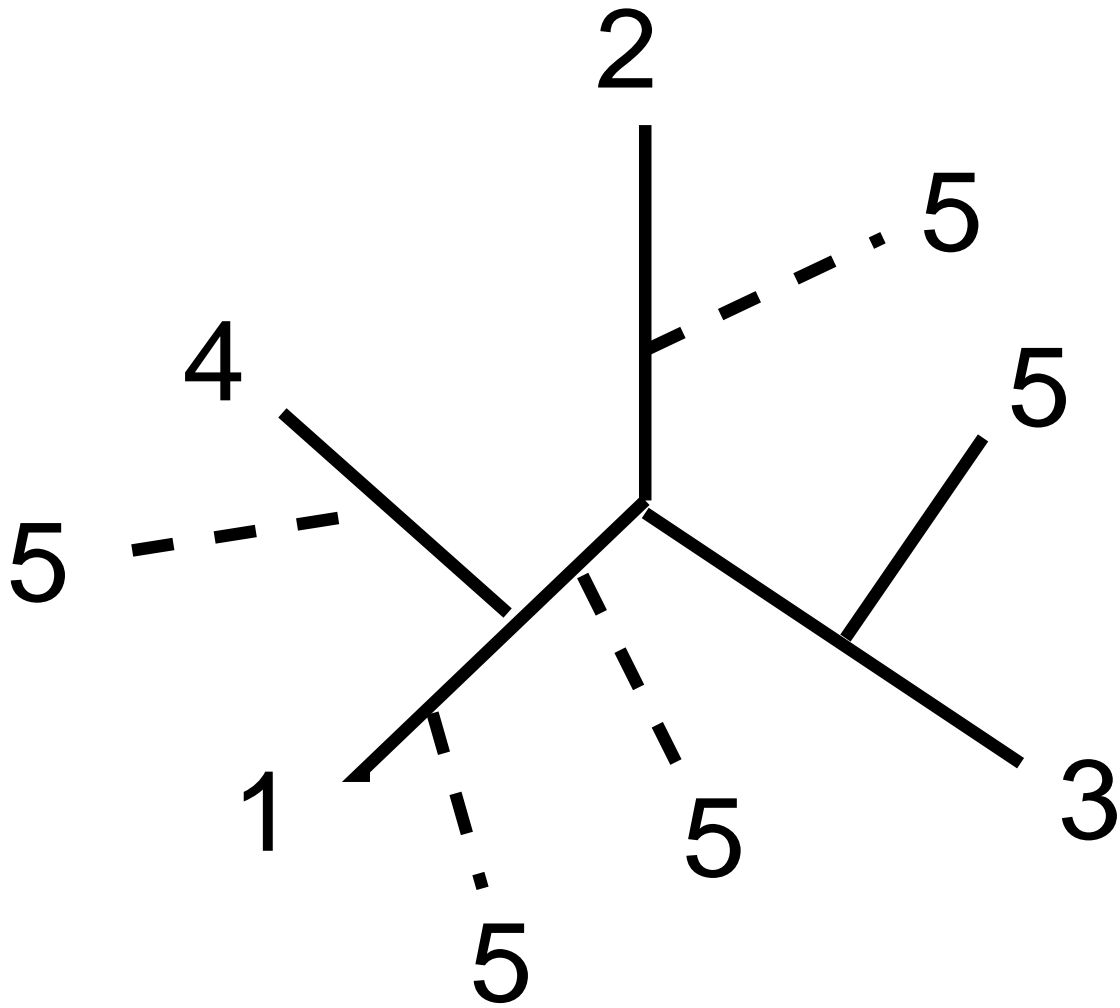


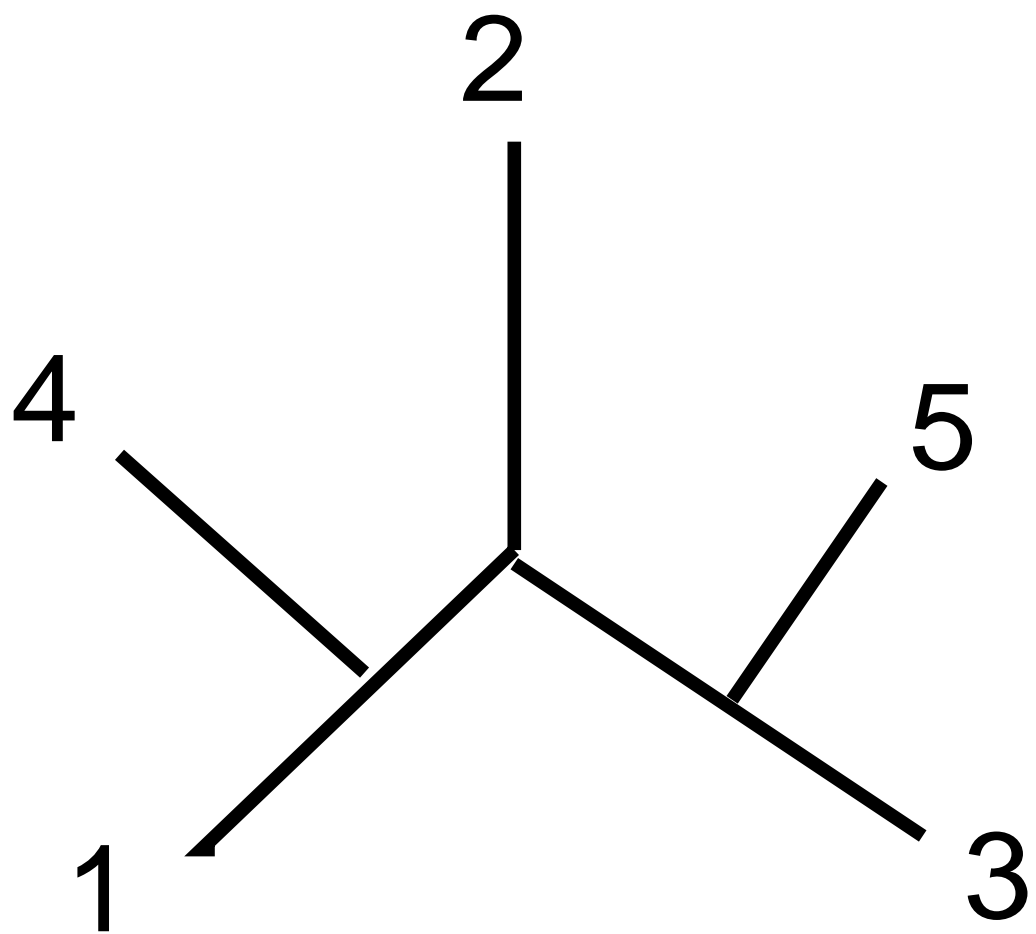


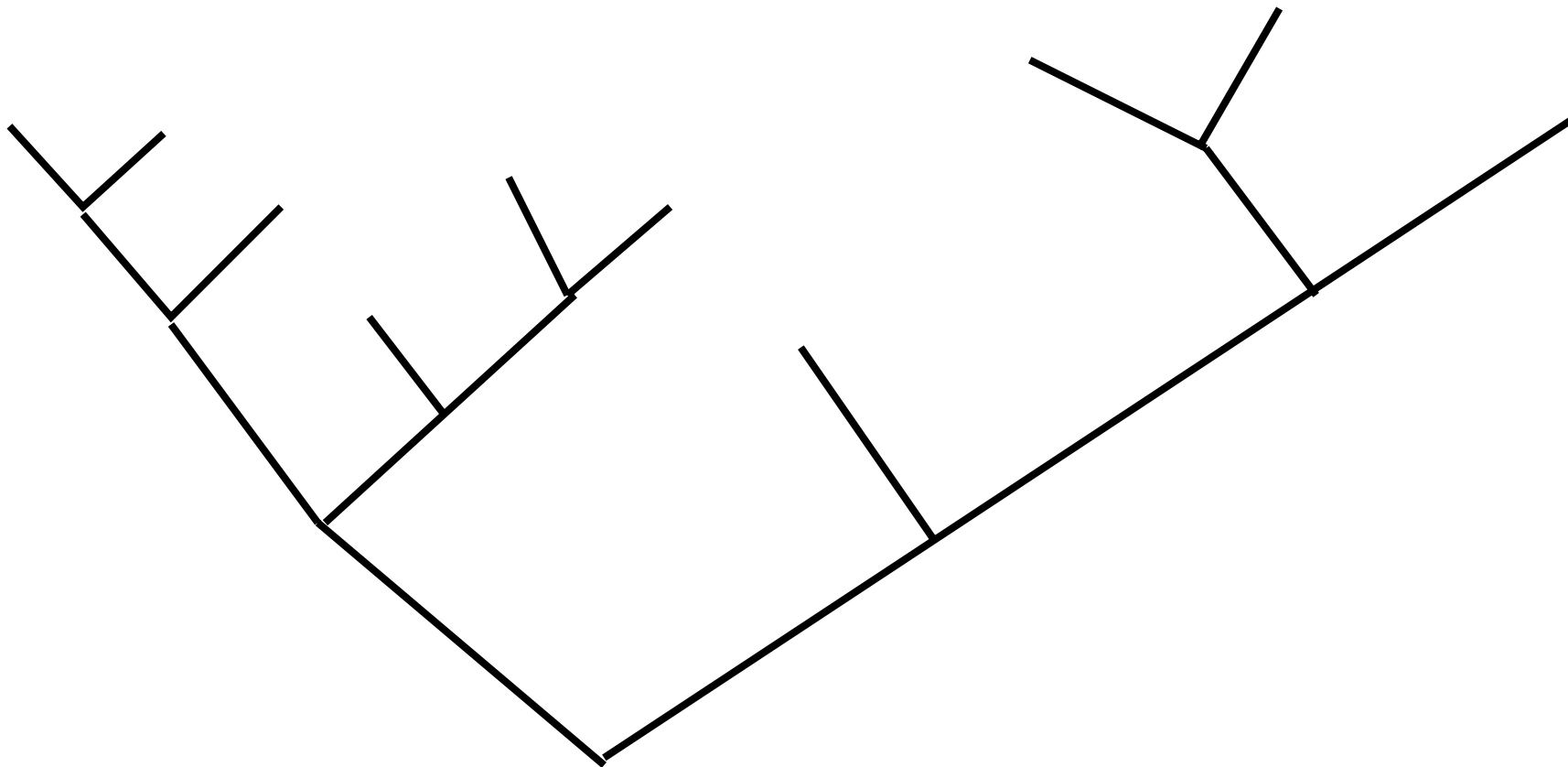




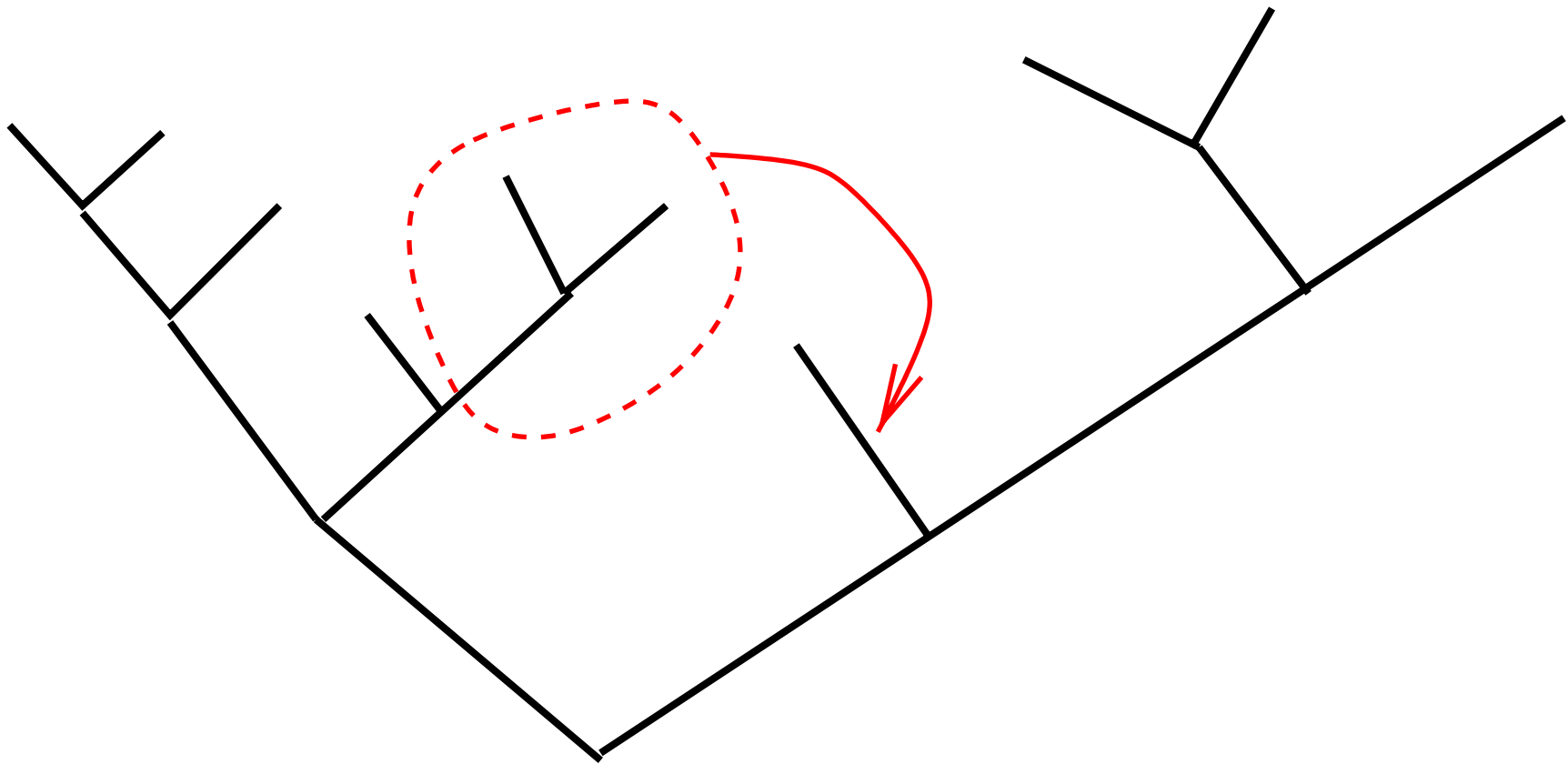




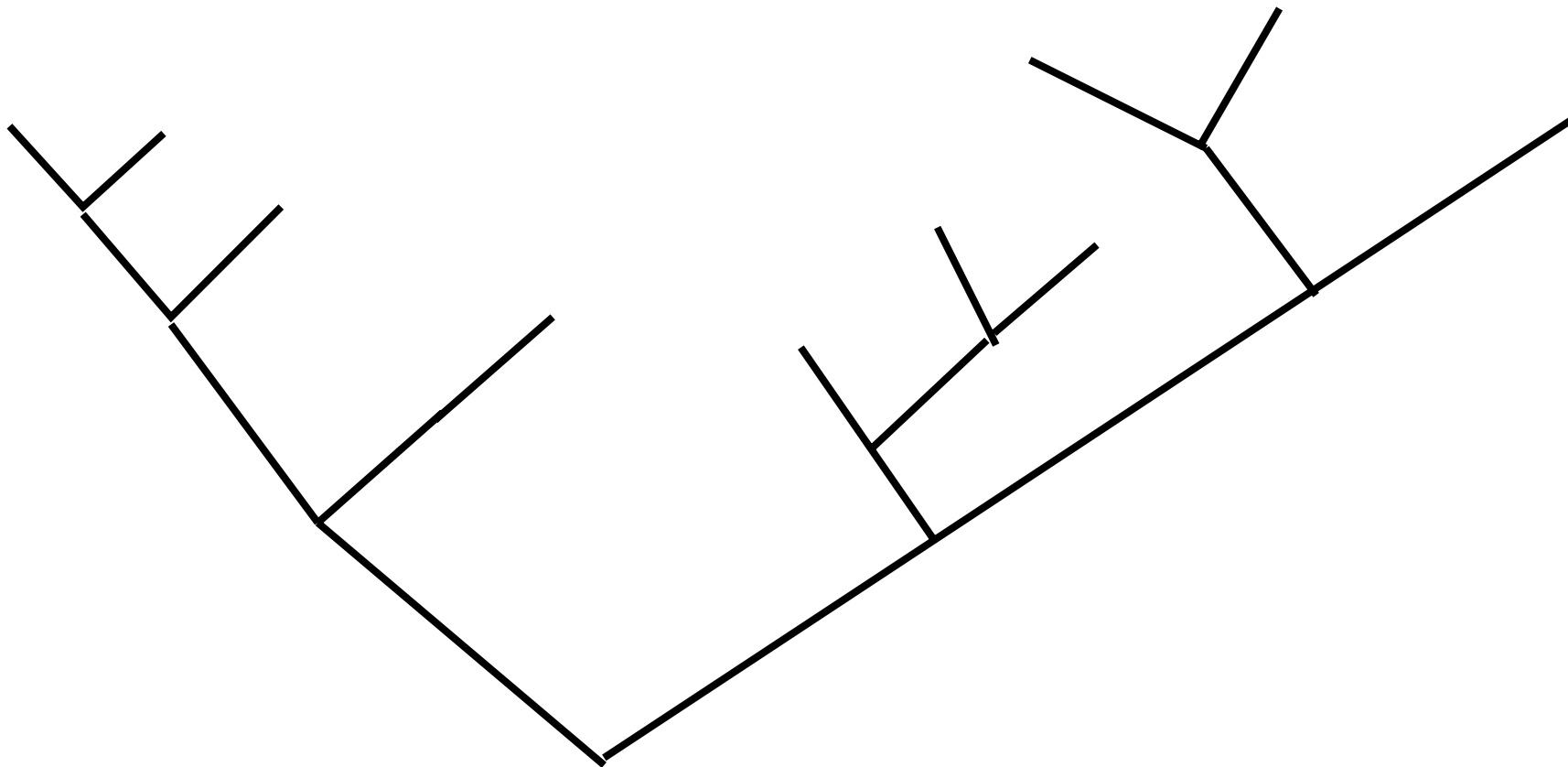




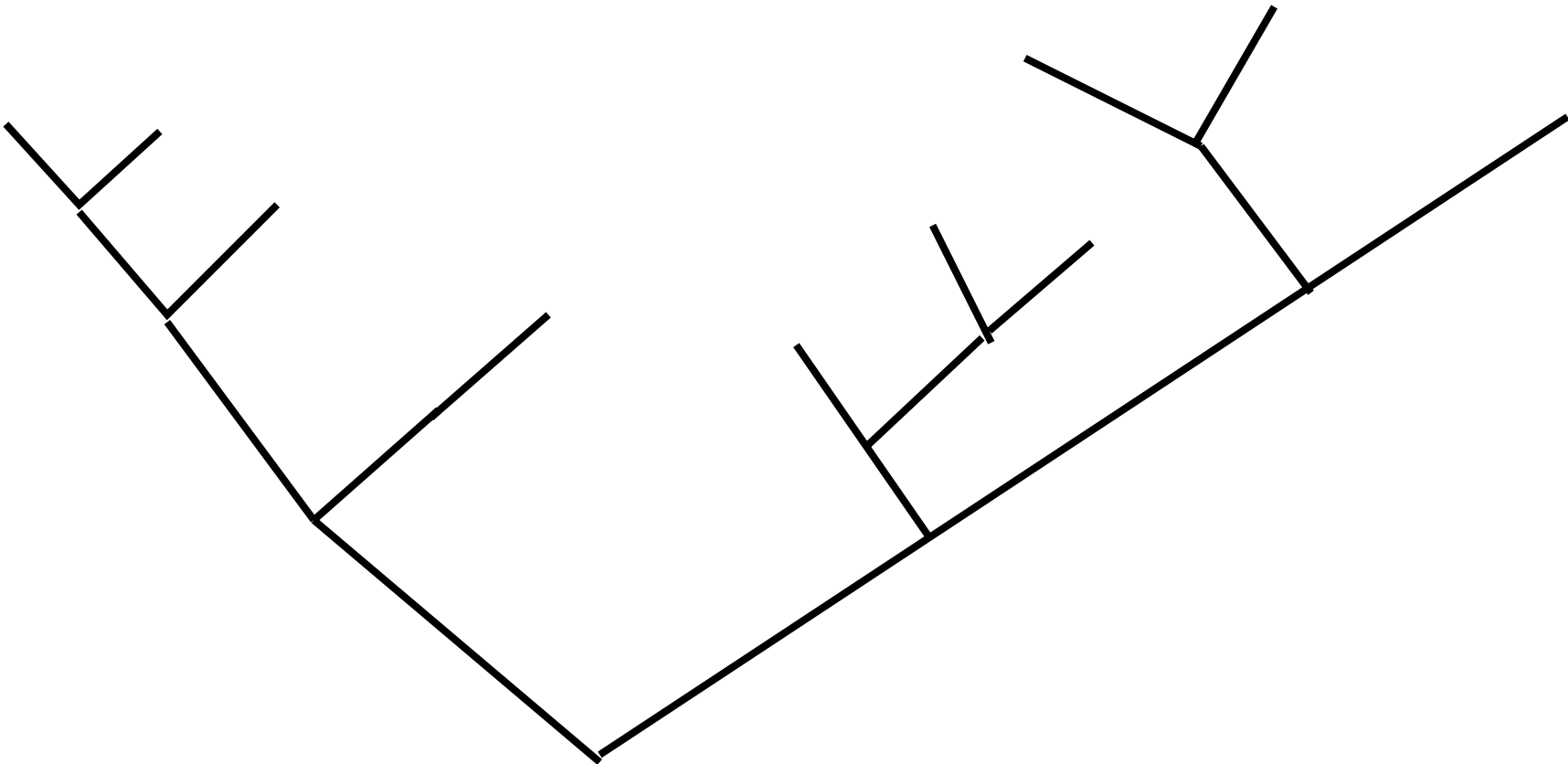
..



..



..



..

Disadvantage: Computational expensive

Heuristic search methods

- DNAML (Felsenstein)
- Tree-Puzzle (Strimmer et al.)

Tree Puzzle

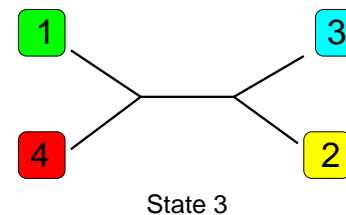
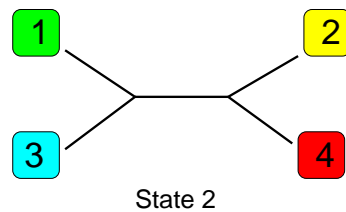
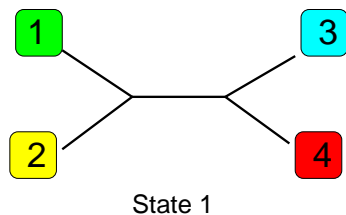
- Strimmer, Haeseler 1996
Molecular Biology and Evolution 13
- Strimmer, Goldman, Haeseler 1997
Molecular Biology and Evolution 14
- Implemented in the program **TREE-PUZZLE**,
<http://www.tree-puzzle.de/>

Tree Puzzle

- Strimmer, Haeseler 1996
Molecular Biology and Evolution 13
- Strimmer, Goldman, Haeseler 1997
Molecular Biology and Evolution 14
- Implemented in the program **TREE-PUZZLE**,
<http://www.tree-puzzle.de/>

Idea:

Maximum likelihood is difficult for $n \gg 1$ taxa,
but easy for $n = 4$ taxa.



Tree puzzle algorithm

Consider an alignment of n taxa.

1. Maximum likelihood step

Reconstruct all possible quartet trees with maximum likelihood: $\longrightarrow \binom{n}{4}$ different trees.

Tree puzzle algorithm

Consider an alignment of n taxa.

1. Maximum likelihood step

Reconstruct all possible quartet trees with maximum likelihood: $\longrightarrow \binom{n}{4}$ different trees.

2. Puzzling step

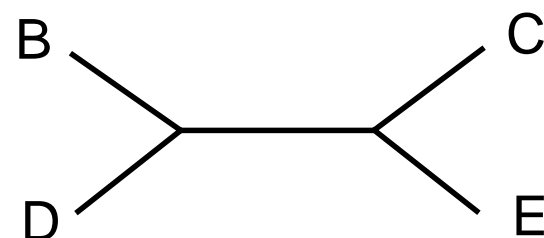
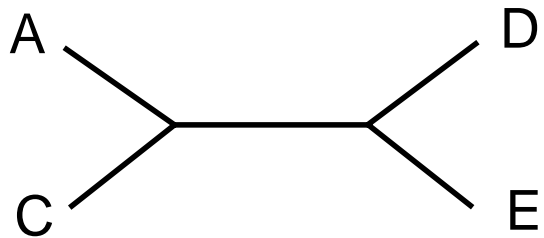
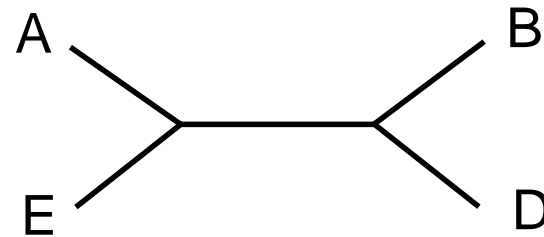
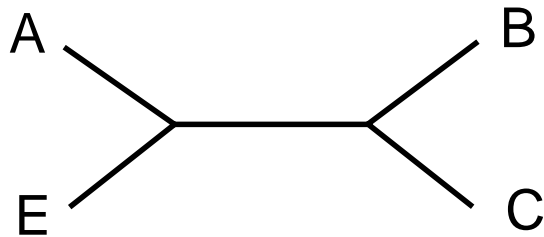
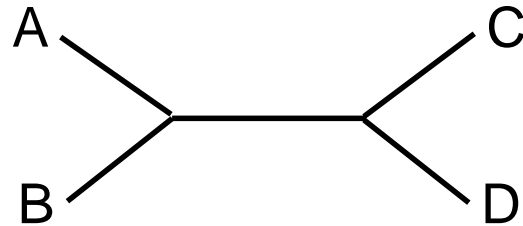
Combine the quartet trees to an overall tree

- The resulting tree depends on the order with which sequences are presented.
- Repeat for different orders of taxa.

Example: 5 taxa A,B,C,D,E

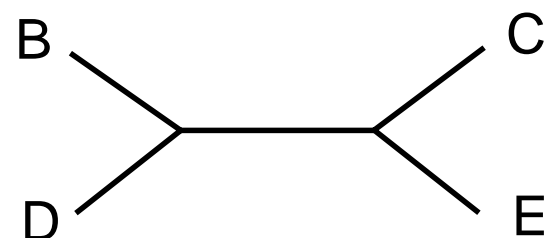
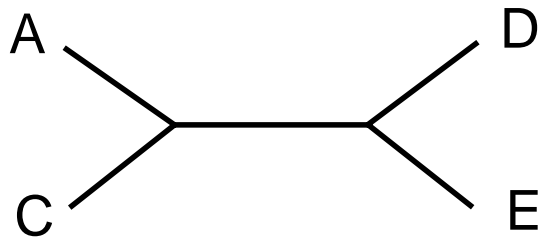
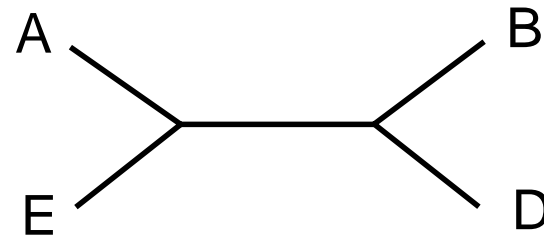
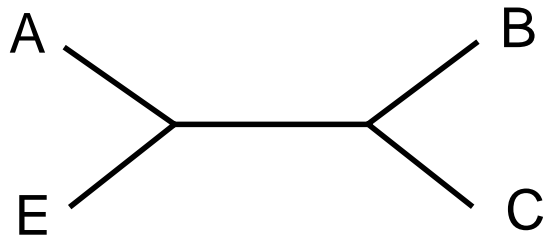
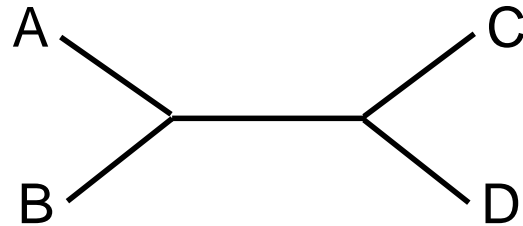
Example: 5 taxa A,B,C,D,E

First step: Reconstruct all possible quartet maximum likelihood trees

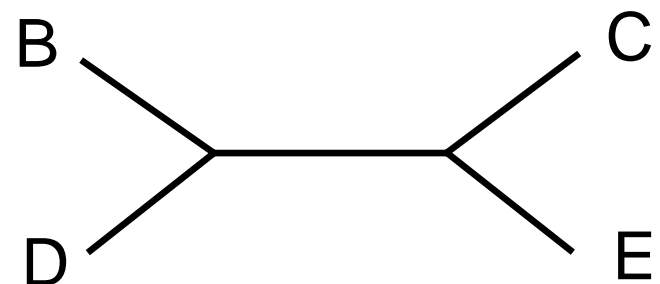
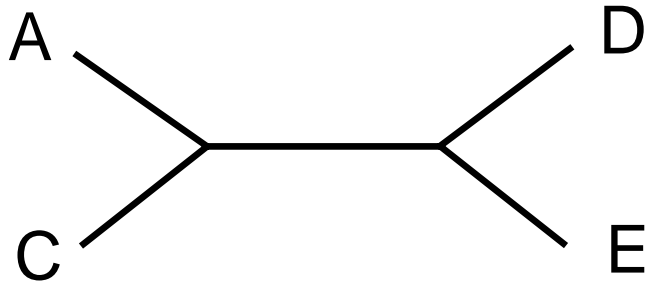
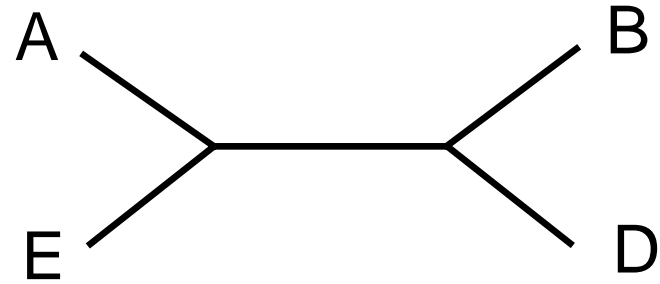
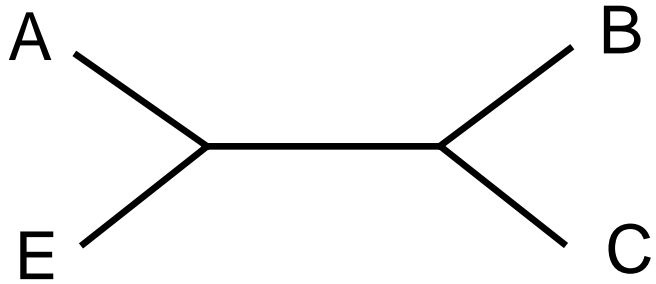
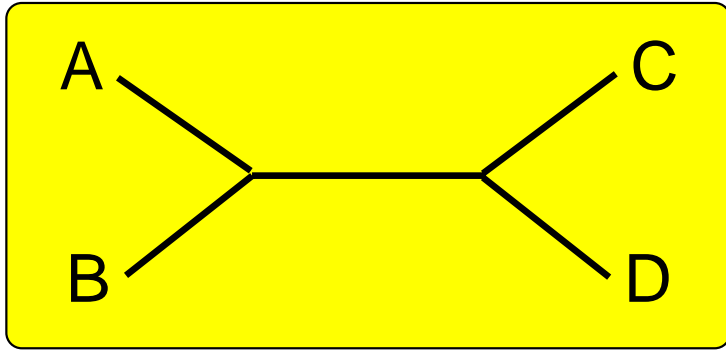


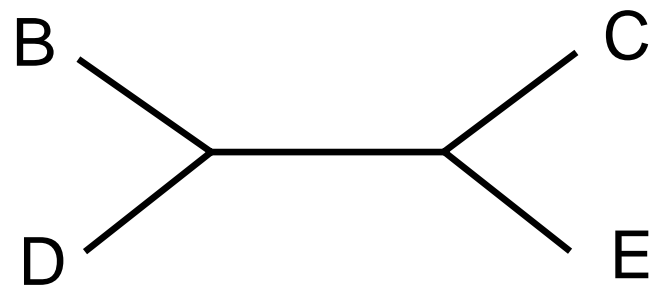
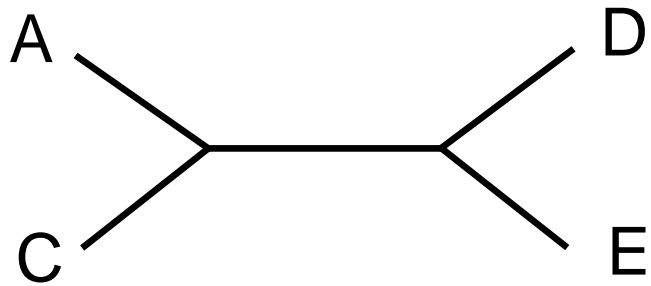
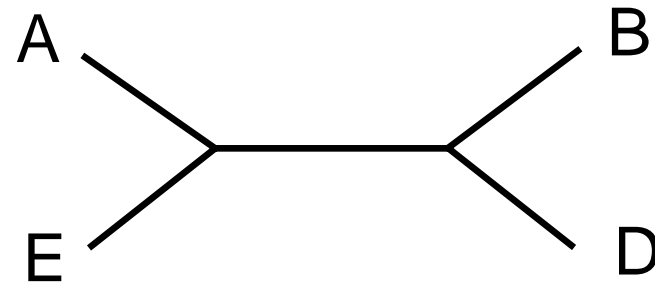
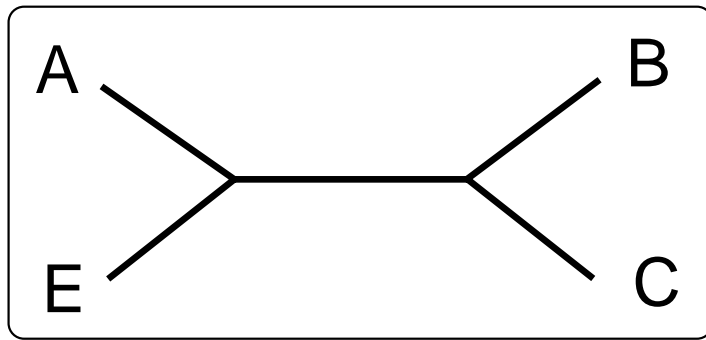
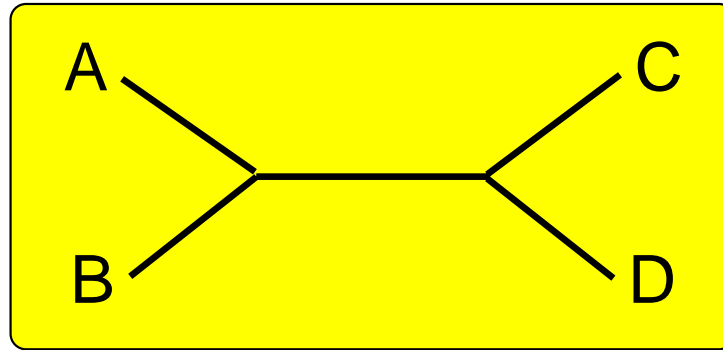
Example: 5 taxa A,B,C,D,E

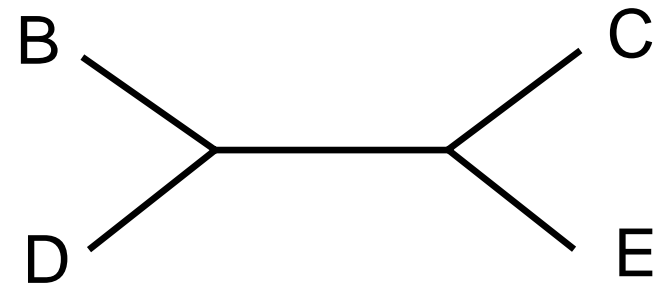
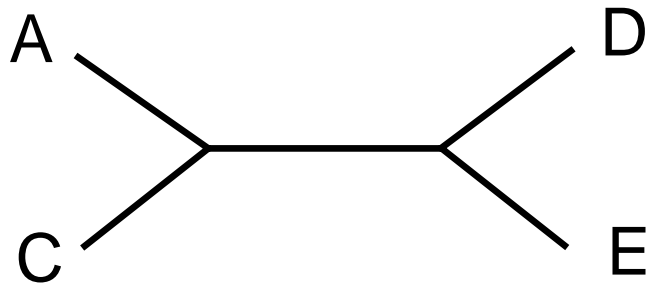
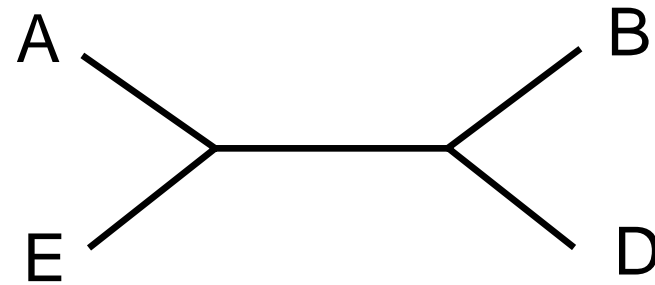
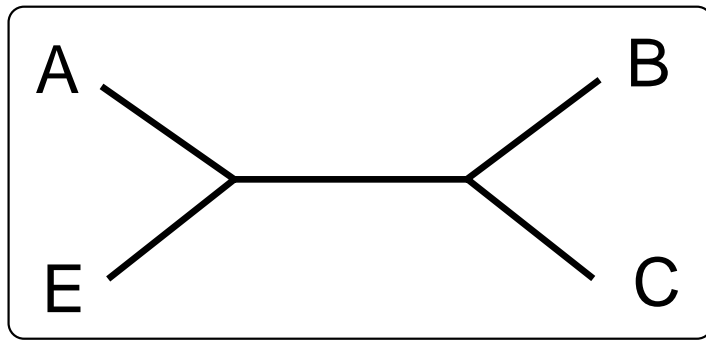
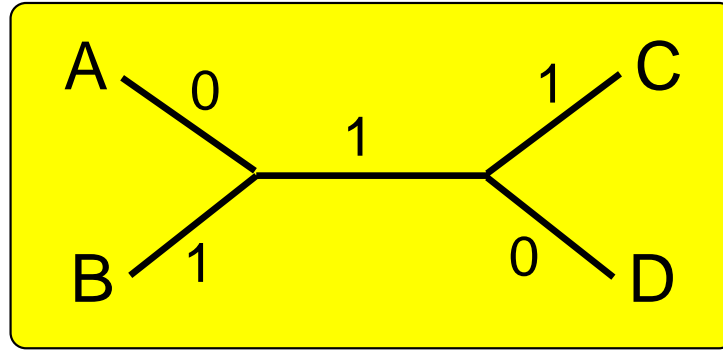
First step: Reconstruct all possible quartet maximum likelihood trees

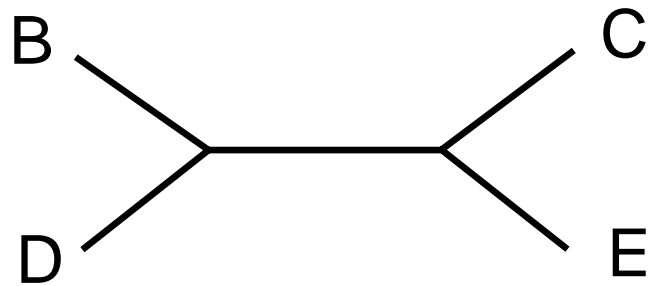
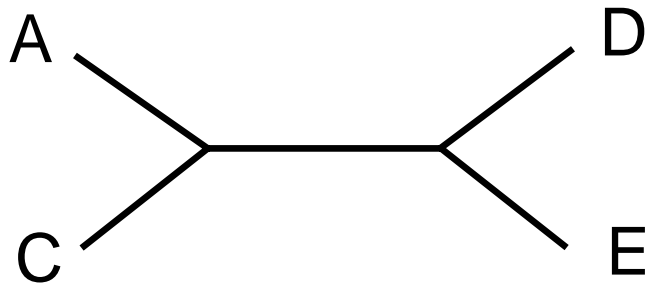
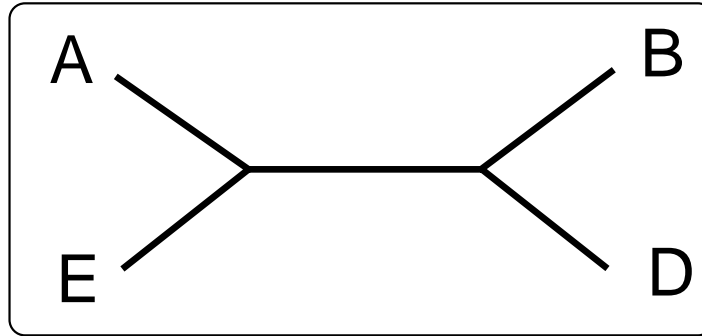
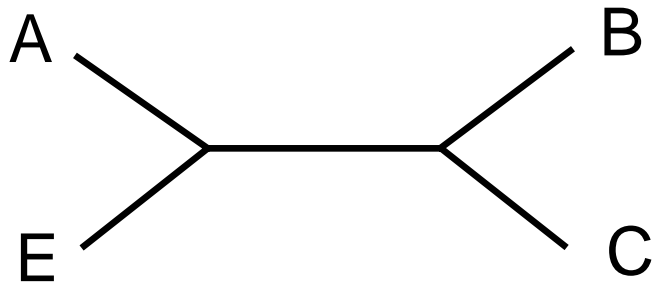
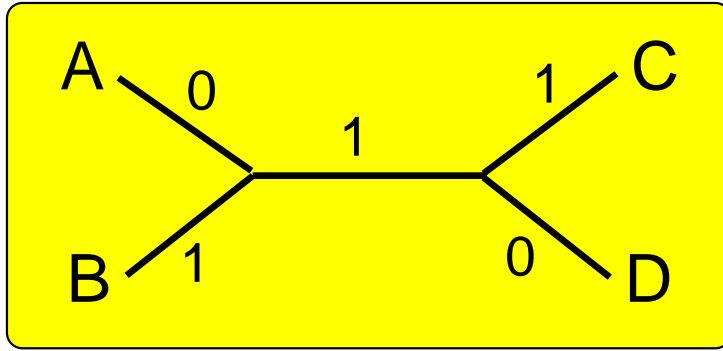


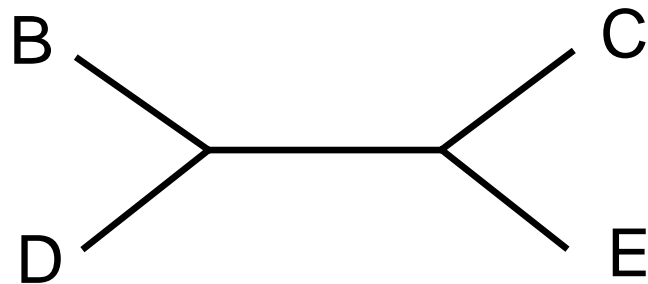
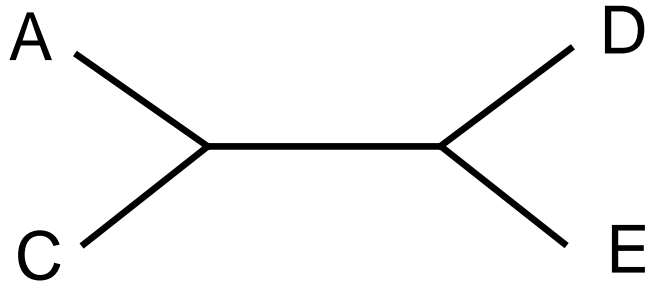
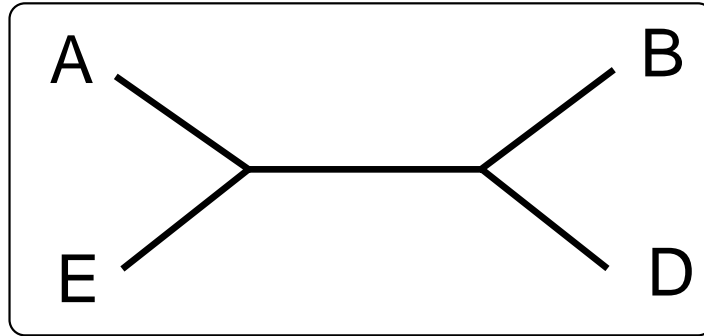
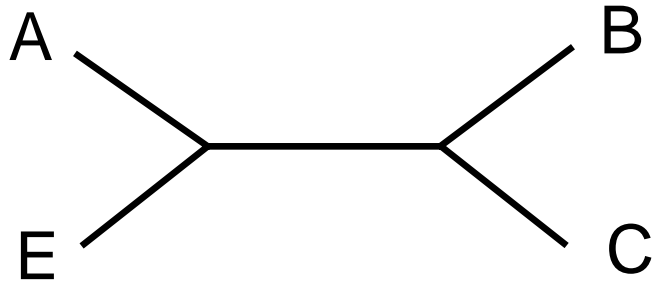
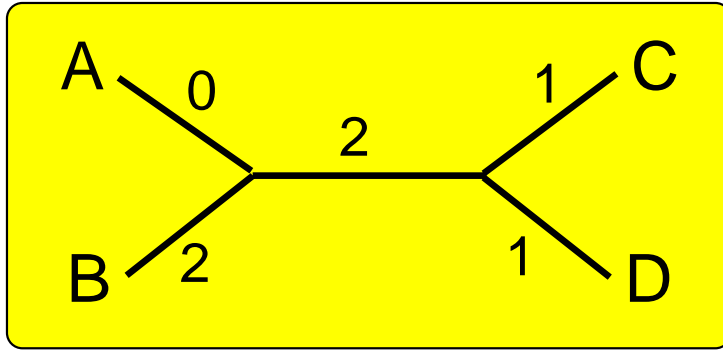
Second step: Puzzling step, for order $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$

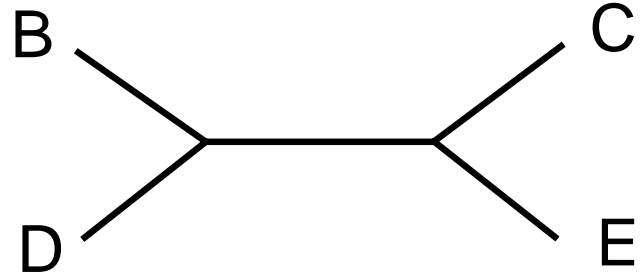
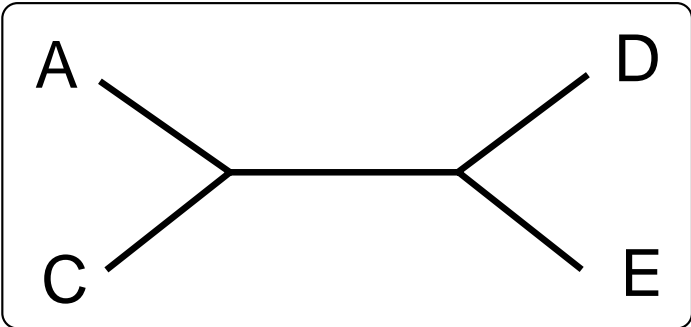
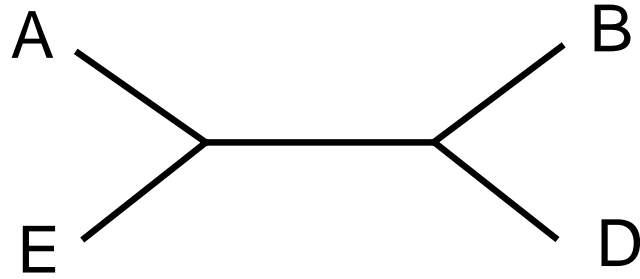
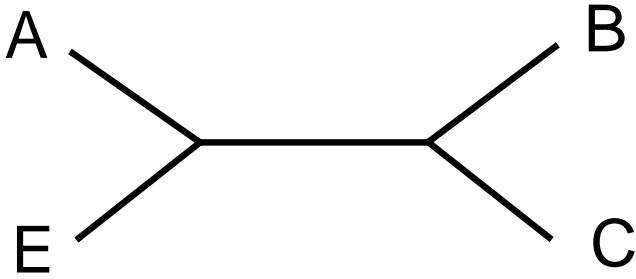
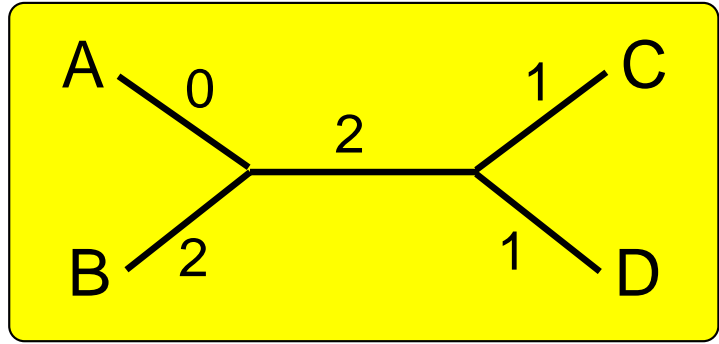


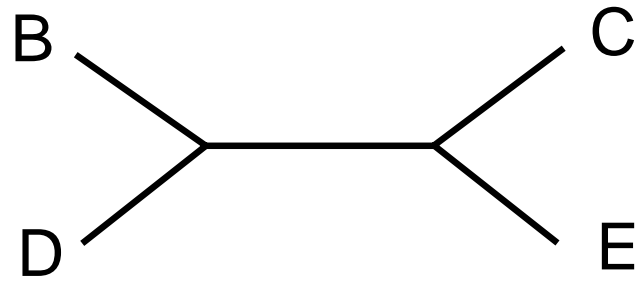
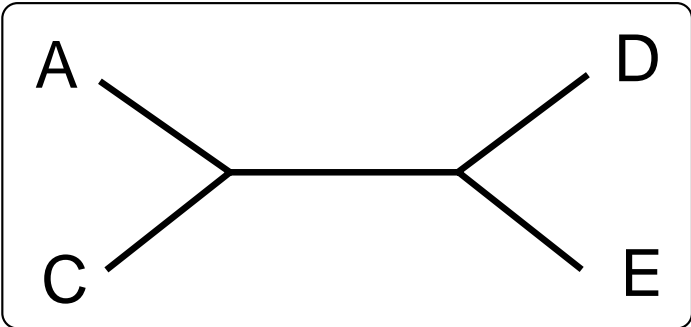
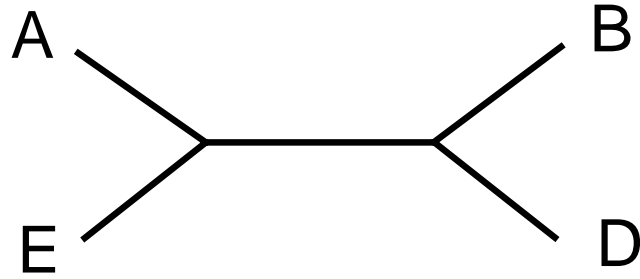
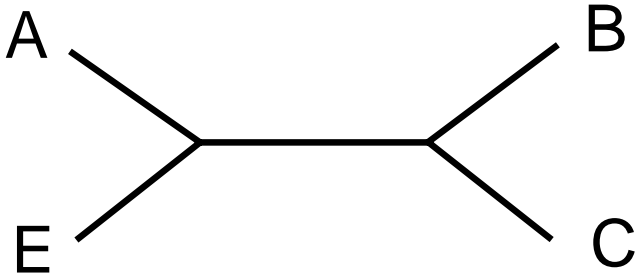
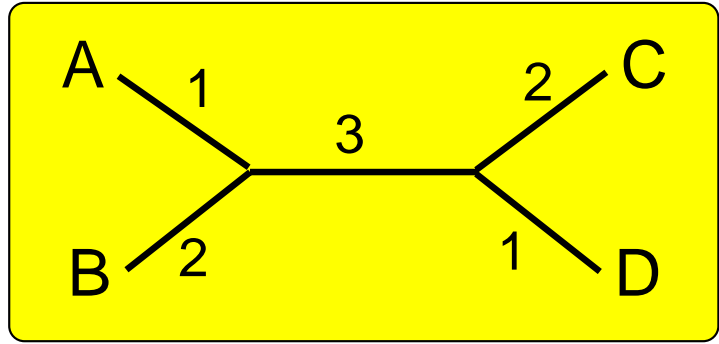


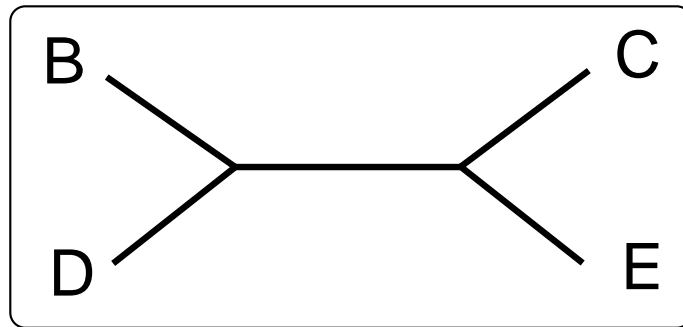
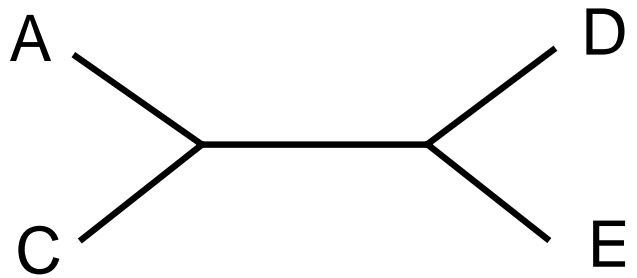
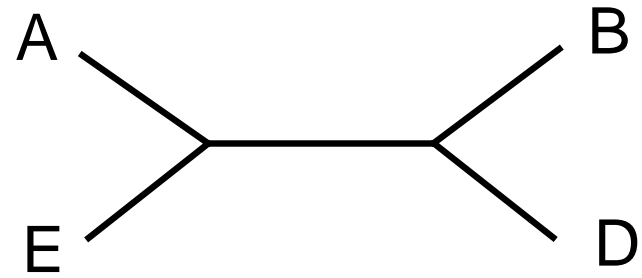
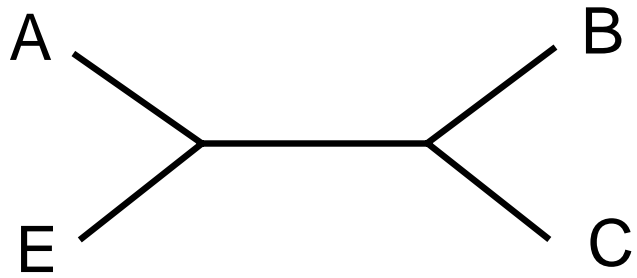
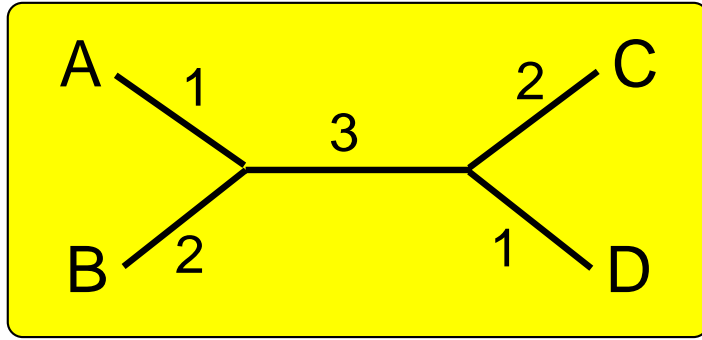


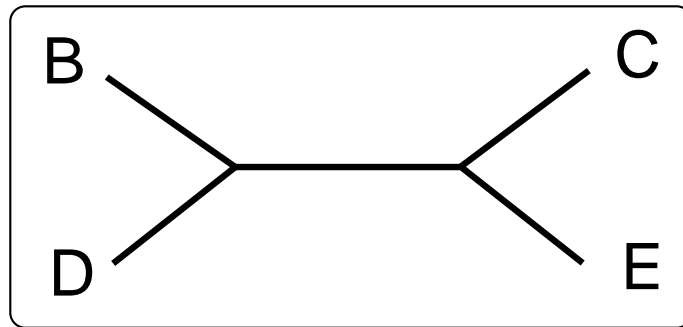
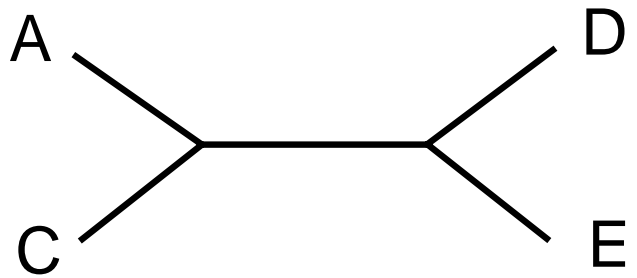
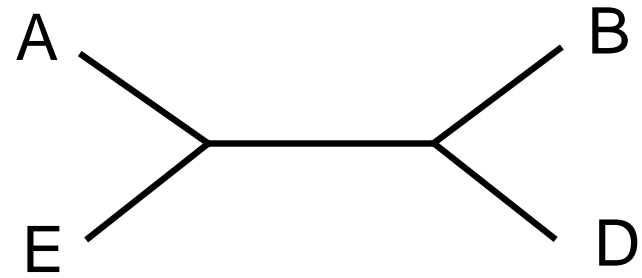
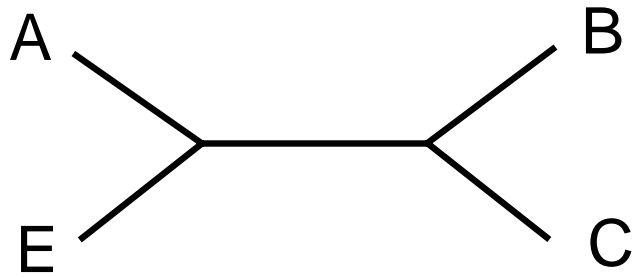
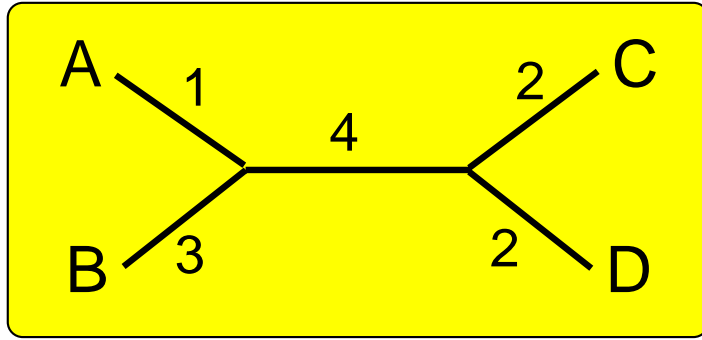


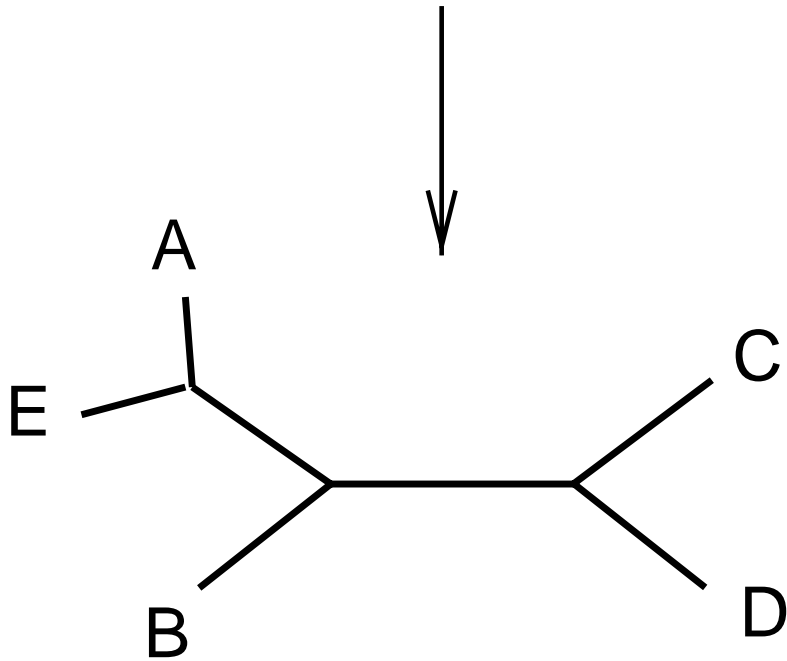
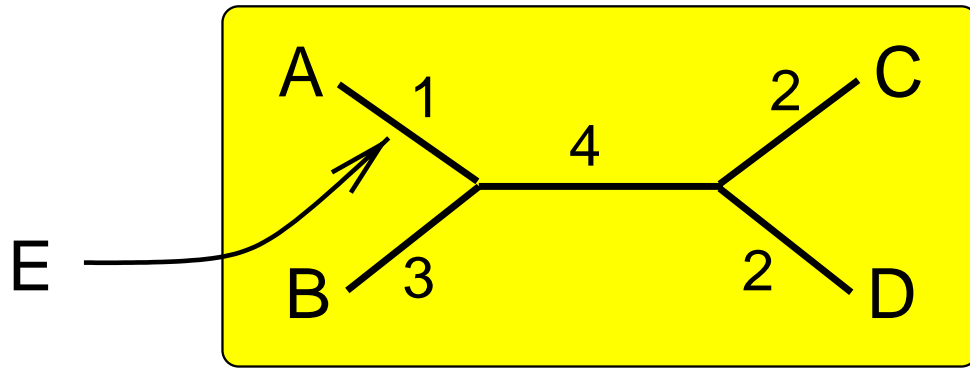


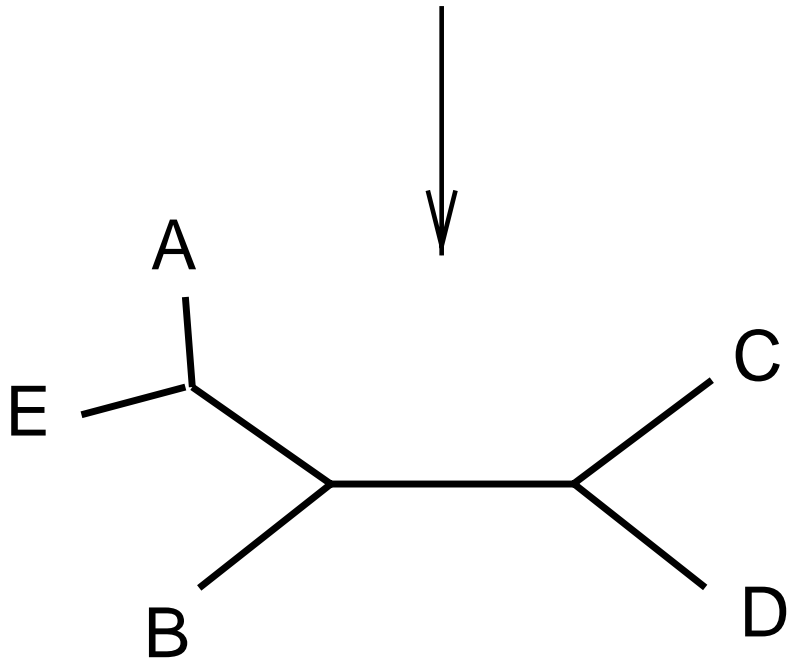
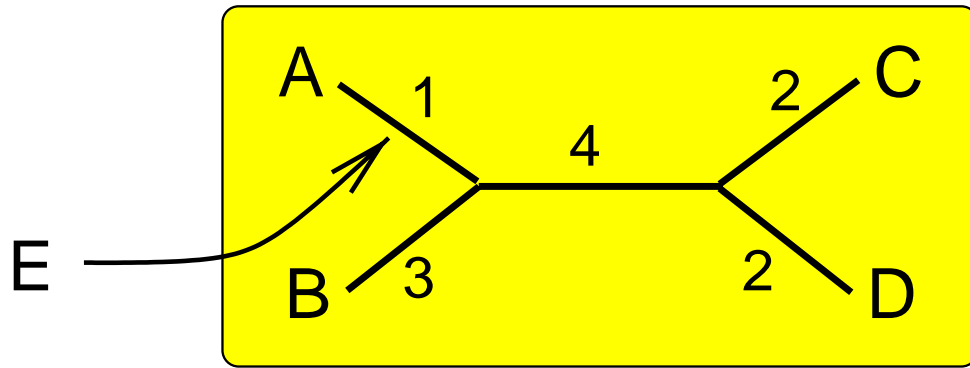










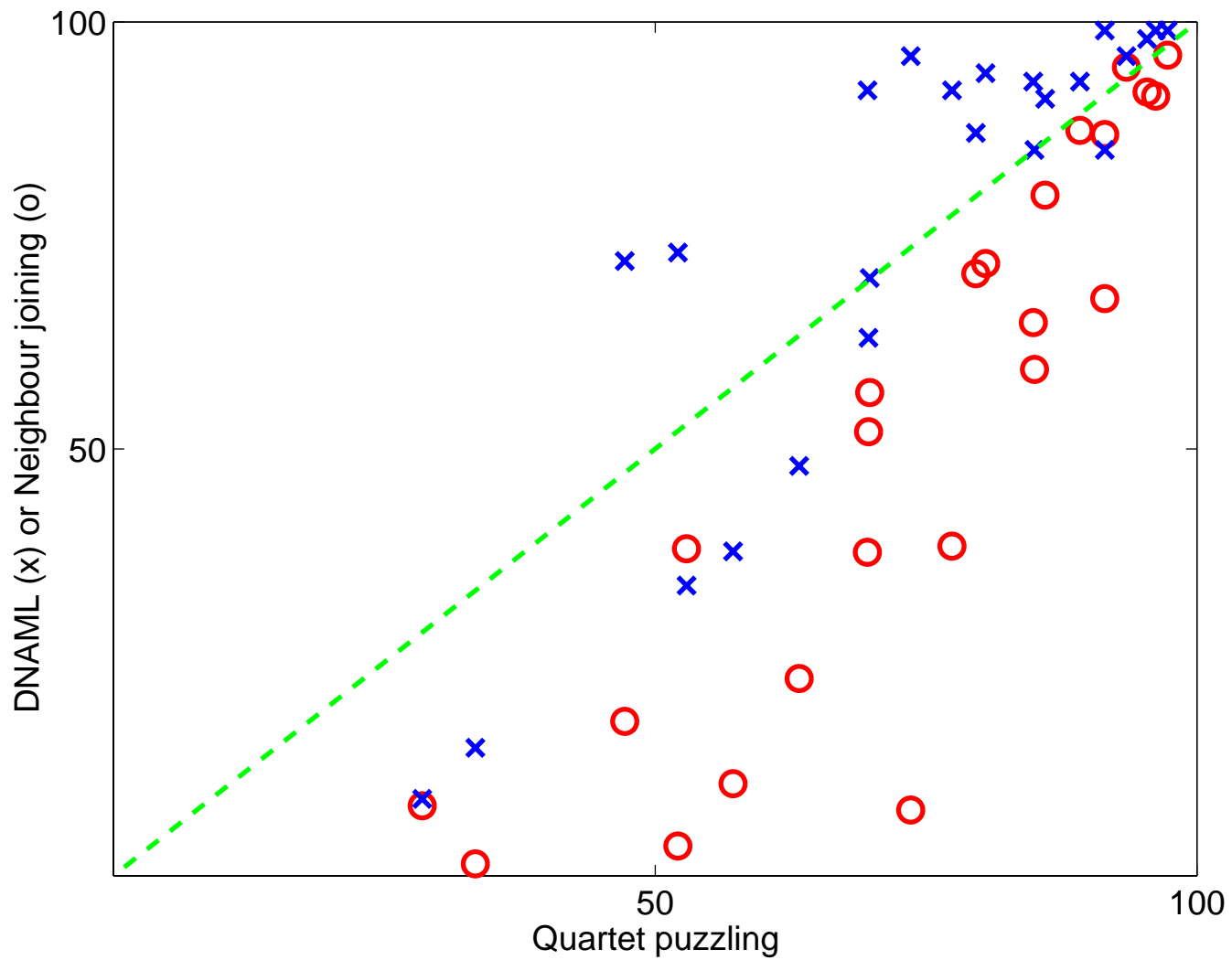


Repeat for different sequence orders

Empirical evaluation (Strimmer, Haeseler, 1996, 1997)

- Comparison between **quartet puzzling**, **DNAML** , and **neighbour joining**.
- 8-taxa **synthetic trees**: with and without molecular clock constraint.
- Sequences simulated with the **Felsenstein 81** and the **Kimura** models of nucleotide substitution.
- Different **sequence lengths**: 500 and 1000.
- Different **branch lengths**.
- For each parameter setting: 1000 simulations (except for DNAML: 100).
- Percentage **accuracy** for reconstructing the true tree.

Comparison of quartet puzzling with DNAML (x) and NJ (o)



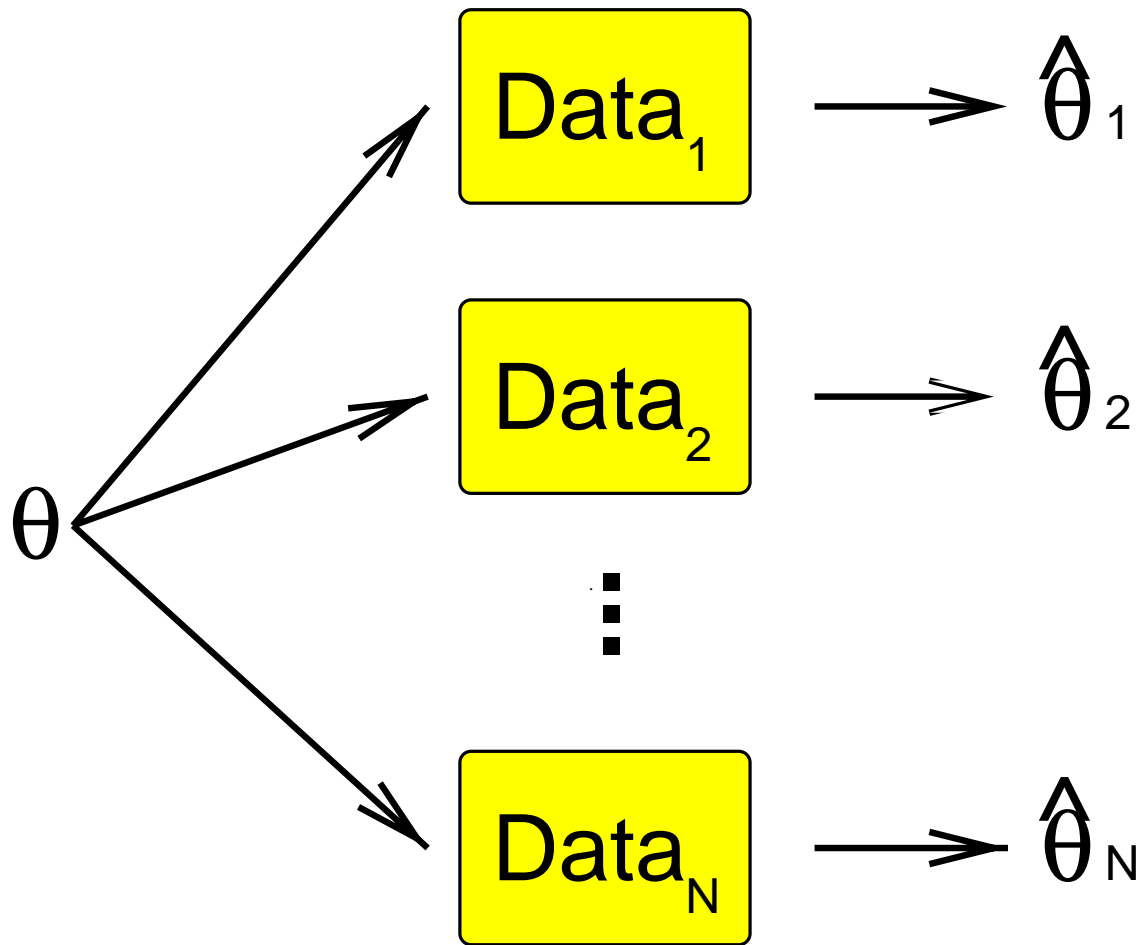
Estimating uncertainty

- Frequentist approach
- Bayesian approach

Statistical inference



Statistical inference





Heads

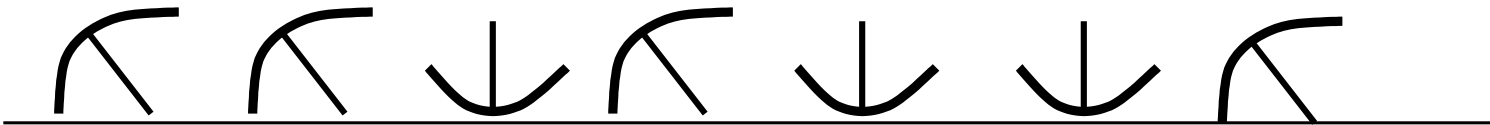
Tails

Probability

θ

$1-\theta$

Data



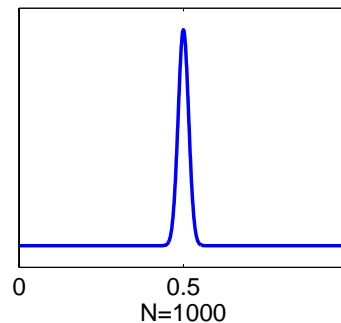
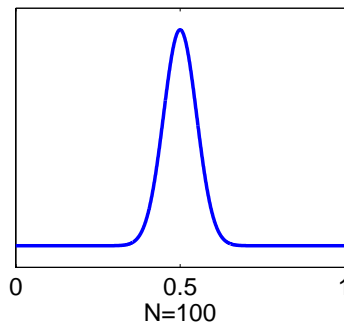
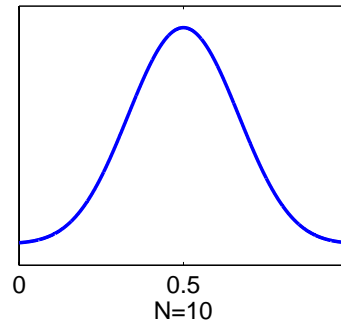
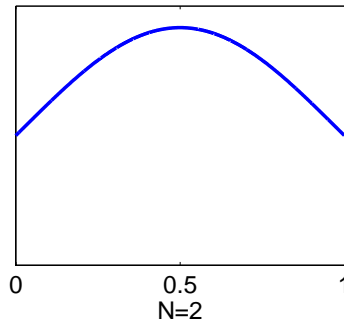
N tosses, k observations of "heads"

$$\hat{\theta} = \frac{k}{N}$$

Thumbnail: $\hat{\theta} = \frac{k}{N}$

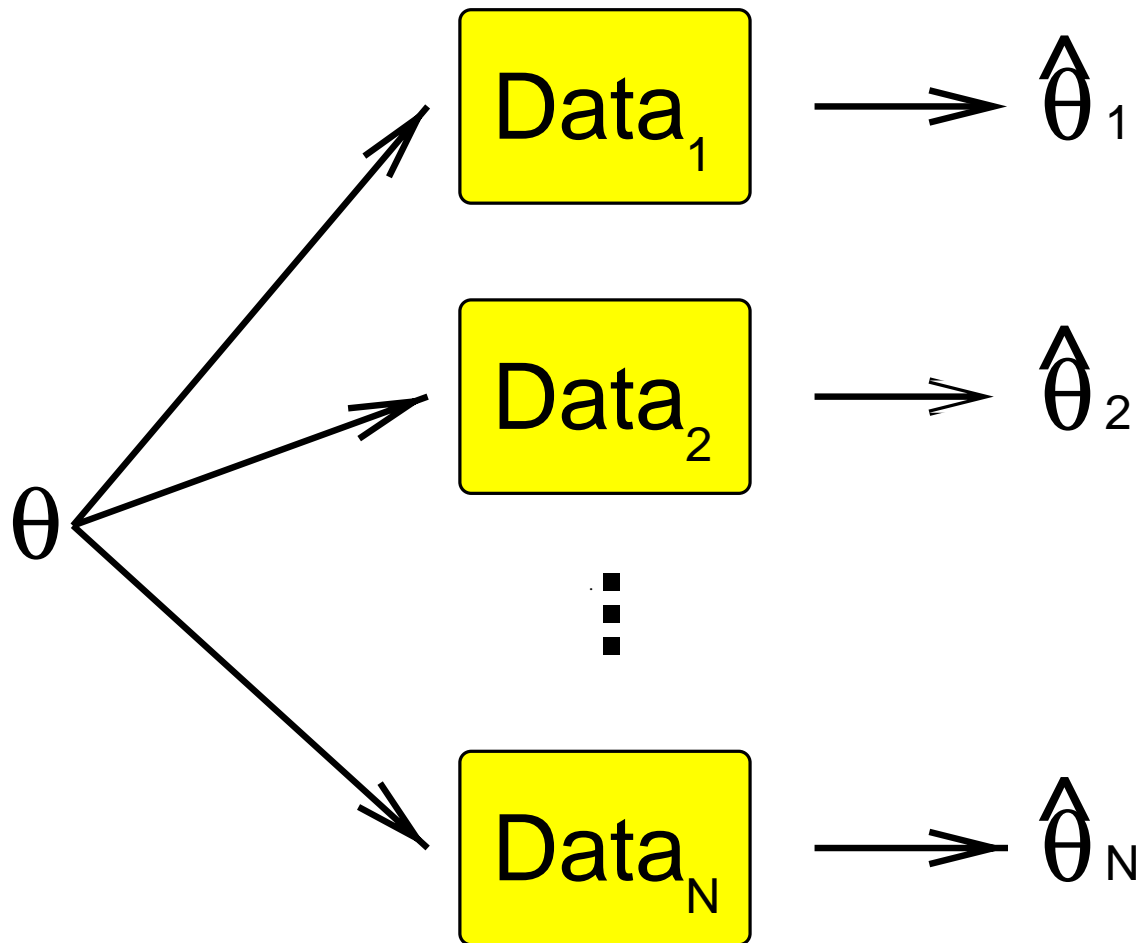
$$P(k) = \theta^k (1 - \theta)^{N-k} \binom{N}{k}$$

$$P(\hat{\theta}) = \theta^{N\hat{\theta}} (1 - \theta)^{N(1-\hat{\theta})} \binom{N}{N\hat{\theta}} (N + 1)$$

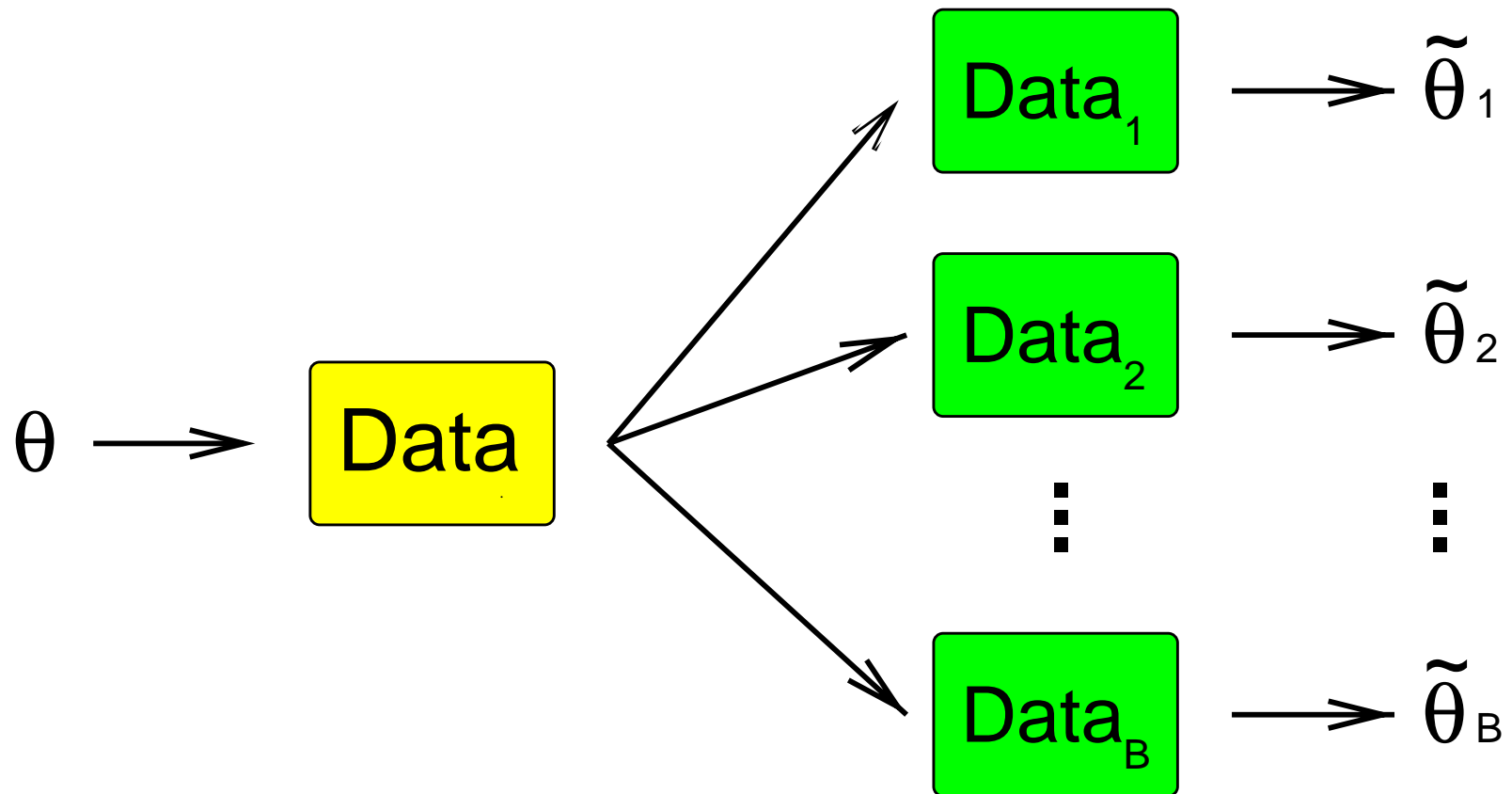


$$E(\hat{\theta}) = \theta, \quad Var(\hat{\theta}) = \frac{\theta(1-\theta)}{N}$$

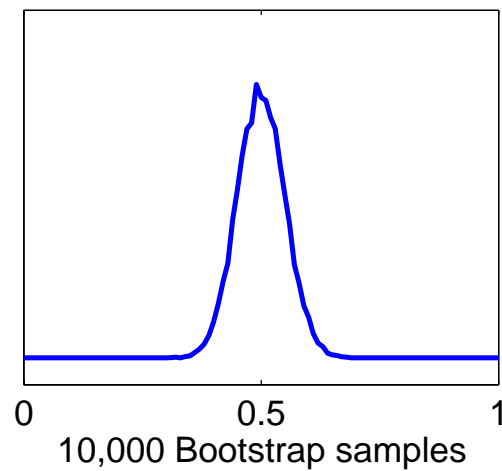
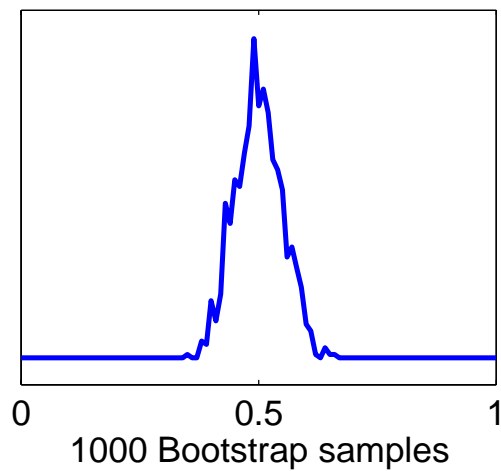
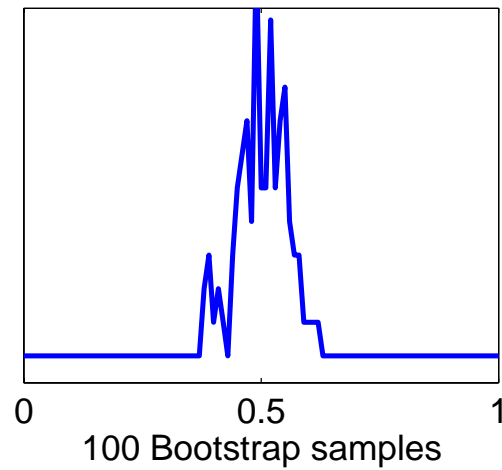
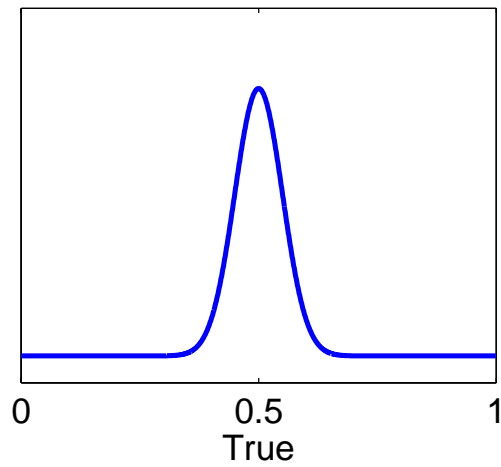
Statistical inference



Bootstrapping

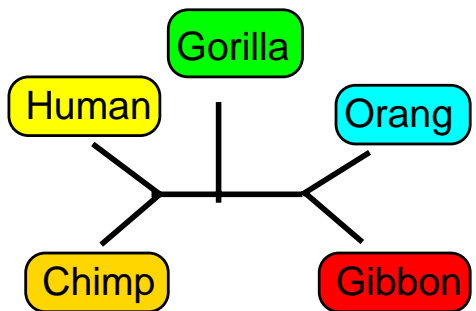


Bootstrapping example: 100 tosses of a thumbnail

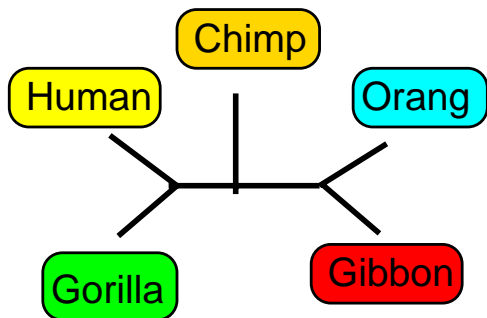


Application of bootstrapping in phylogenetics

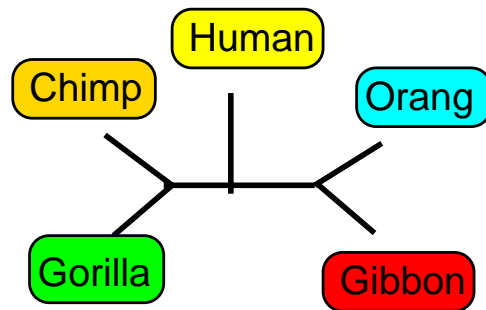
- Sample columns from the alignment with replacement.
- Estimate a phylogenetic tree on the bootstrap sample with ML.
- Compute probabilities for clades.



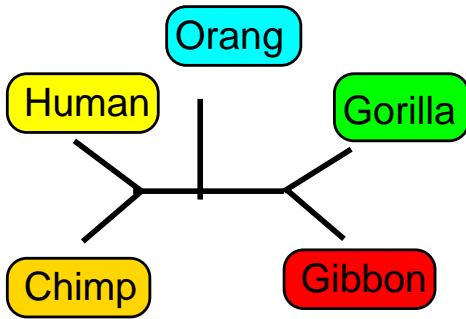
921



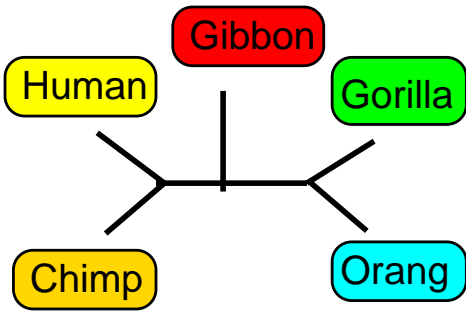
39



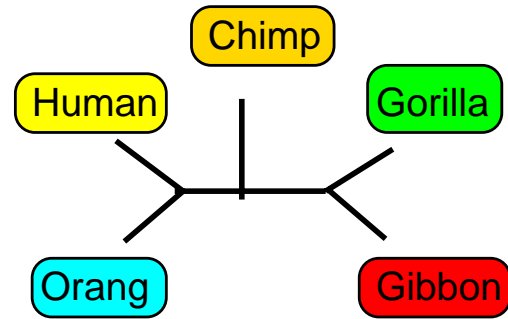
28



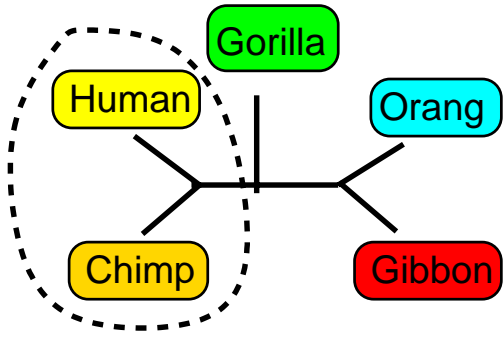
7



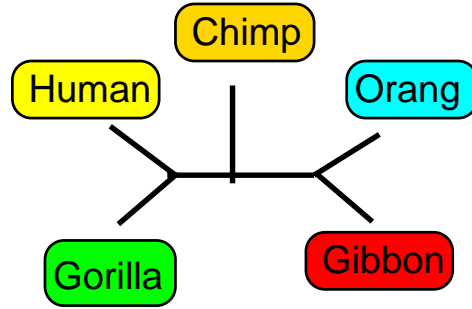
4



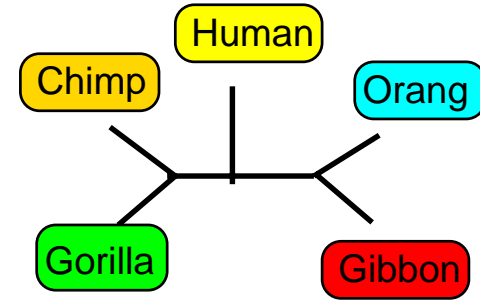
1



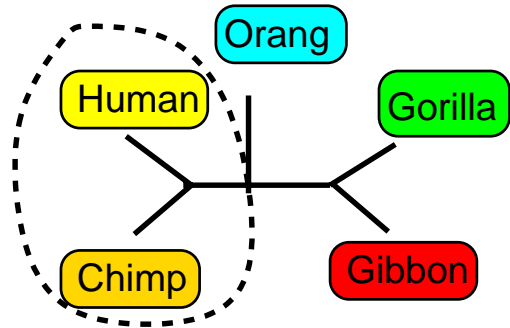
921



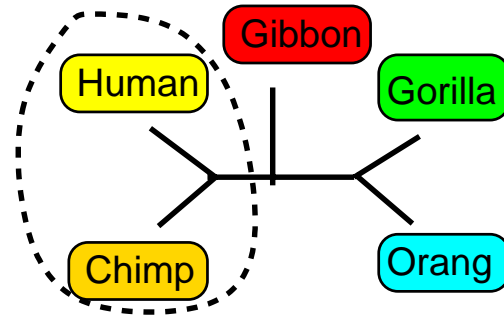
39



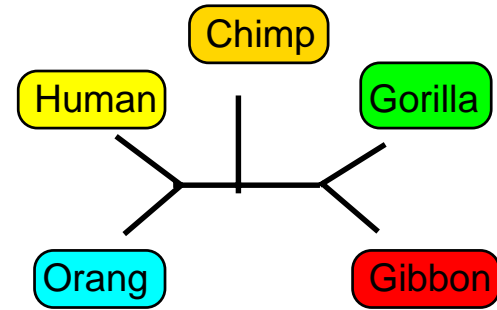
28



7

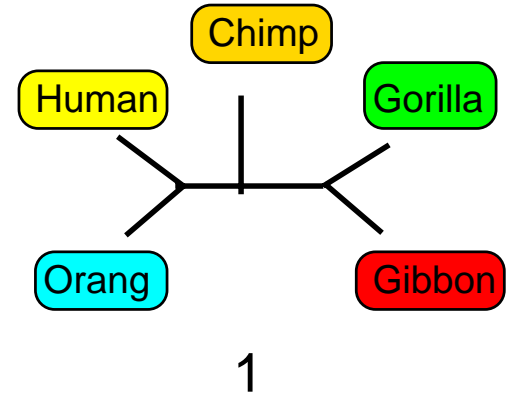
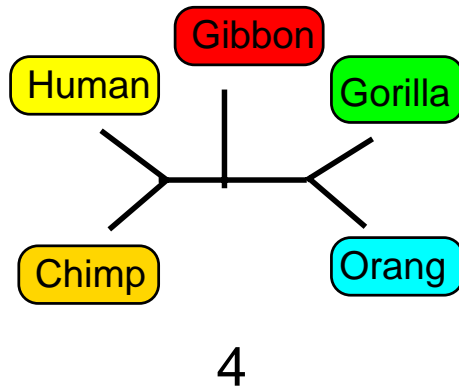
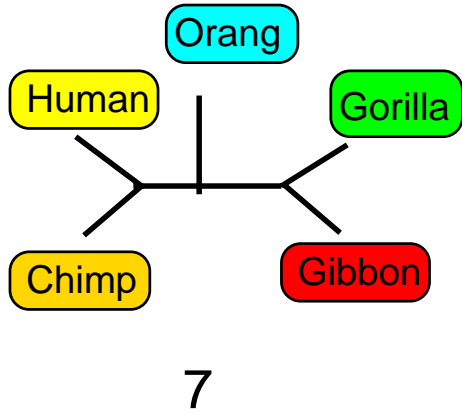
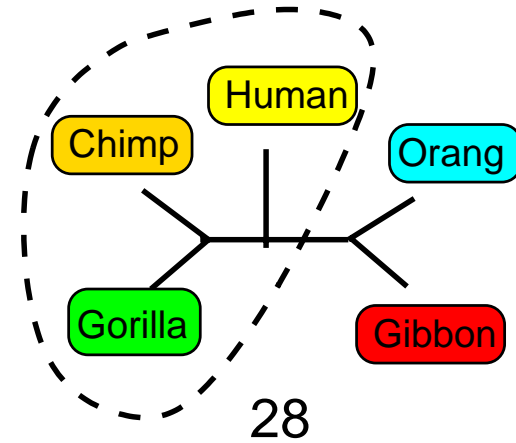
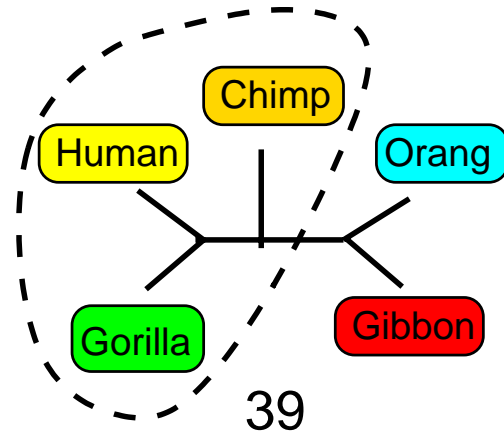
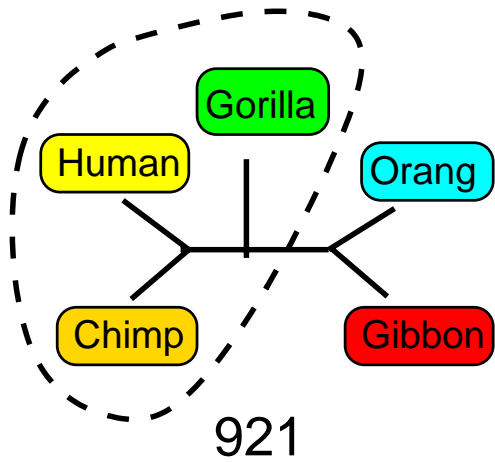


4

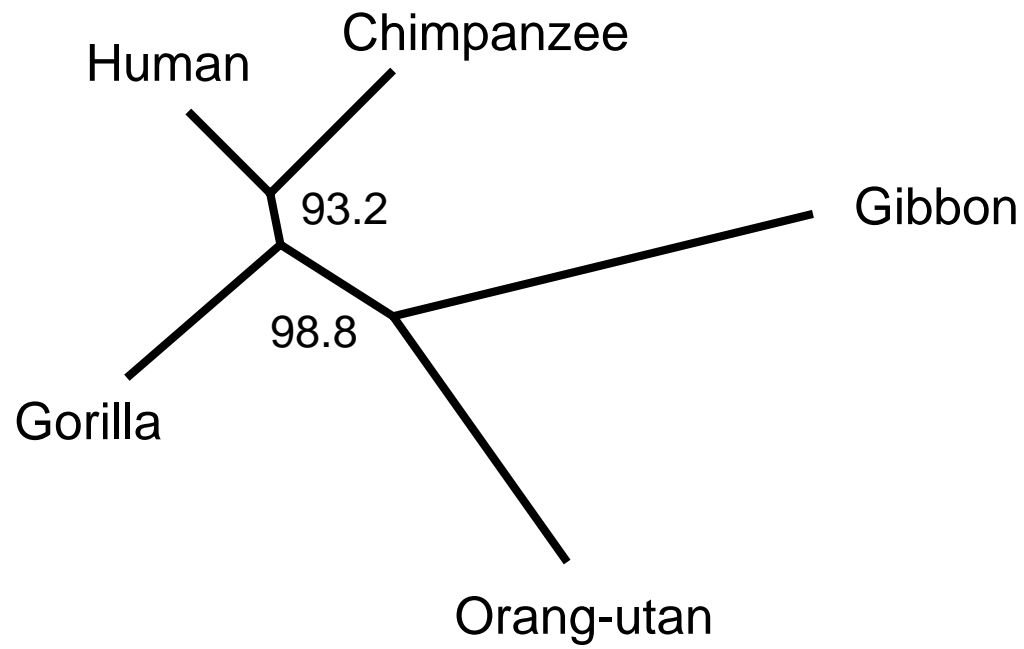


1

$$P = (921 + 7 + 4) / 1000 = 0.932$$



$$P = (921 + 39 + 28) / 1000 = 0.988$$

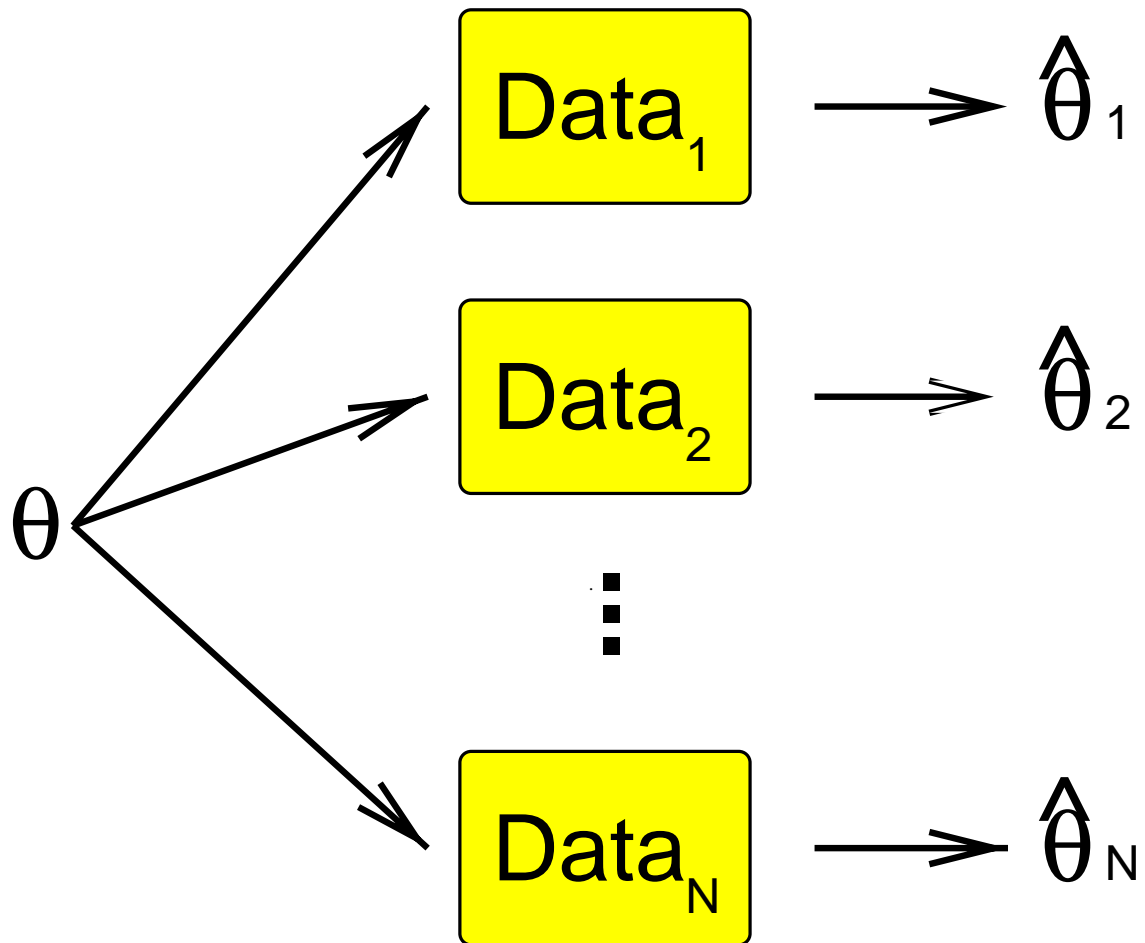


Clade	Probability
(Human Chimp)	0.932
(Human Chimp Gorilla)	0.988

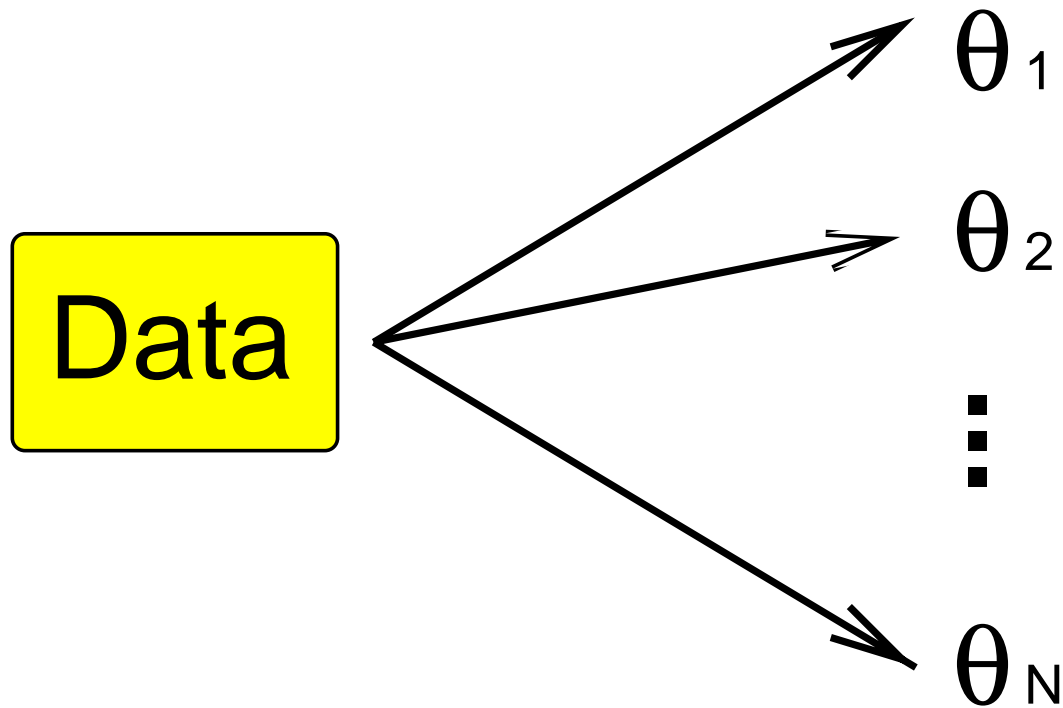
Estimating uncertainty

- Frequentist approach
- Bayesian approach

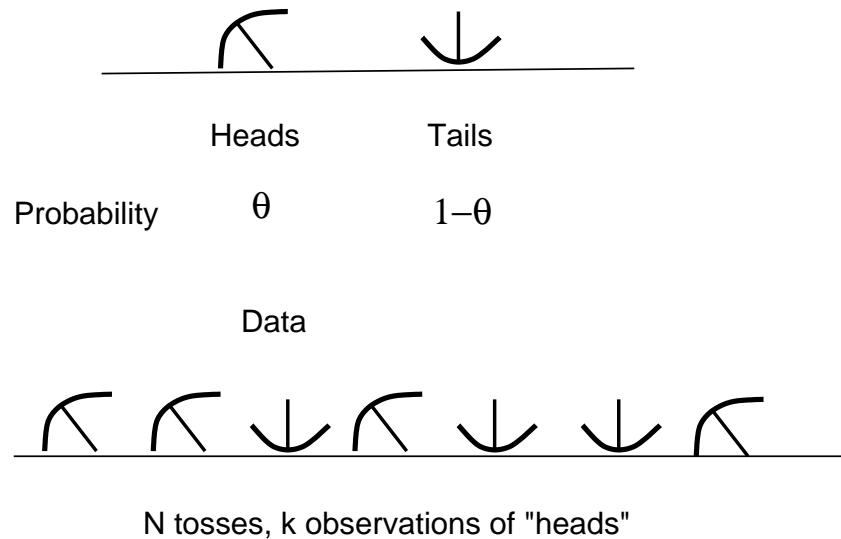
Frequentist statistics



Bayesian statistics



Posterior probability: $P(\theta|\text{Data}) \propto P(\text{Data}|\theta)P(\theta)$



$$P(\theta|D) \propto P(D|\theta)P(\theta)$$

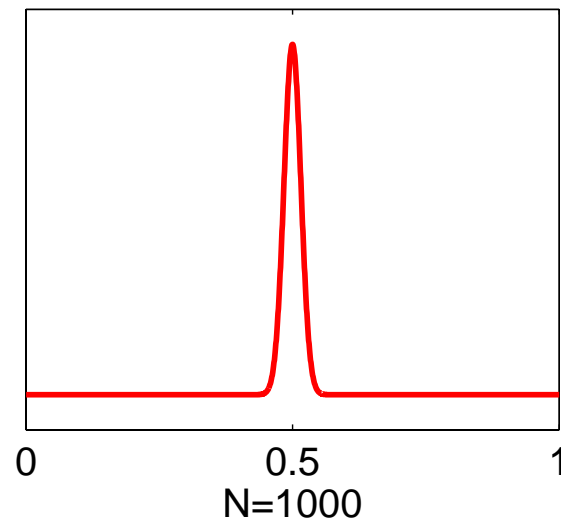
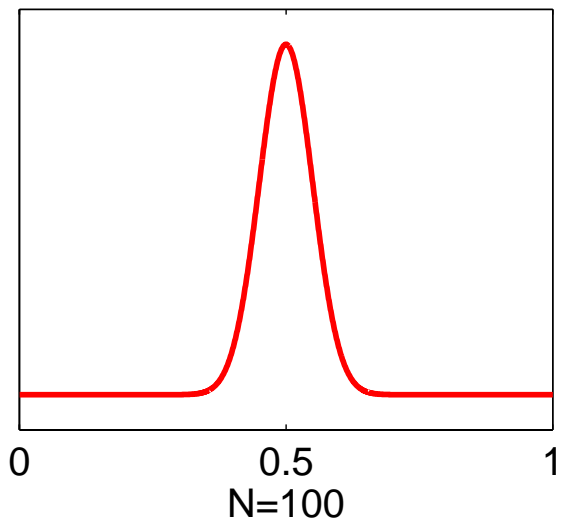
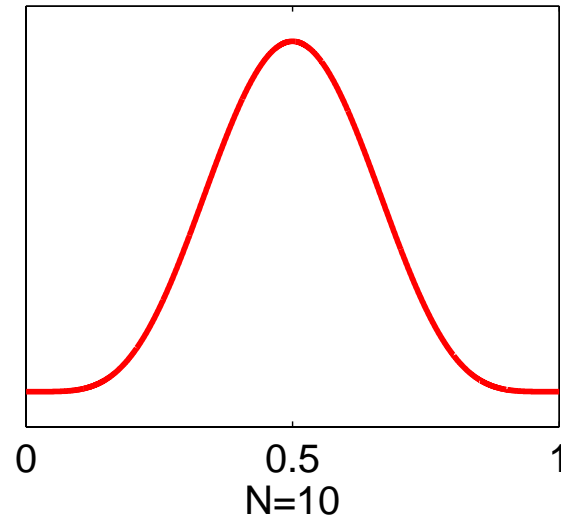
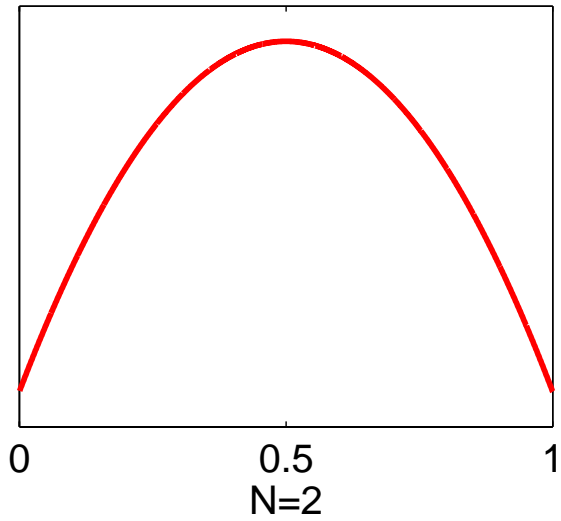
$$P(D|\theta) \propto \theta^k(1 - \theta)^{N-k}$$

$$P(\theta) = B^{-1}(\alpha, \beta)\theta^{\alpha-1}(1 - \theta)^{\beta-1}$$

$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha-1}(1 - \theta)^{\beta-1}d\theta$$

$$P(\theta|D) = B^{-1}(k + \alpha, N - k + \beta)\theta^{k+\alpha-1}(1 - \theta)^{N-k+\beta-1}$$

Example: $P(\theta|D)$ for equal numbers of heads and tails



Markov chain Monte Carlo (MCMC)

- Objective: Sample from the **posterior distribution**

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$$

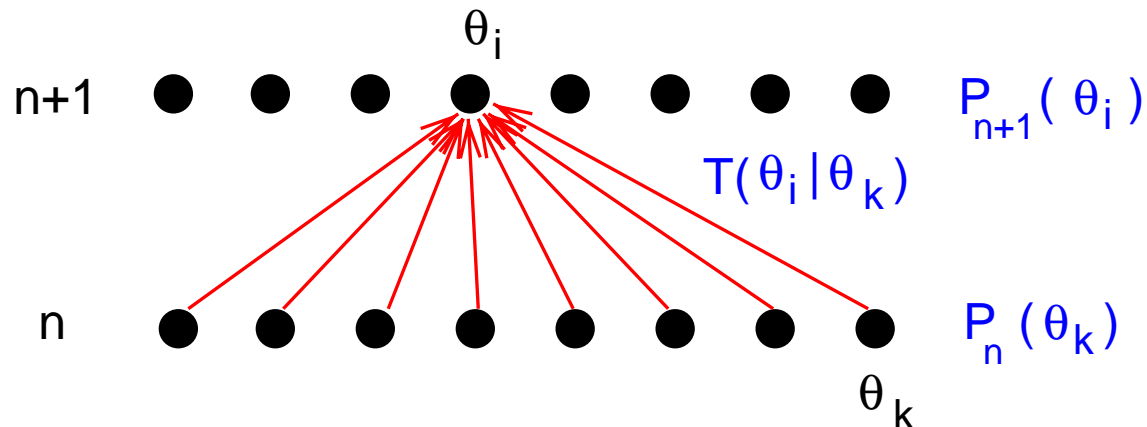
- Direct approach intractable due to $\int P(D|\theta)P(\theta)d\theta$

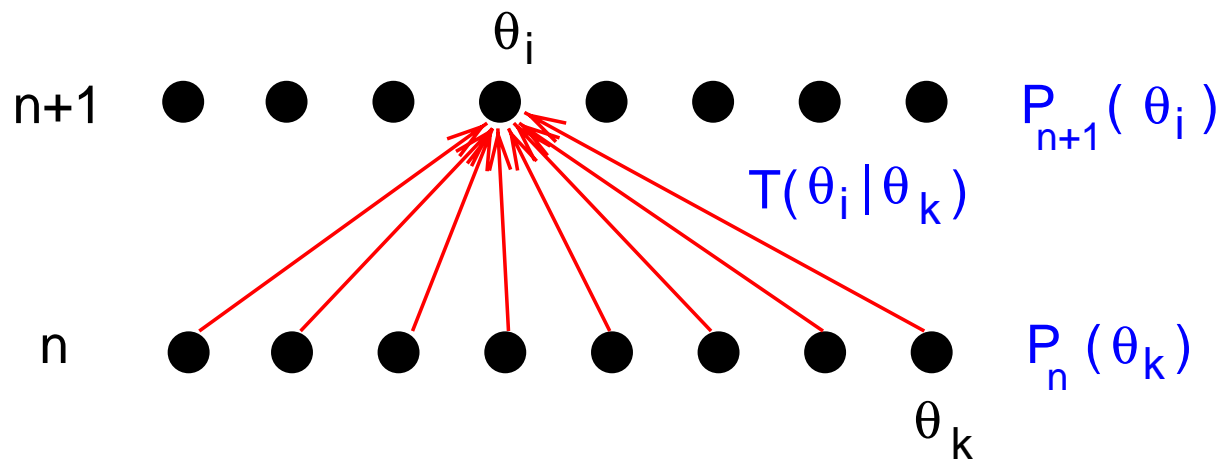
Markov chain Monte Carlo (MCMC)

- Objective: Sample from the posterior distribution

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} = \frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)d\theta}$$

- Direct approach intractable due to $\int P(D|\theta)P(\theta)d\theta$
- Devise a Markov chain $P_{n+1}(\theta_i) = \sum_k T(\theta_i|\theta_k)P_n(\theta_k)$ that converges in distribution to $P(\theta|D)$: $P_n(\theta) \rightarrow P(\theta|D)$





- Theorem: An **ergodic** Markov chain converges to its **stationary distribution** irrespective of its **initialization**.
- Stationary distribution: $P(\theta_i) = \sum_k T(\theta_i|\theta_k)P(\theta_k)$
- Design the **Markov transition matrix** $T(\theta_i|\theta_k)$ such that $P(\theta|D)$ is the stationary distribution.
- Detailed balance: $\frac{T(\theta_k|\theta_i)}{T(\theta_i|\theta_k)} = \frac{P(\theta_k|D)}{P(\theta_i|D)} = \frac{P(D|\theta_k)P(\theta_k)}{P(D|\theta_i)P(\theta_i)}$

Proof

Show that $\sum_k T(\theta_i|\theta_k)P(\theta_k|D) = P(\theta_i|D)$

Detailed balance: $\frac{T(\theta_k|\theta_i)}{T(\theta_i|\theta_k)} = \frac{P(\theta_k|D)}{P(\theta_i|D)} \implies$

$$T(\theta_k|\theta_i)P(\theta_i|D) = T(\theta_i|\theta_k)P(\theta_k|D) \implies$$

$$\begin{aligned} \sum_k T(\theta_i|\theta_k)P(\theta_k|D) &= \sum_k T(\theta_k|\theta_i)P(\theta_i|D) = \\ &P(\theta_i|D) \sum_k T(\theta_k|\theta_i) = P(\theta_i|D) \end{aligned}$$

Metropolis-Hastings algorithm

$$\frac{T(\theta_k|\theta_i)}{T(\theta_i|\theta_k)} = \frac{P(D|\theta_k)P(\theta_k)}{P(D|\theta_i)P(\theta_i)}$$

Transition Probability = Proposal Probability \times Acceptance Probability

$$T(\theta_k|\theta_i) = q(\theta_k|\theta_i)a(\theta_k|\theta_i)$$

Metropolis-Hastings algorithm

$$\frac{T(\theta_k|\theta_i)}{T(\theta_i|\theta_k)} = \frac{P(D|\theta_k)P(\theta_k)}{P(D|\theta_i)P(\theta_i)}$$

Transition Probability = Proposal Probability \times Acceptance Probability

$$T(\theta_k|\theta_i) = q(\theta_k|\theta_i)a(\theta_k|\theta_i)$$

Acceptance Probabilities:

$$\frac{a(\theta_k|\theta_i)}{a(\theta_i|\theta_k)} = \frac{P(D|\theta_k)P(\theta_k)q(\theta_i|\theta_k)}{P(D|\theta_i)P(\theta_i)q(\theta_k|\theta_i)}$$

$$a(\theta_k|\theta_i) = \min \left\{ \frac{P(D|\theta_k)P(\theta_k)q(\theta_i|\theta_k)}{P(D|\theta_i)P(\theta_i)q(\theta_k|\theta_i)}, 1 \right\}$$

Metropolis-Hastings algorithm

- **Start** from initial θ_0
- **Iterate** $n = 1 \dots N$
 1. Obtain new $\theta^{(n)}$ from proposal distribution $q(\theta^{(n)}|\theta^{(n-1)})$
 2. Accept with probability $a(\theta^{(n)}|\theta^{(n-1)})$, otherwise leave unchanged: $\theta^{(n)} = \theta^{(n-1)}$
- **Discard** $\theta_1, \dots, \theta_{N/2}$ (burn-in period)
- **Sample** from $\theta_{N/2+1}, \dots, \theta_N$

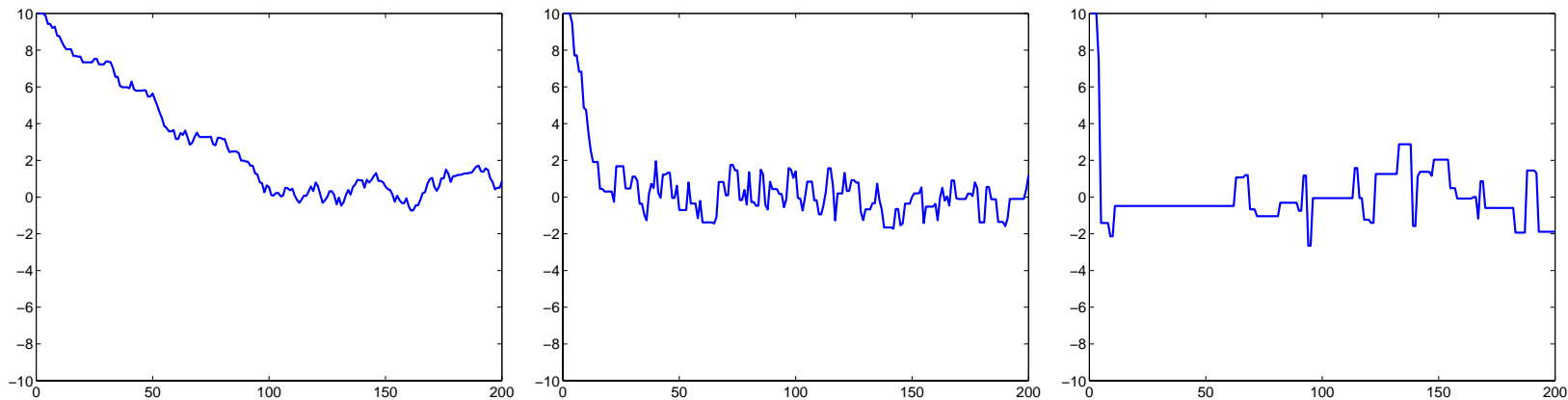
$$\bullet \int f(\theta)P(\theta|D)d\theta = \frac{2}{N} \sum_{n=N/2+1}^N f(\theta_n)$$

Simple MCMC example: Mean of a normal distribution

Proposal distribution: $X_{move} = X_{old} + 2\lambda(U - 0.5)$

U : uniform RV from $[0, 1]$, λ : step size

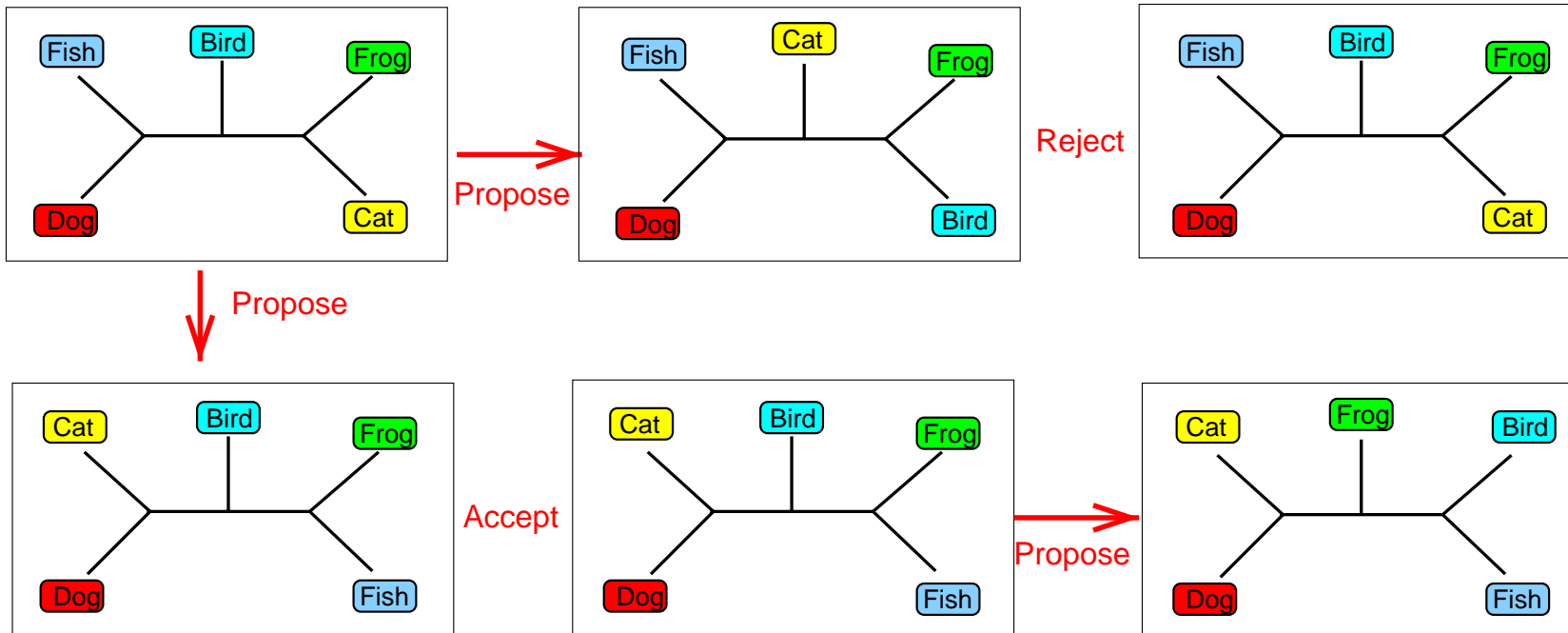
MCMC sample size: $N = 200$



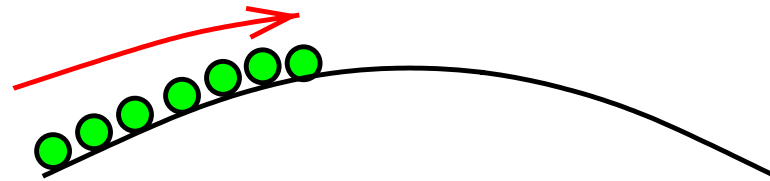
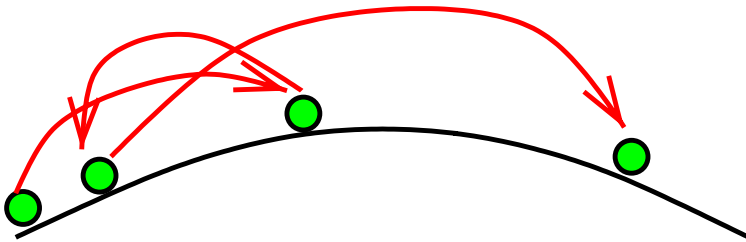
λ	AR	Absolute error
0.5	79%	0.51
2	58%	0.19
10	17%	0.22

For $N = 2000$, all errors are below 0.04.

Illustration of MCMC applied to phylogenetics

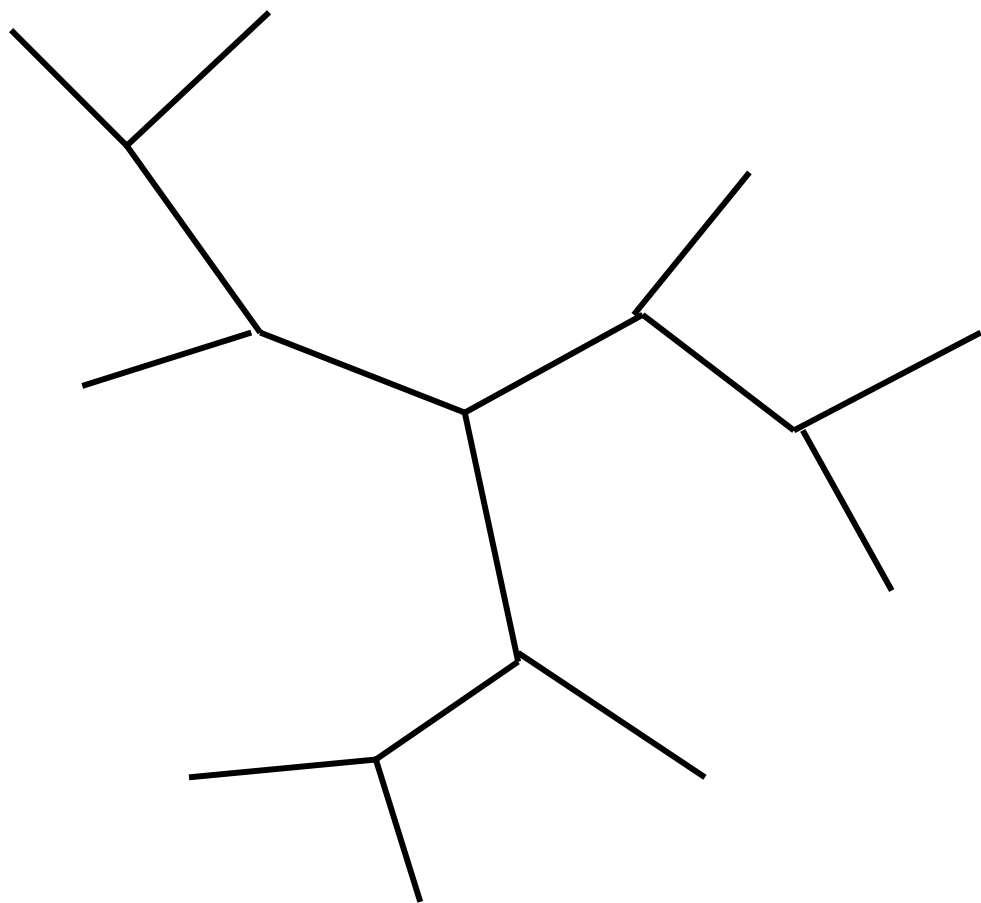


Comparison between MCMC and bootstrapping

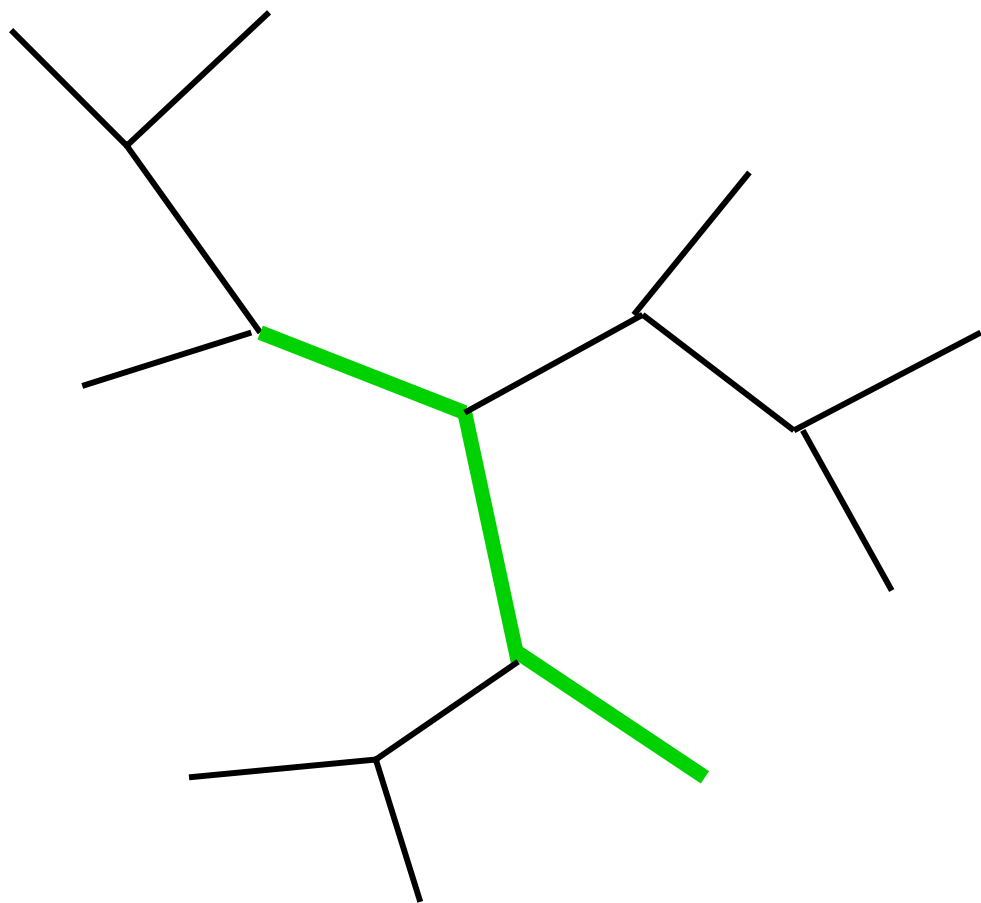


Implementation

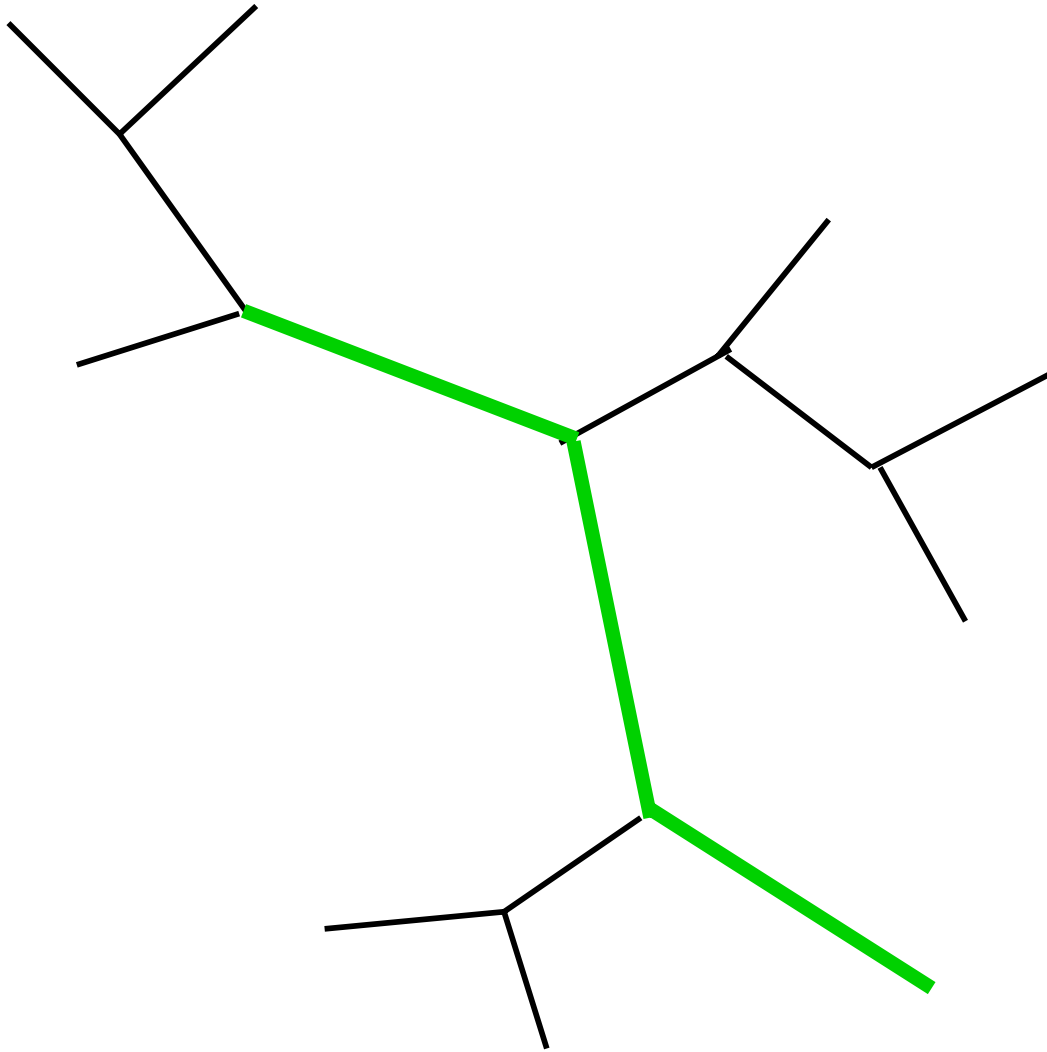
- B. Larget, D. L. Simon (1999)
Molecular Biology and Evolution 6 (16), 750-759
- BAMBE:
Bayesian Analysis in Molecular Biology and Evolution
<http://www.mathcs.duq.edu/larget/bambe.html>



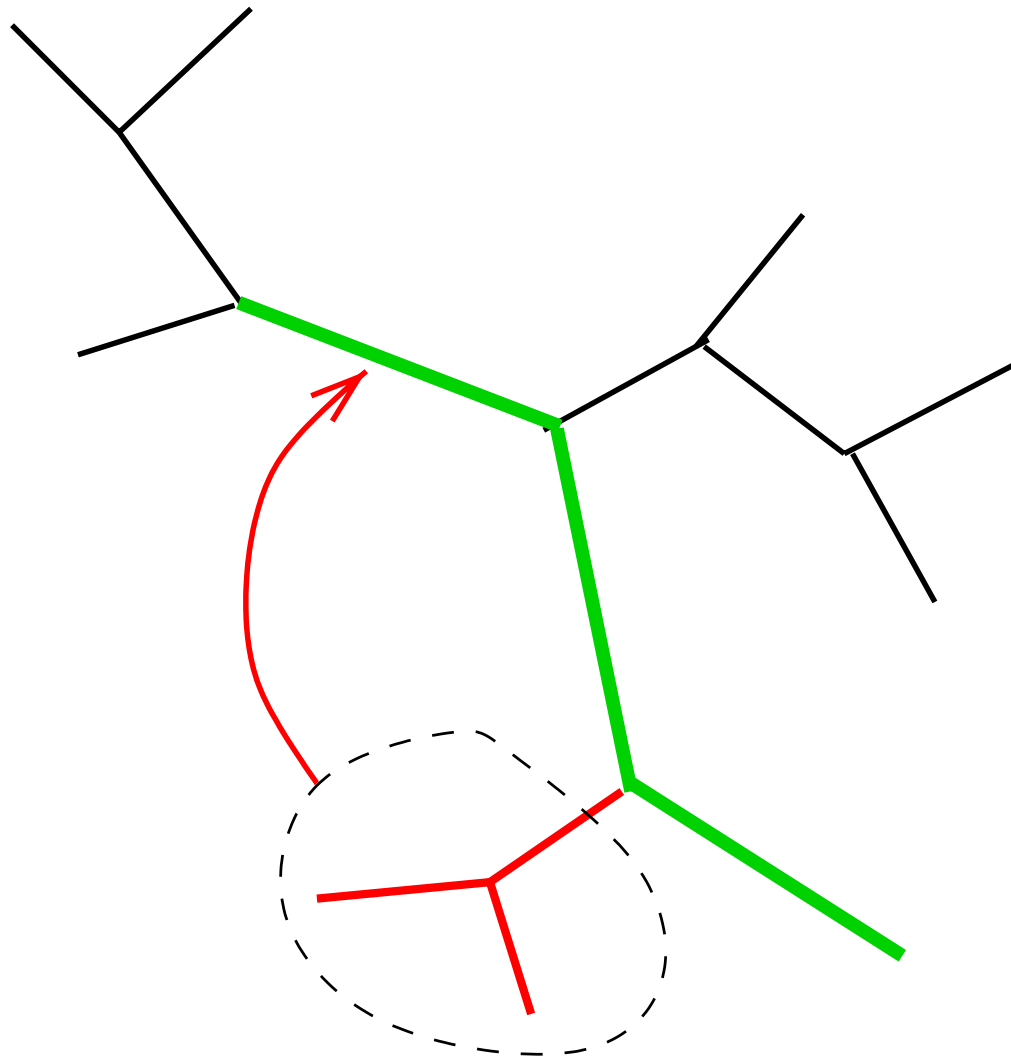
..

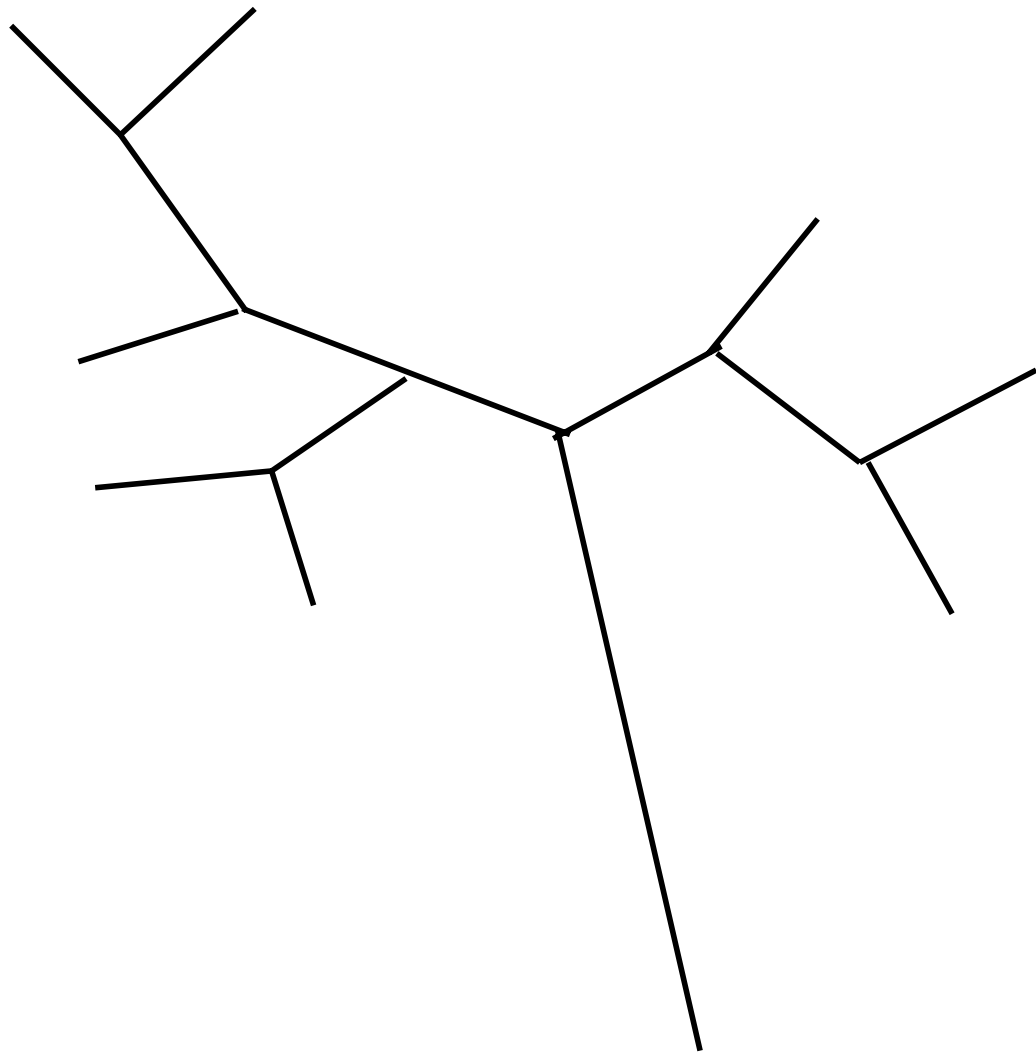


..



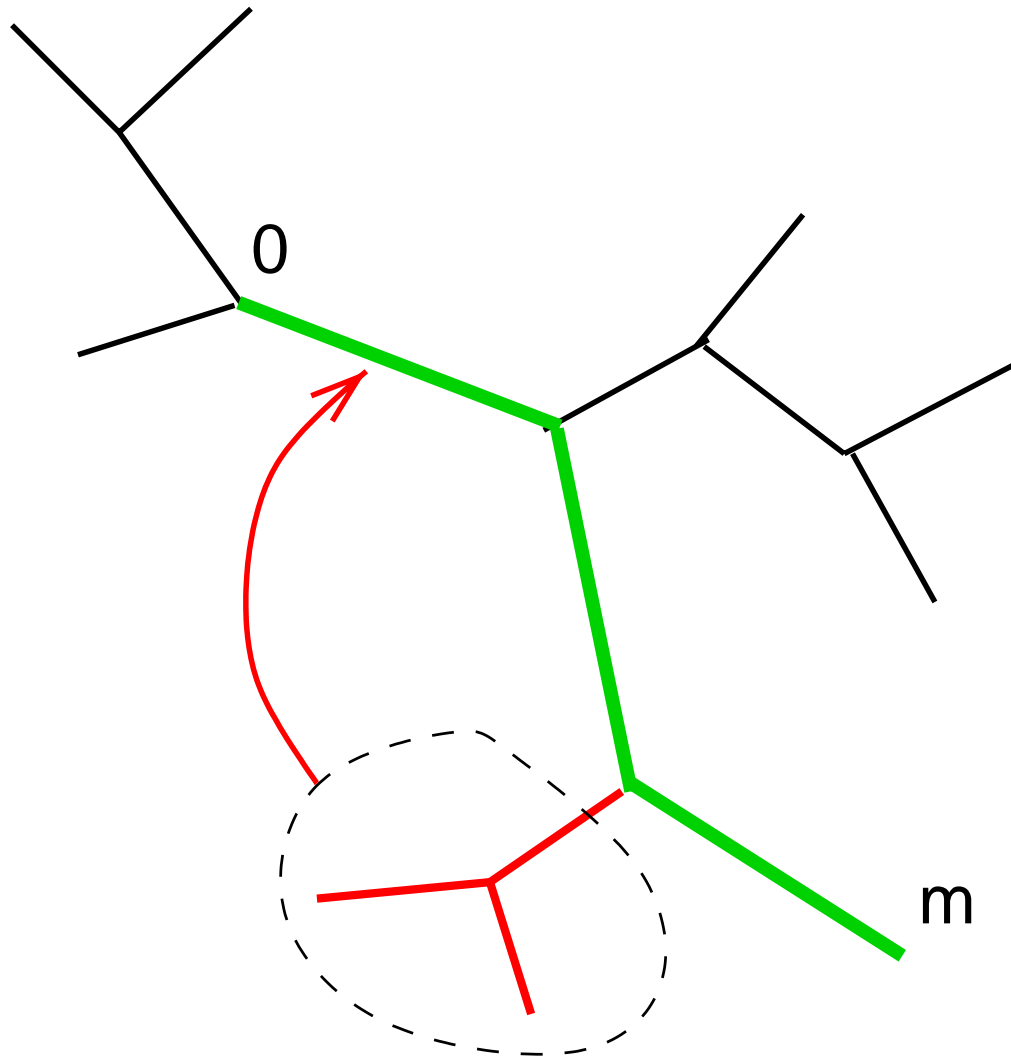
..





- Old length: m
- New length: $m' = m \exp(\lambda(U - 0.5))$
 U : Uniform random variable in $[0, 1]$
 $\lambda > 0$: Tuning parameter
- Select one of the branches with equal probabilities
- Regraft this branch; new position chosen uniformly from $[0, m']$.

What is the **Hastings ratio**: $\frac{P(\text{backward move})}{P(\text{forward move})}$?



Forward move

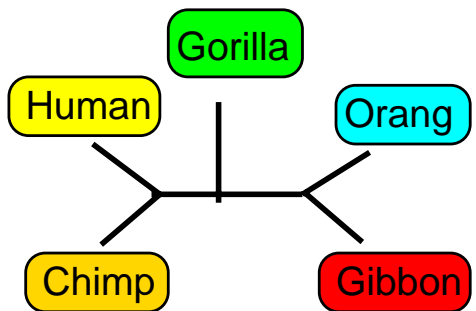
$$y' \in [0, m'] \quad \text{with probability} \quad P(y') = \frac{1}{m'}$$
$$x' = \frac{m'}{m}x \quad \text{with probability} \quad P(x') = P(x) \frac{dx}{dx'} = P(x) \frac{m}{m'}$$

Backward move

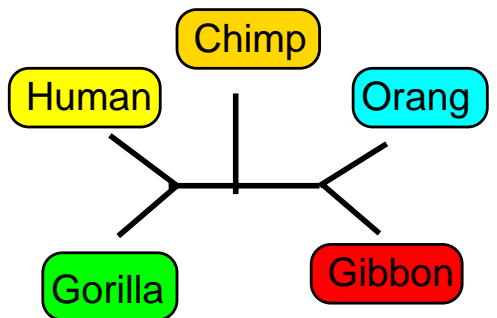
$$y \in [0, m] \quad \text{with probability} \quad P(y) = \frac{1}{m}$$
$$x \quad \text{with probability} \quad P(x)$$

Hastings ratio

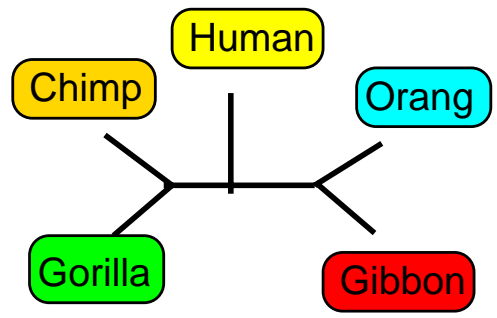
$$\frac{P(\text{backward move})}{P(\text{forward move})} = \frac{P(y)P(x)}{P(y')P(x')} = \left(\frac{m'}{m}\right)^2$$



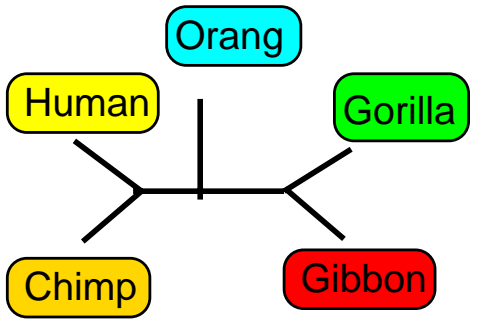
0.919



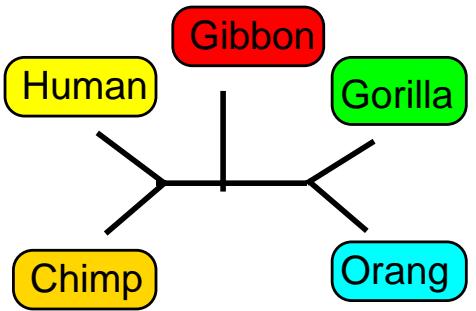
0.038



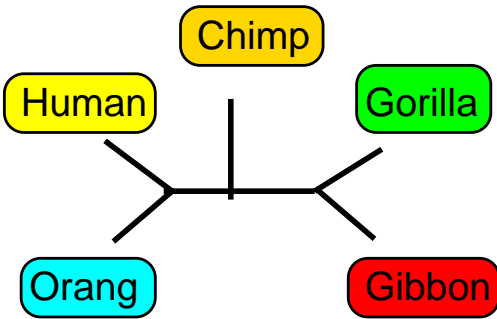
0.027



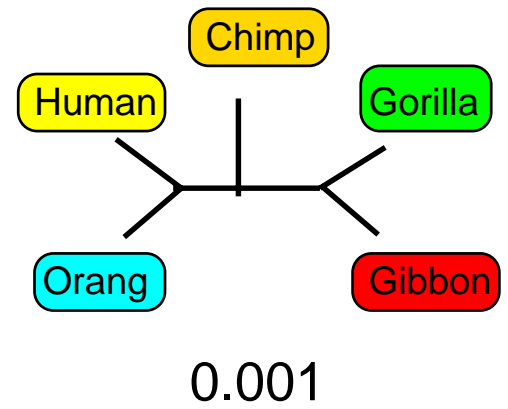
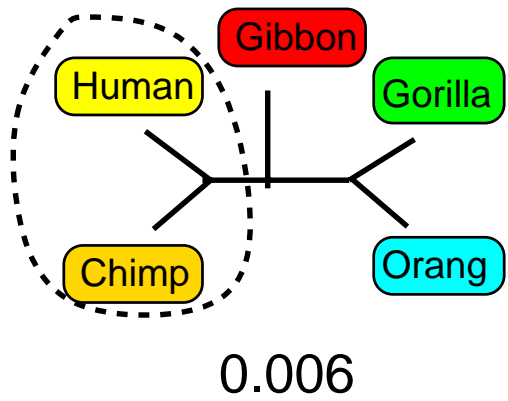
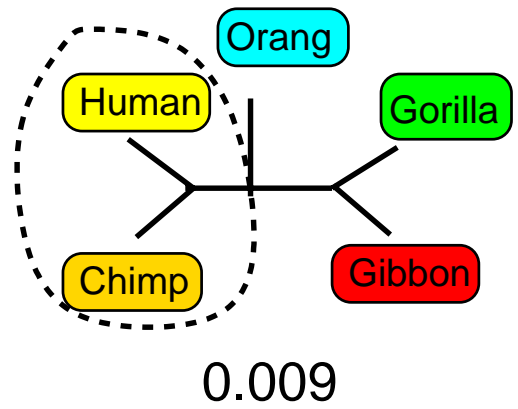
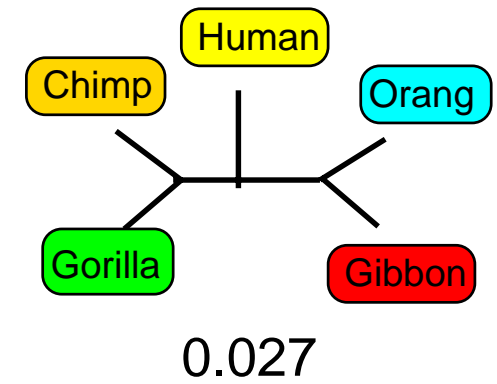
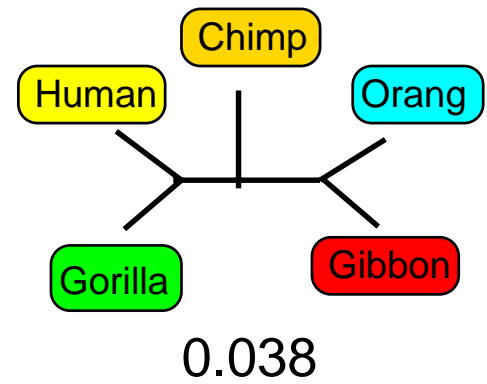
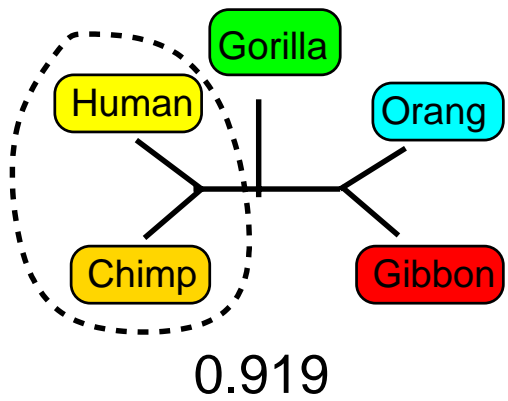
0.009



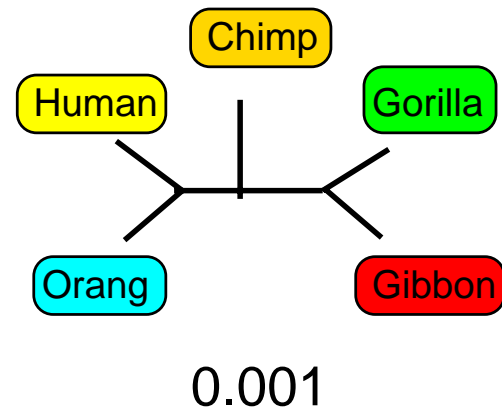
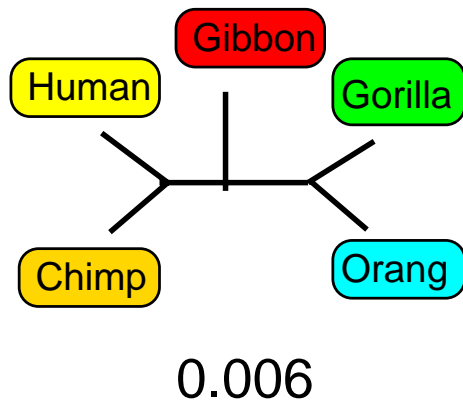
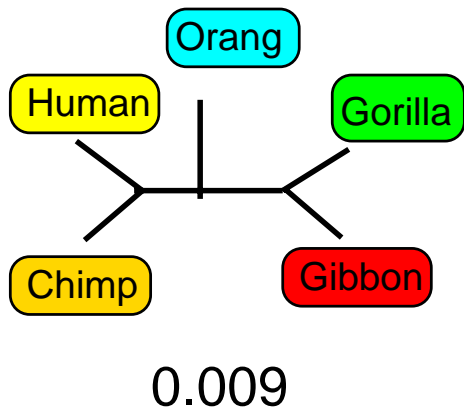
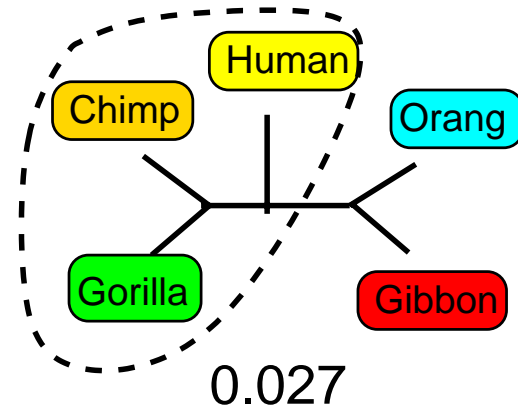
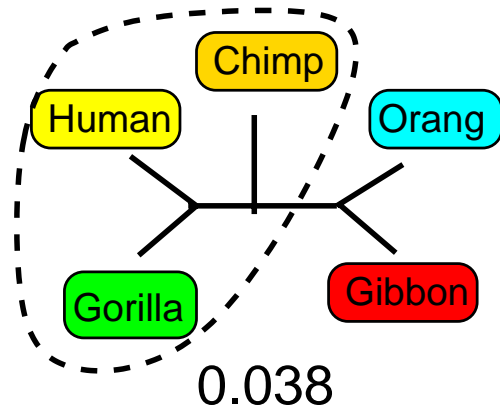
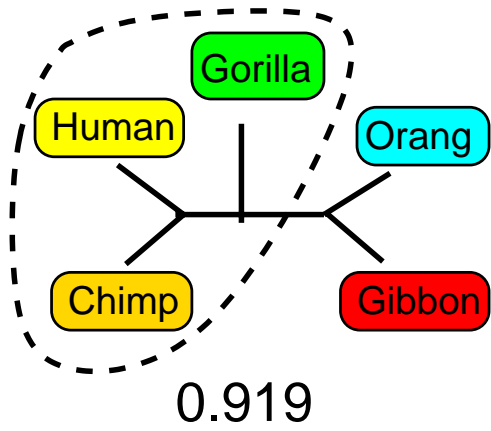
0.006



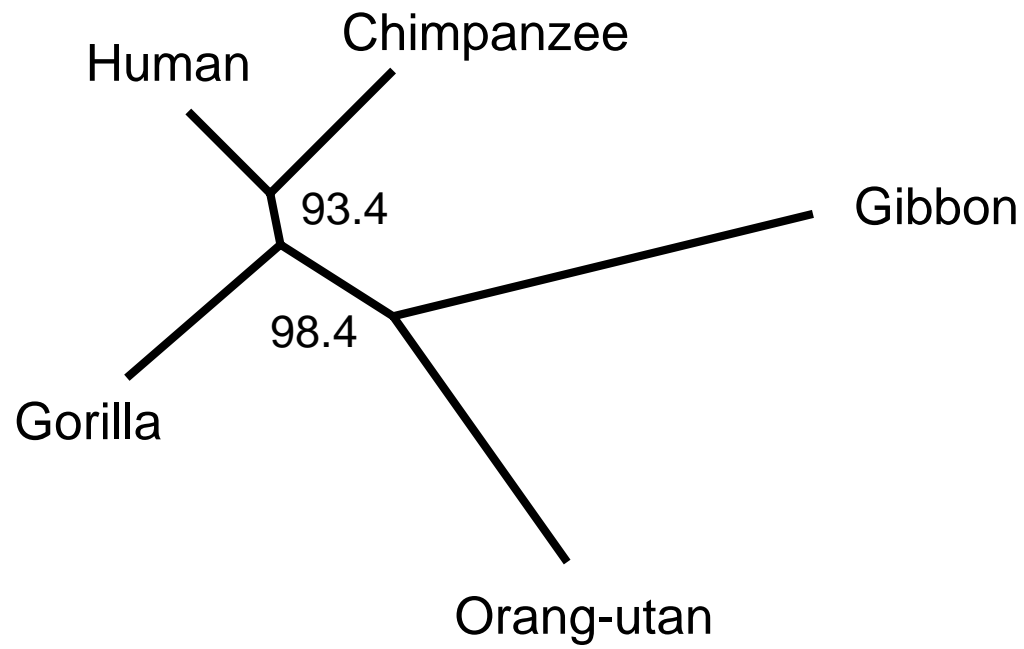
0.001



$$P = 0.919 + 0.009 + 0.006 = 0.934$$

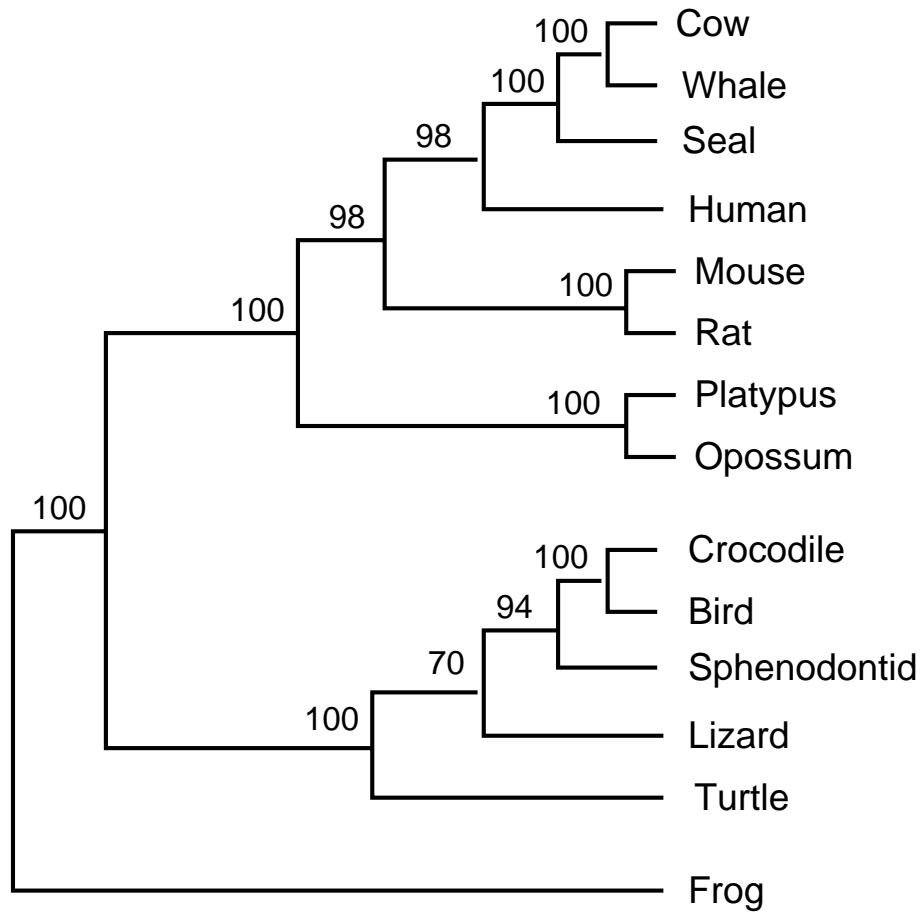


$$P = 0.919 + 0.038 + 0.027 = 0.984$$



Clade	Probability
(Human Chimp)	0.934
(Human Chimp Gorilla)	0.984

- Concatenated sequences of mitochondrial 12S rRNA, 16S rRNA, and tRNA^{Val} genes. Total length: 2439 sites.



Summary

Methods of phylogenetic inference

- Clustering
- Parsimony
- Likelihood

Estimating uncertainty

- Frequentist approach
- Bayesian approach