
Introduction to Phylogenetics

Part 1

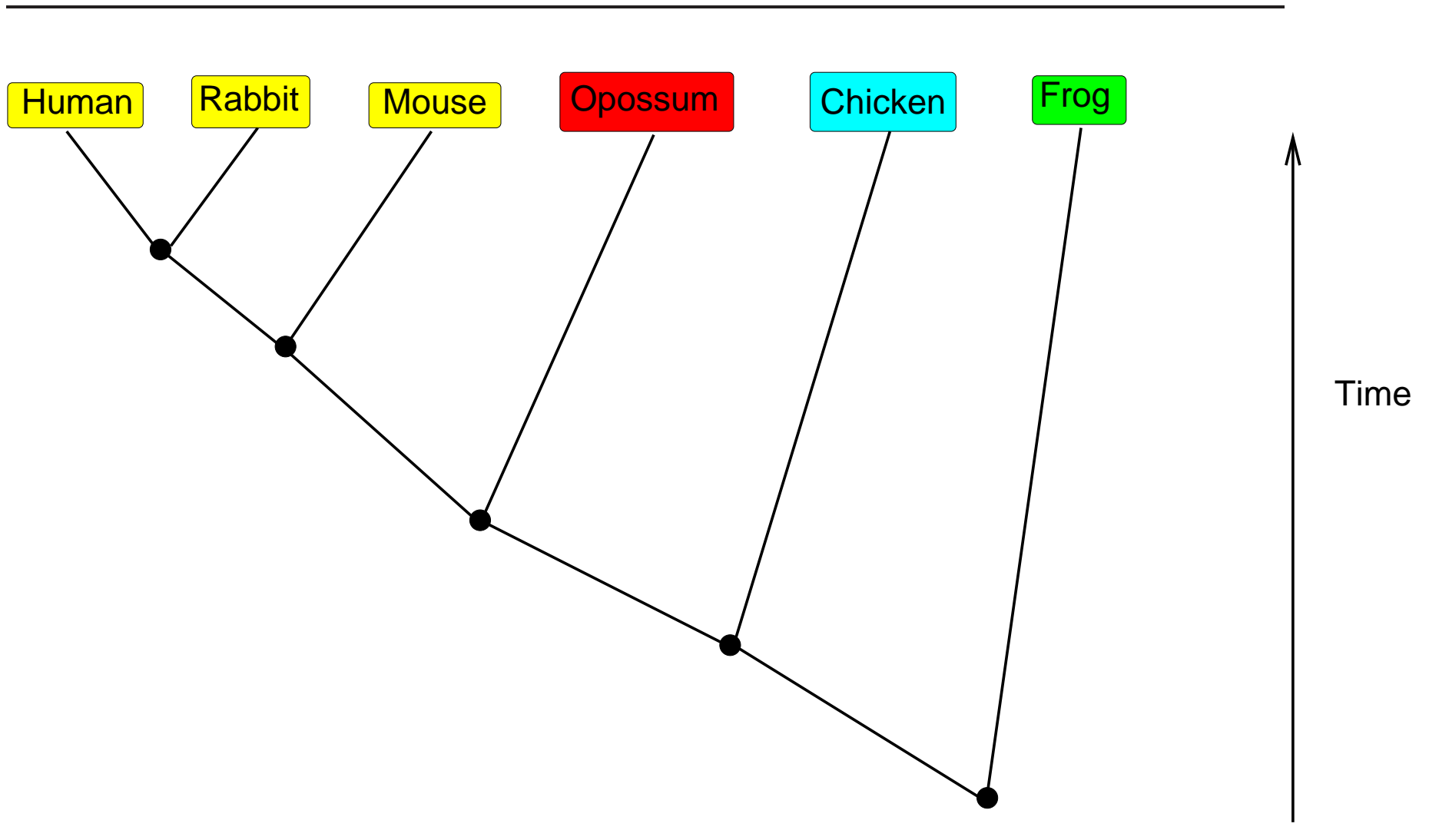
Dirk Husmeier

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

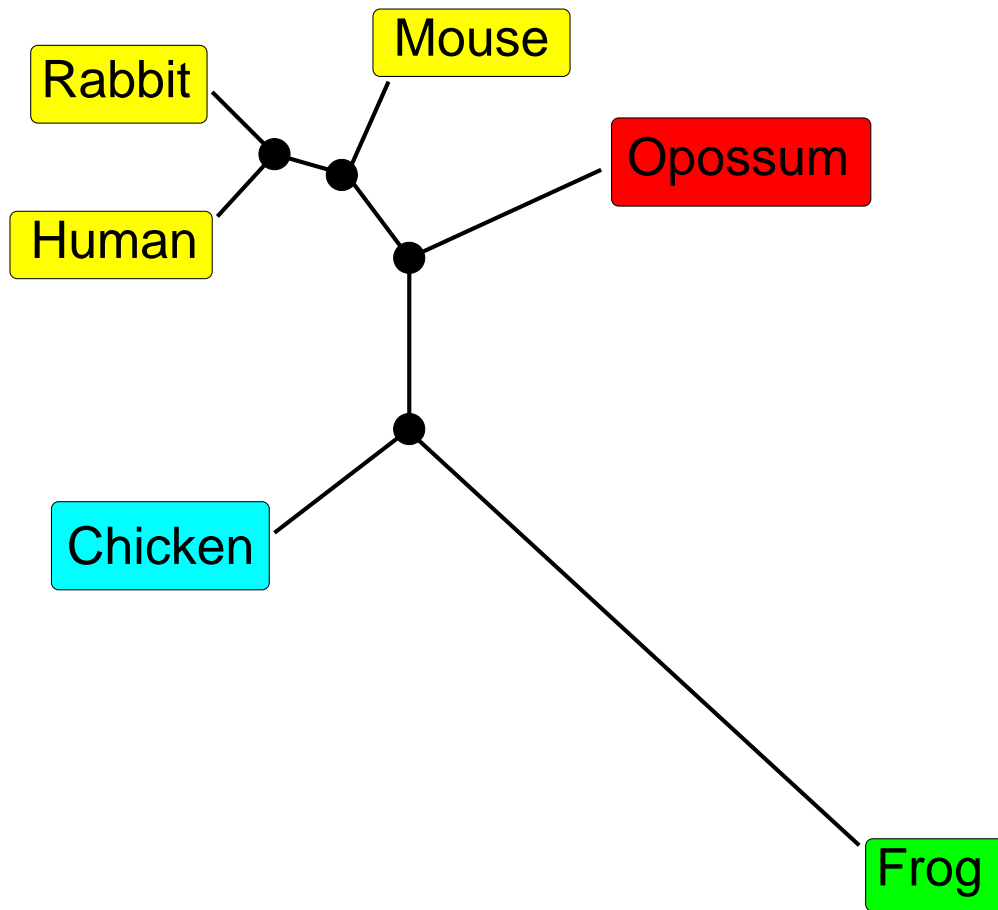
Email: dirk@bioss.ac.uk

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

Rooted Phylogenetic Tree



Unrooted Phylogenetic Tree



--> Topology
--> Branch lengths

Inferring Phylogeny from Pairwise Distances

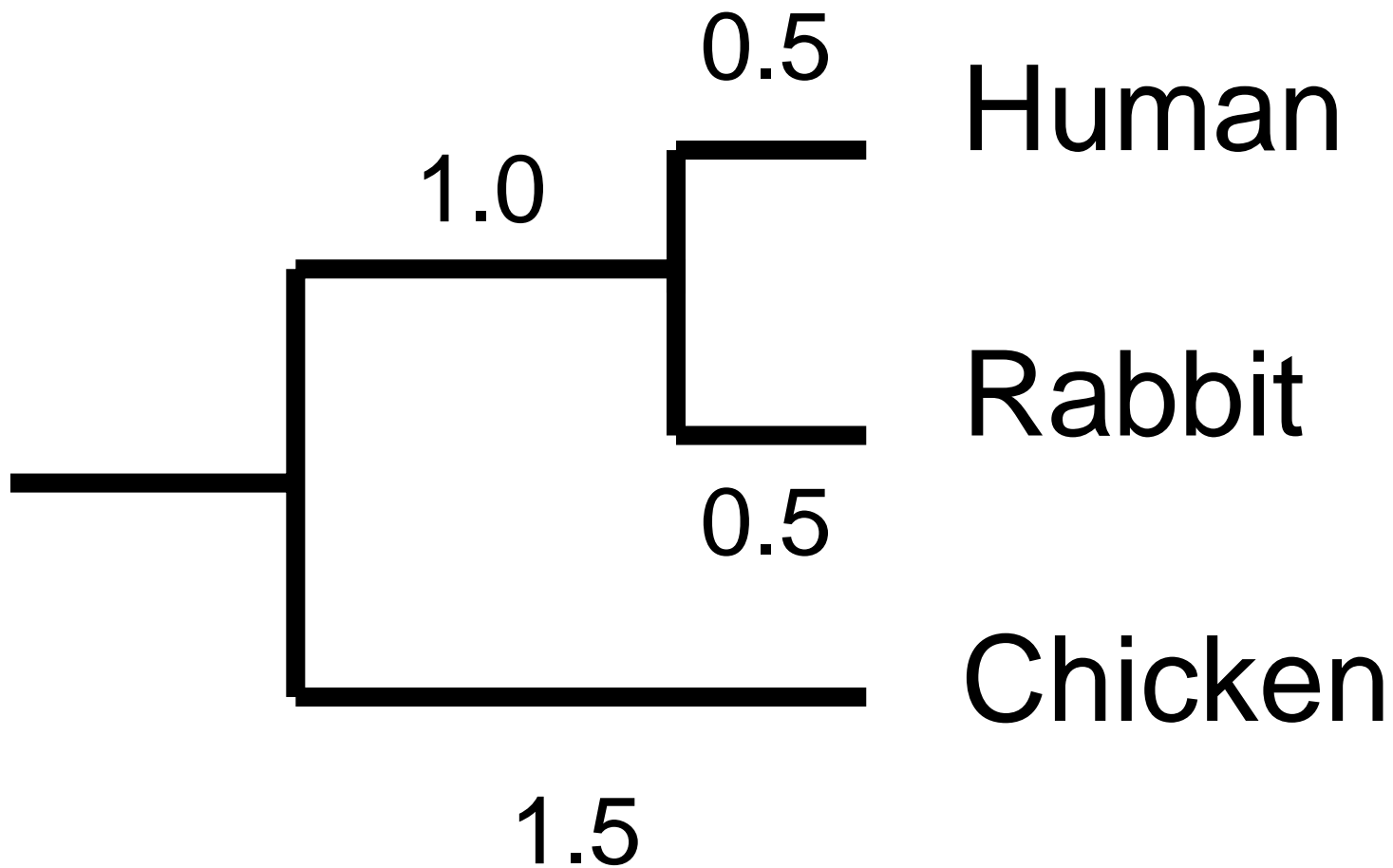
Human ... T G T A T C G C T C ...
 Rabbit ... T G T G T C G C T C ...

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

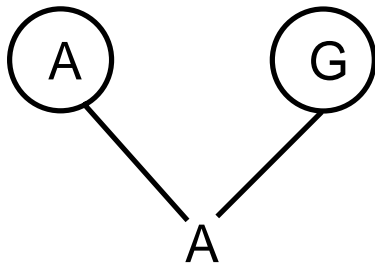
- Naive distance measure: **Hamming distance** $d_0 =$
Proportion of sites at which the two sequences differ.

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**.

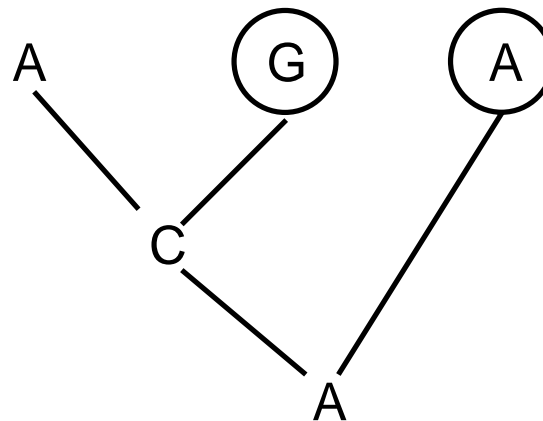
Single substitution

1 change, 1 difference



Multiple substitution

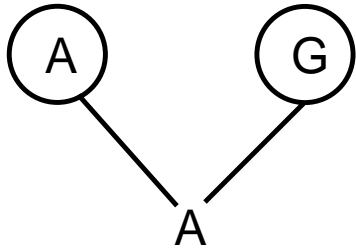
2 changes, 1 difference



Multiple Hits and Reversals

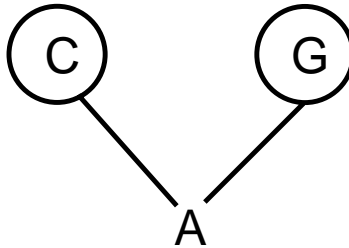
Single substitution

1 change, 1 difference



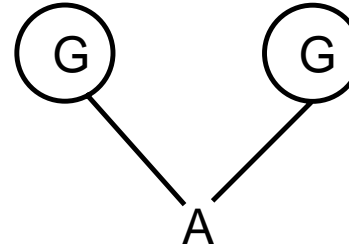
Coincidental substitution

2 changes, 1 difference



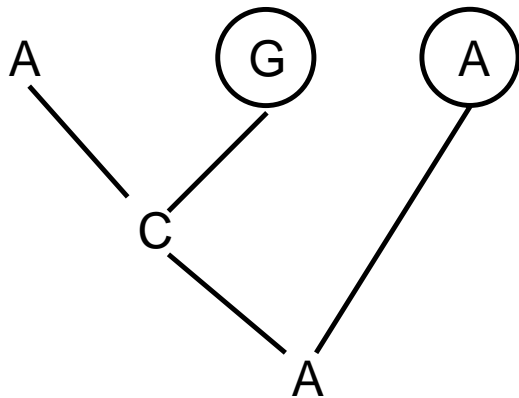
Parallel substitution

2 changes, no difference



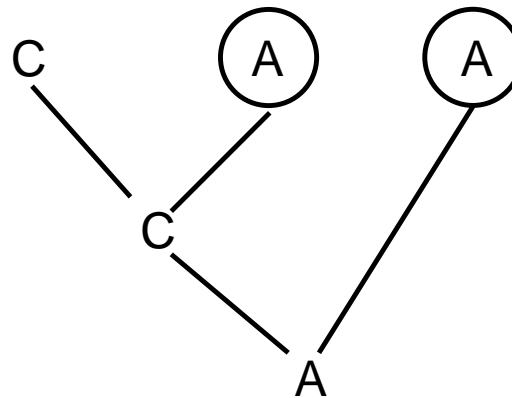
Multiple substitution

2 changes, 1 difference



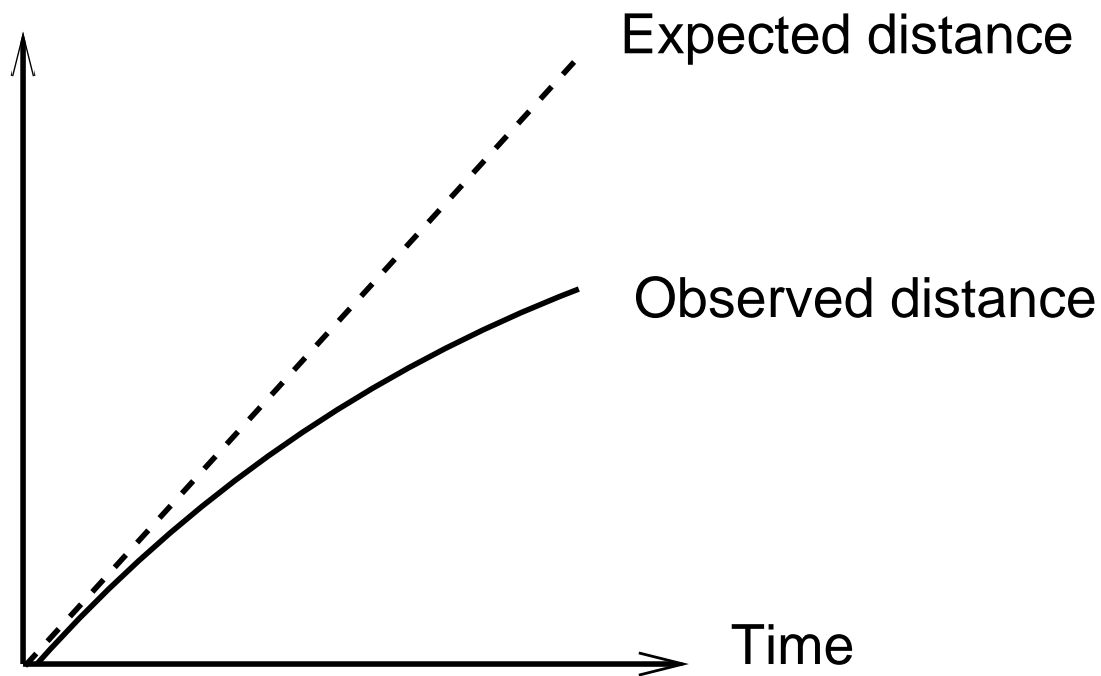
Back substitution

2 changes, no difference



Observed and Expected Genetic Distances

Sequence distance



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

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

Inferring Phylogeny by Clustering: UPGMA

Definition

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

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

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.

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 .
-

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.

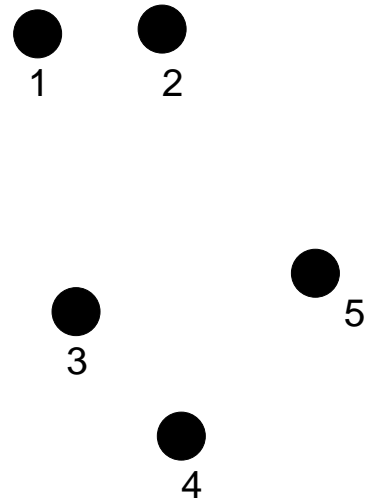
Iteration

- 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

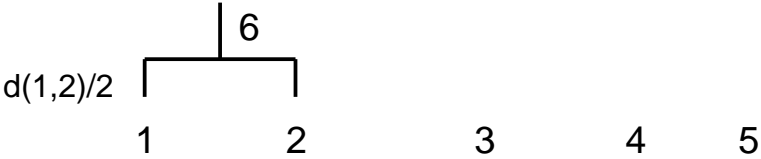
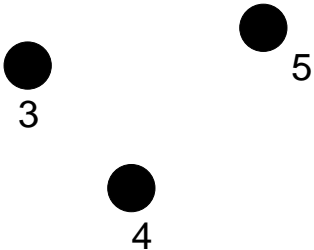
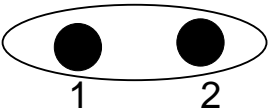
- When only two clusters i, j remain, place the root at height $d_{ij}/2$.
-

Inferring Phylogeny by Clustering: UPGMA

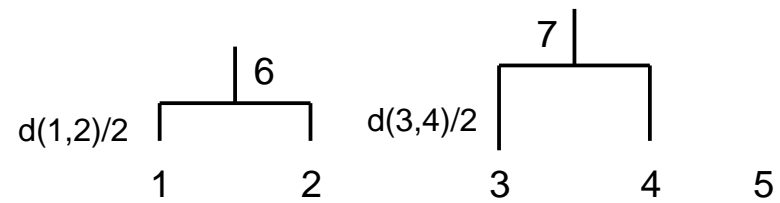
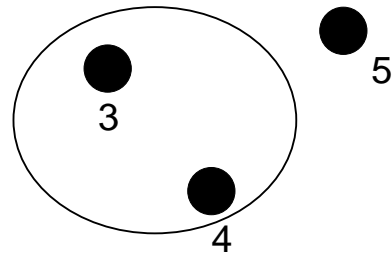
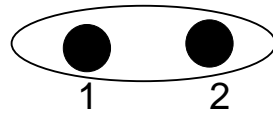


1 2 3 4 5

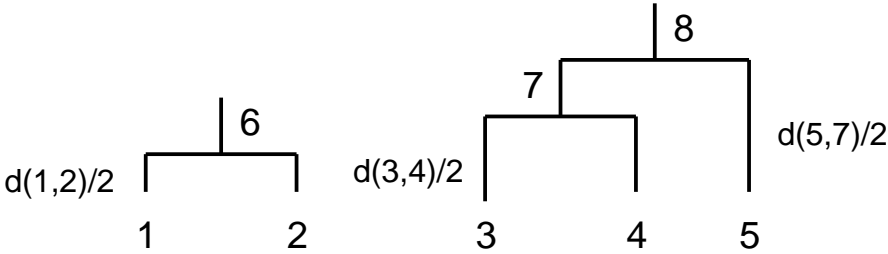
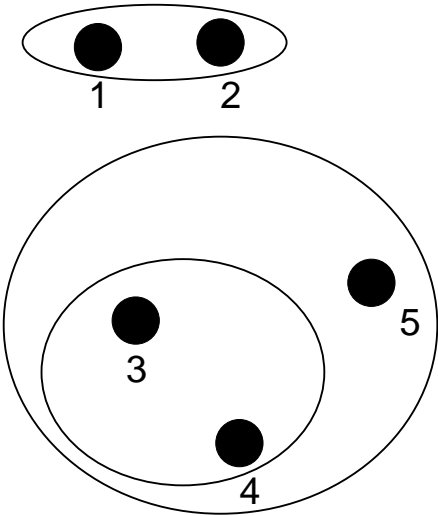
Inferring Phylogeny by Clustering: UPGMA



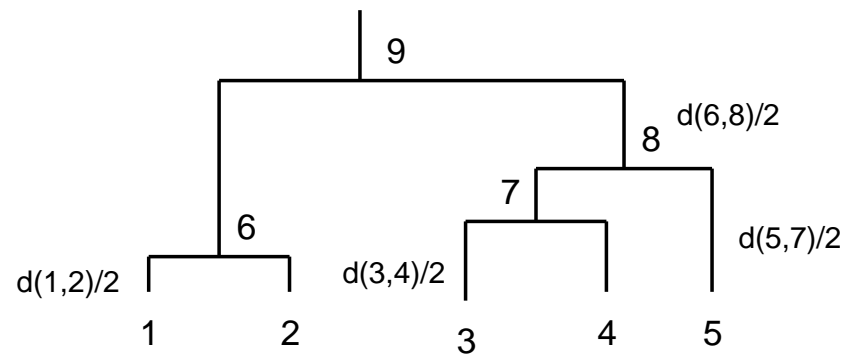
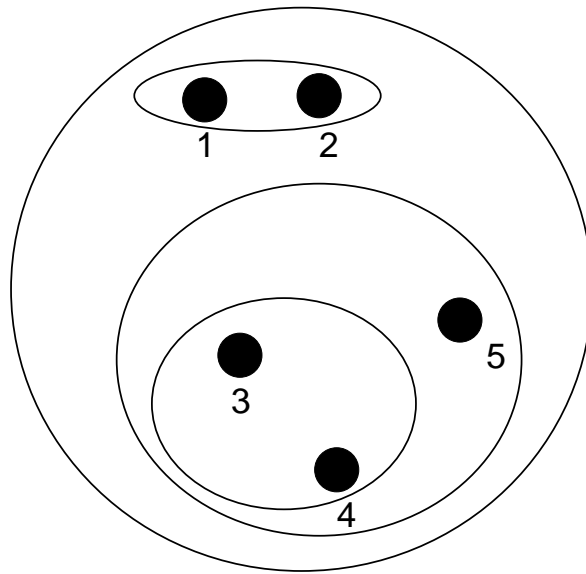
Inferring Phylogeny by Clustering: UPGMA



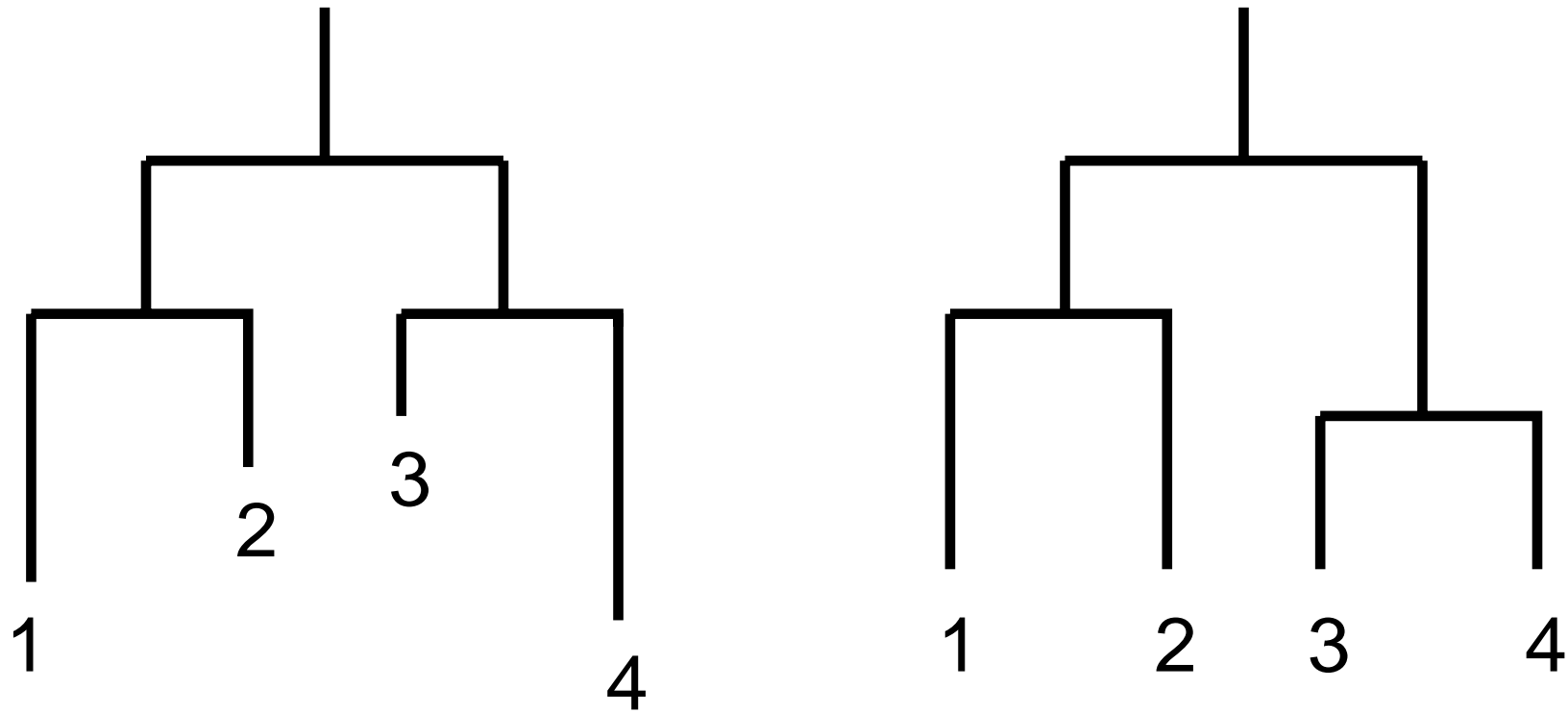
Inferring Phylogeny by Clustering: UPGMA



Inferring Phylogeny by Clustering: UPGMA



Limitation of UPGMA: Ultrametric Trees



Tree Metric

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

Tree Metric

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

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

Tree Metric

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

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

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

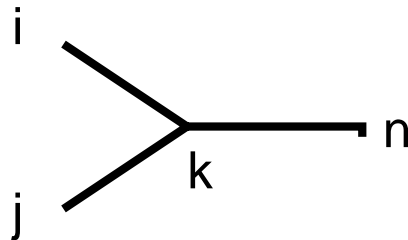
Tree Metric

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

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

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

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



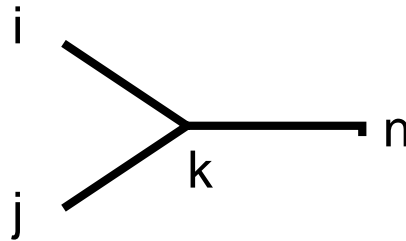
Tree Metric

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

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

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

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



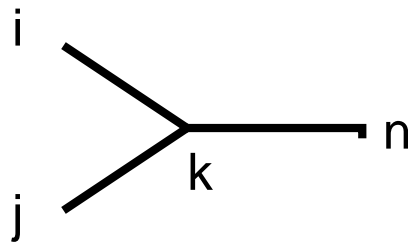
$$d_{in} = d_{ik} + d_{kn}$$

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

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

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

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

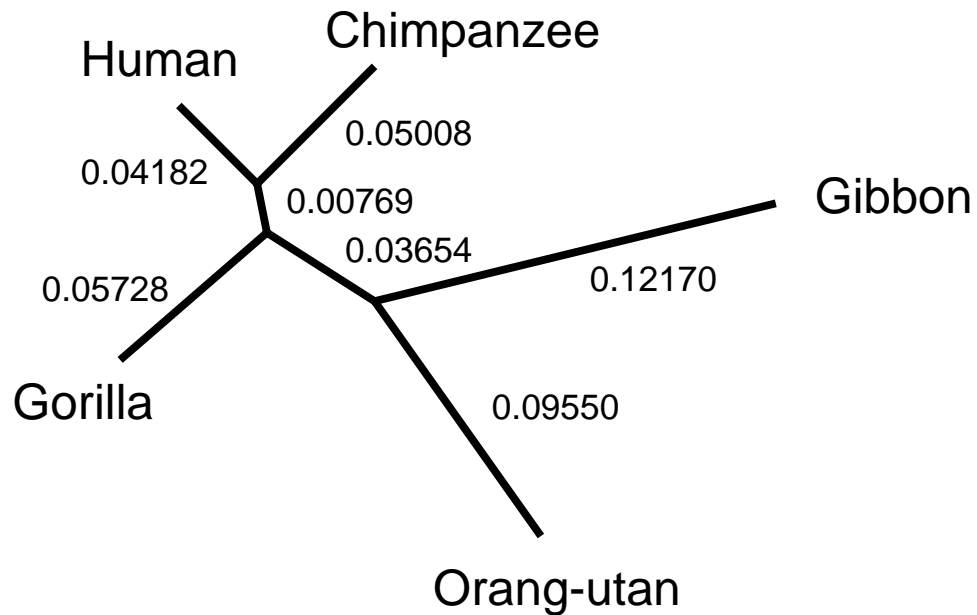
- Find pair of node (i, j) that minimize D_{ij} .
- Replace (i, j) by new node k with new distances:

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

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.



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
-