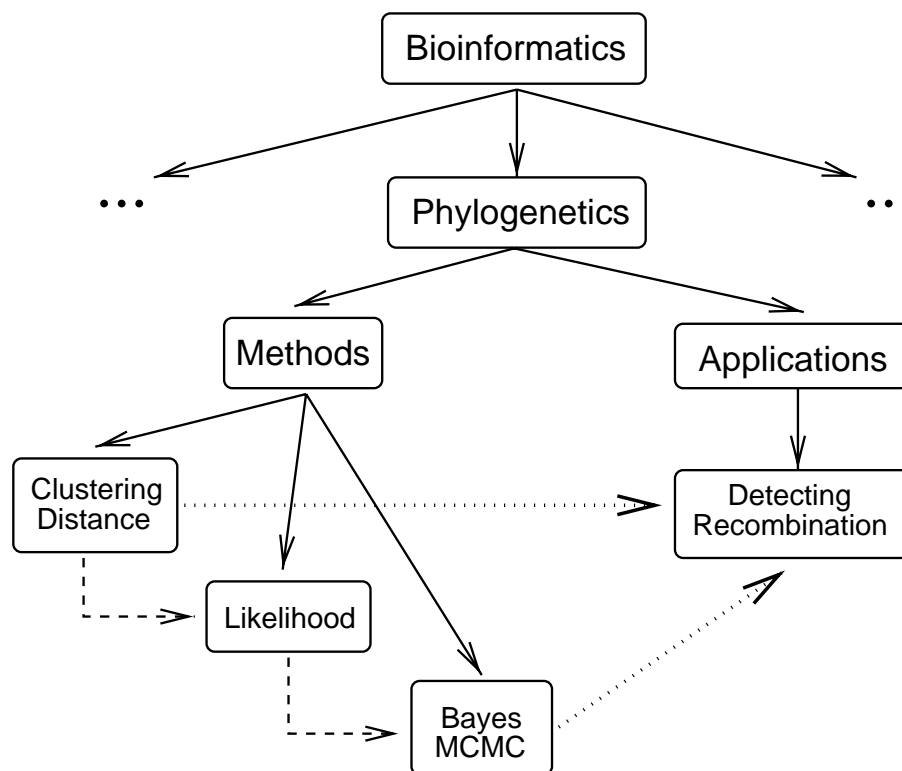
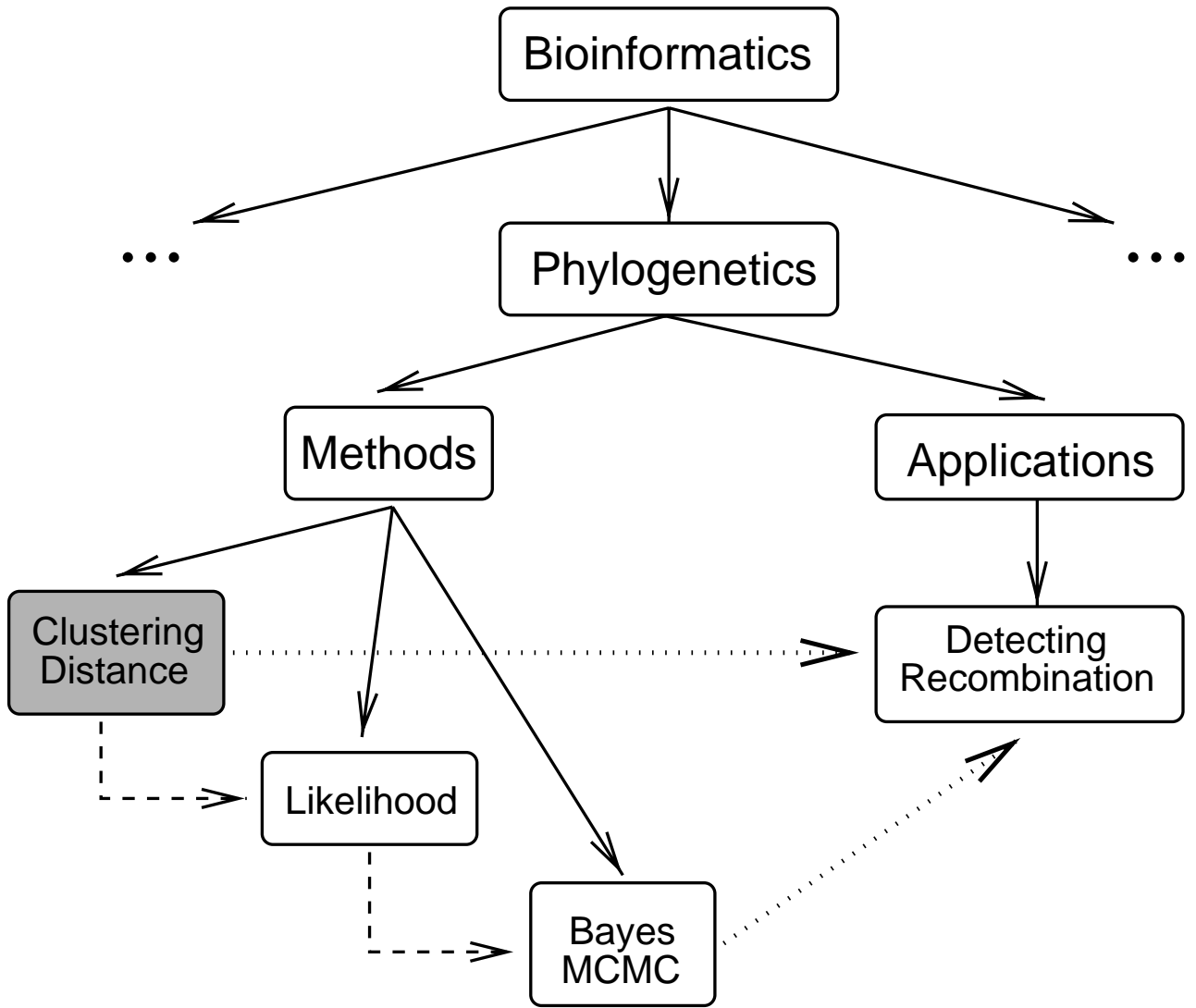

Paradigm Shifts in Statistical Bioinformatics

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>





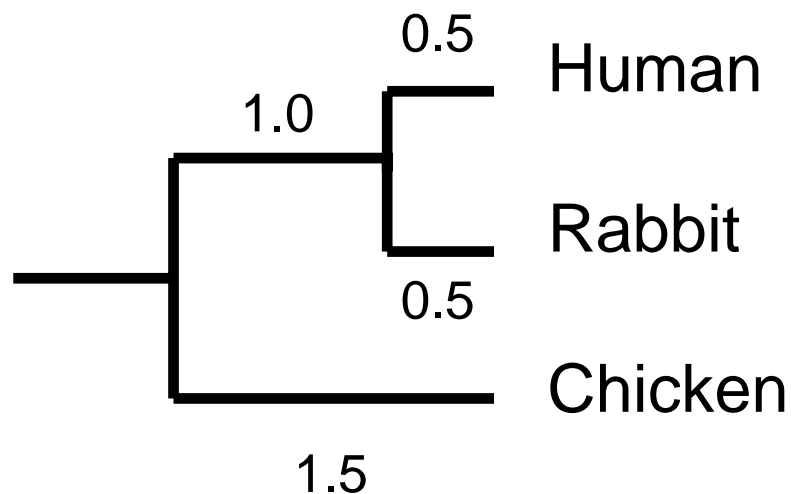
Inferring Phylogeny from Pairwise Distances

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

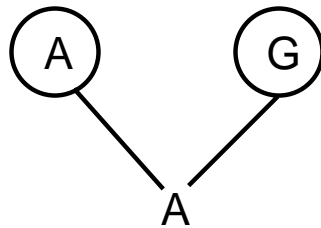


Genetic Distance

- 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** .
 $d_0(t \rightarrow \infty) = 3/4$.

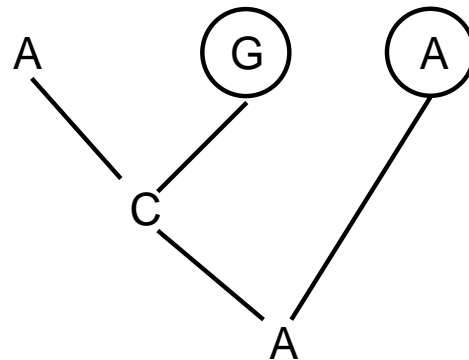
Single substitution

1 change, 1 difference

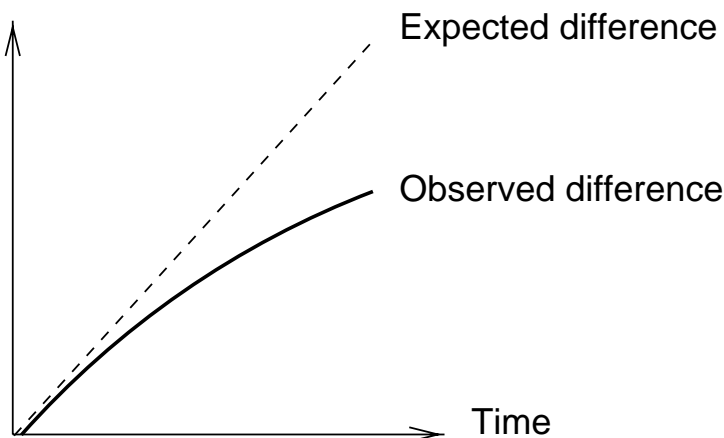


Multiple substitution

2 changes, 1 difference



Sequence difference



$$d = -\frac{3}{4} \ln \left(1 - \frac{4}{3} d_0 \right)$$

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

Initialisation

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

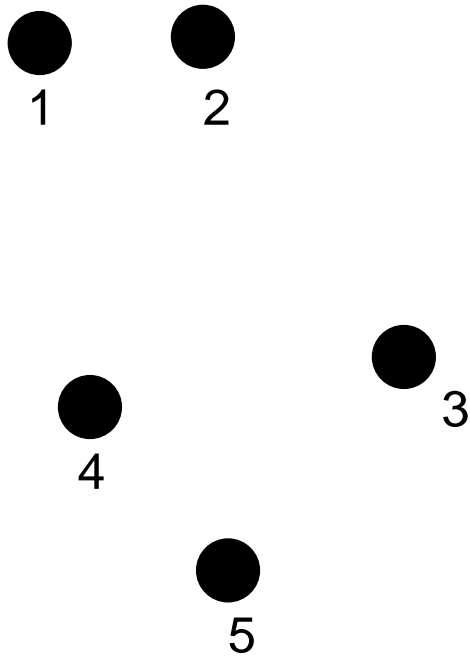
Iteration

- 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

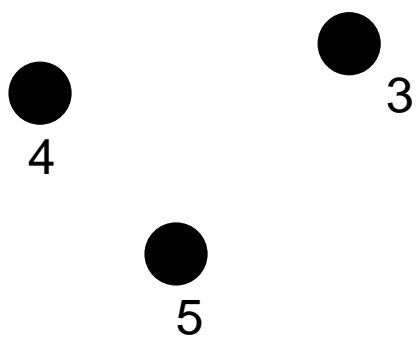
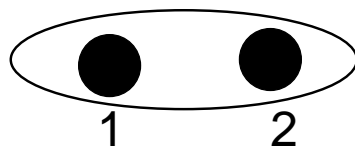
- When only two clusters i, j remain, place the root at height $d_{ij}/2$.

Inferring Phylogeny by Clustering: UPGMA

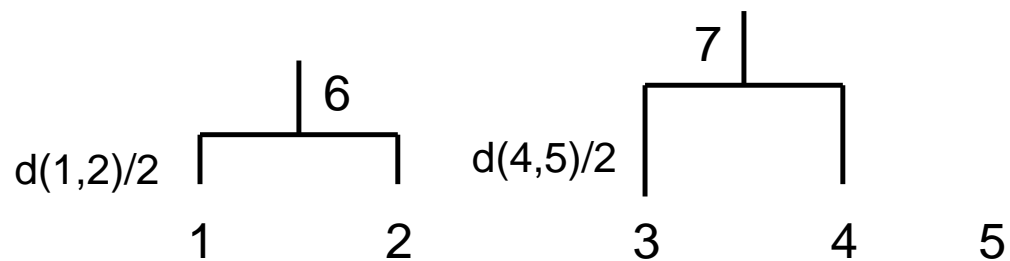
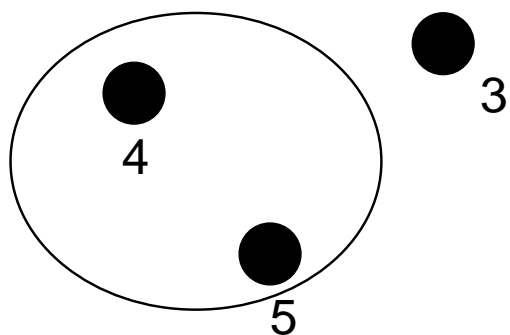
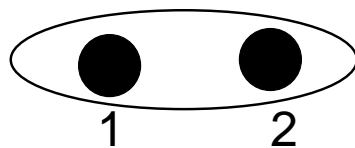


1 2 3 4 5

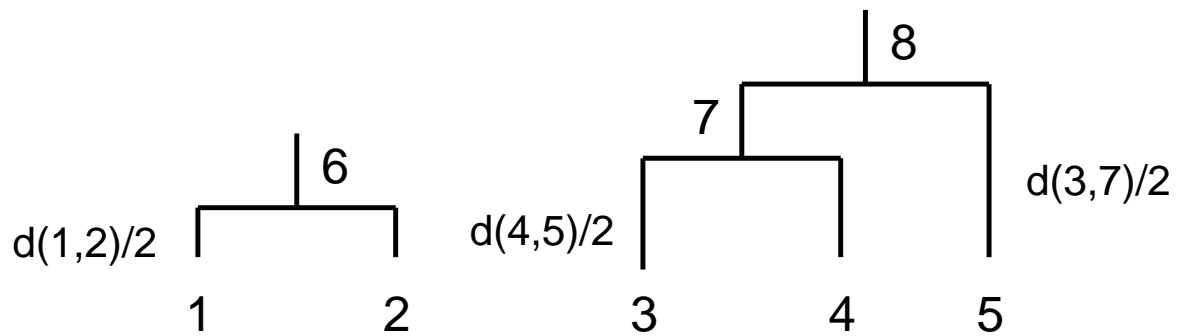
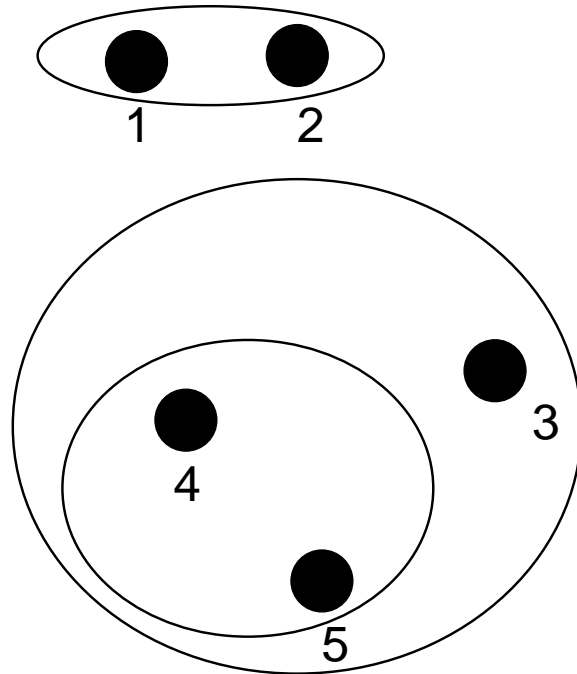
Inferring Phylogeny by Clustering: UPGMA



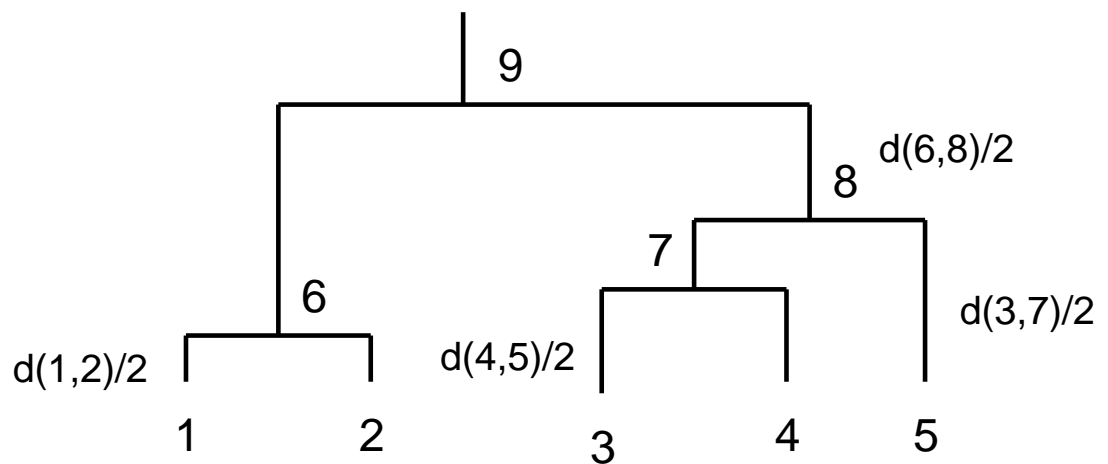
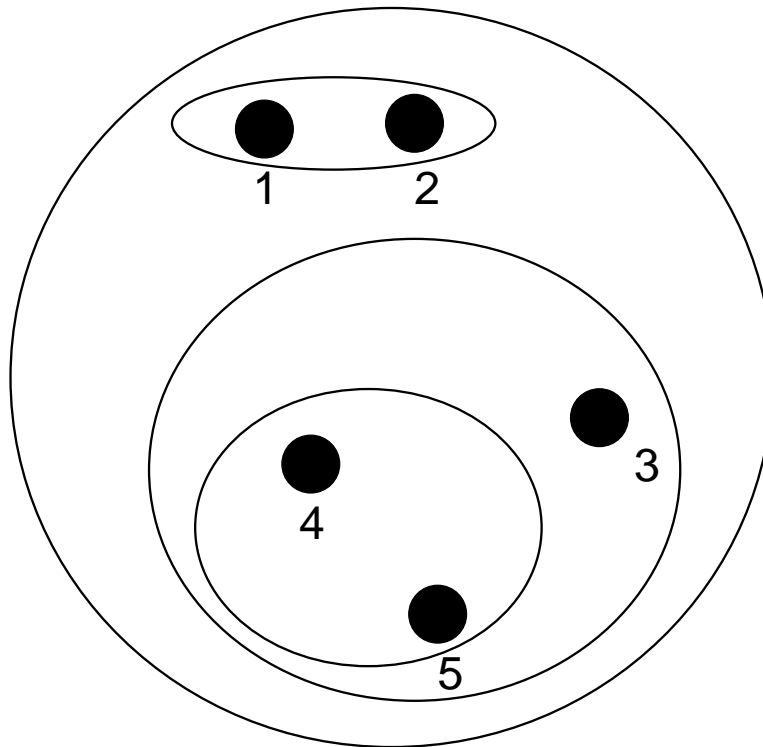
Inferring Phylogeny by Clustering: UPGMA



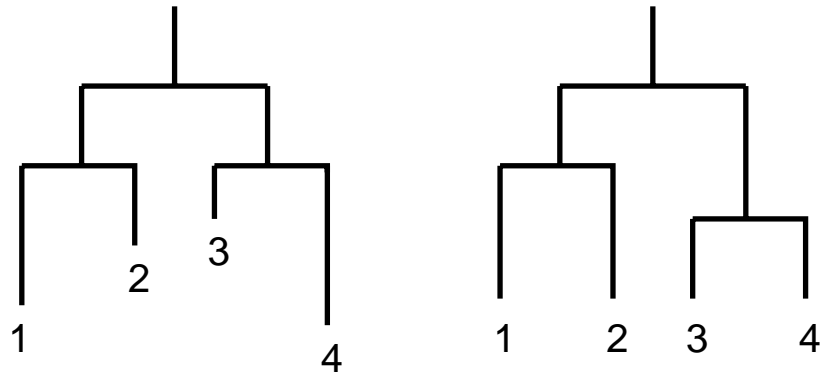
Inferring Phylogeny by Clustering: UPGMA



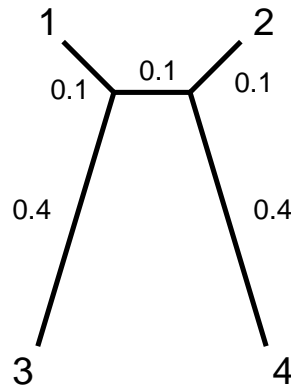
Inferring Phylogeny by Clustering: UPGMA



Limitation of UPGMA: Ultrametric Trees



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



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

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

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

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

Inferring Phylogeny by Clustering: Neighbour Joining

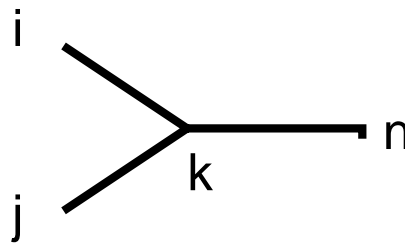
Tree metric

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

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

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

Triangle Inequality: $d_{ac} \leq d_{ab} + d_{bc} \longrightarrow d_{ac} = d_{ab} + d_{bc}$



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

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

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

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

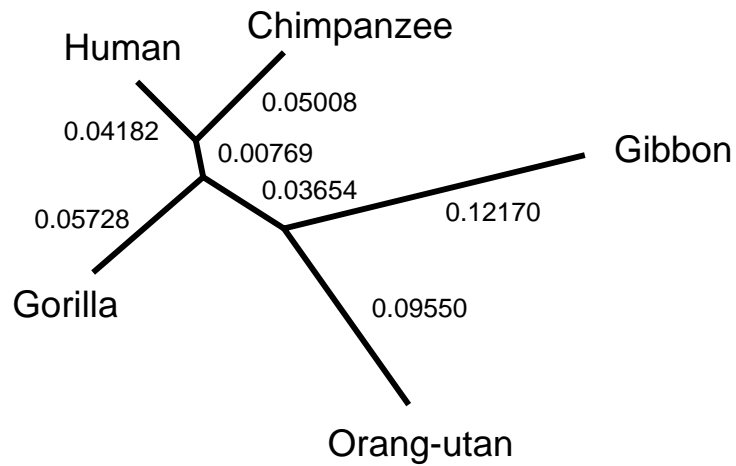
Iteration

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

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

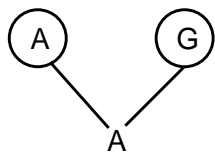
Application of Neighbour Joining

	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	-



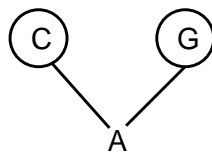
Single substitution

1 change, 1 difference



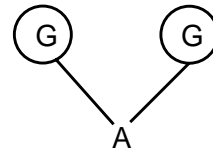
Coincidental substitution

2 changes, 1 difference



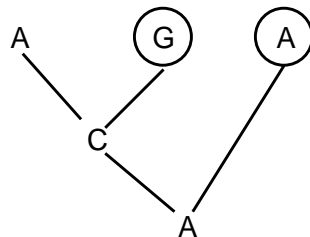
Parallel substitution

2 changes, no difference



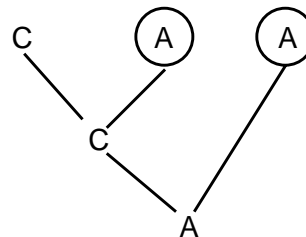
Multiple substitution

2 changes, 1 difference



Back substitution

2 changes, no difference



Objections to Distance Methods and Clustering

- 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

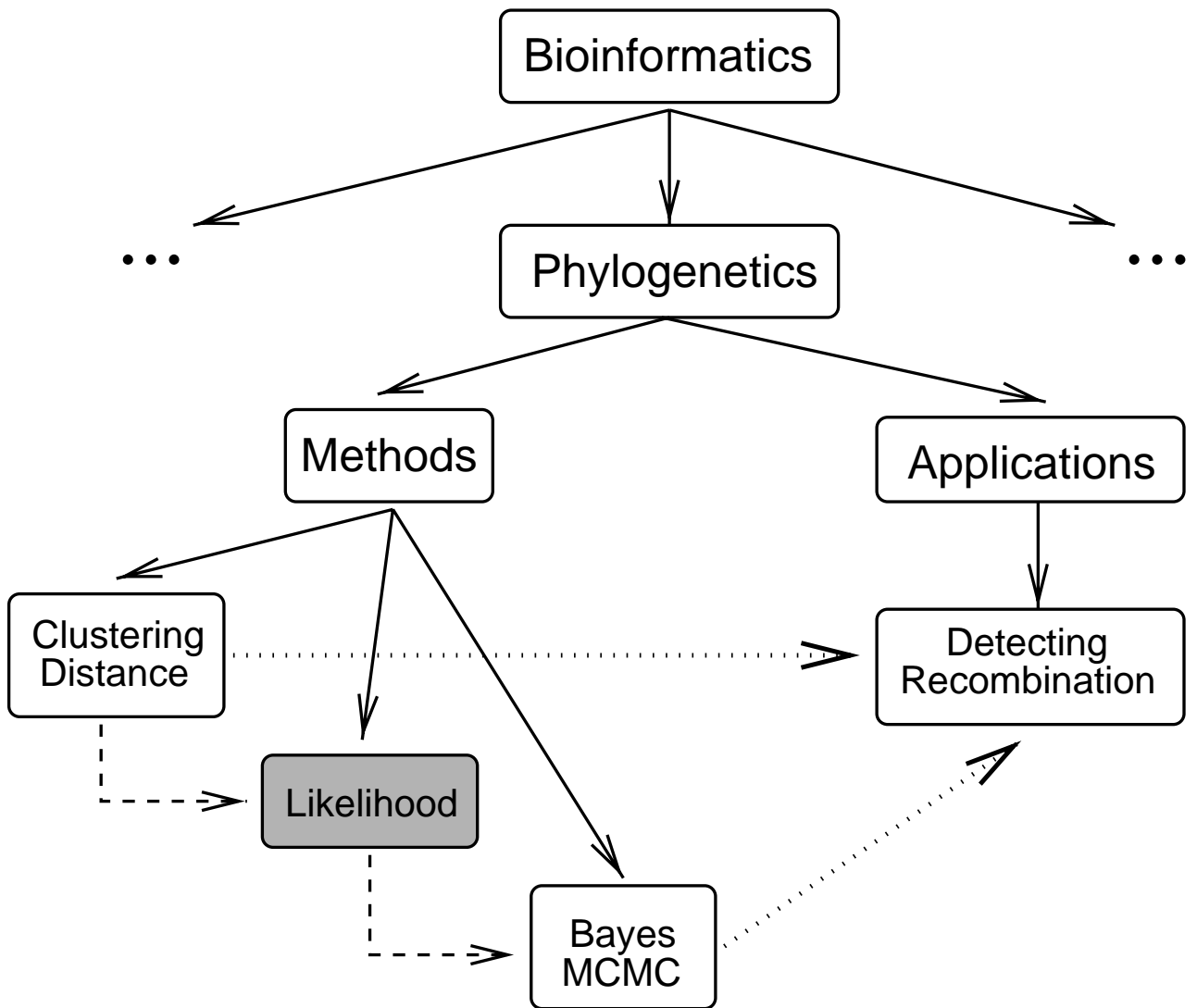
- Uninterpretable branch lengths

- $d_{ij}^{tree} < d_{ij}^{obs}$ 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

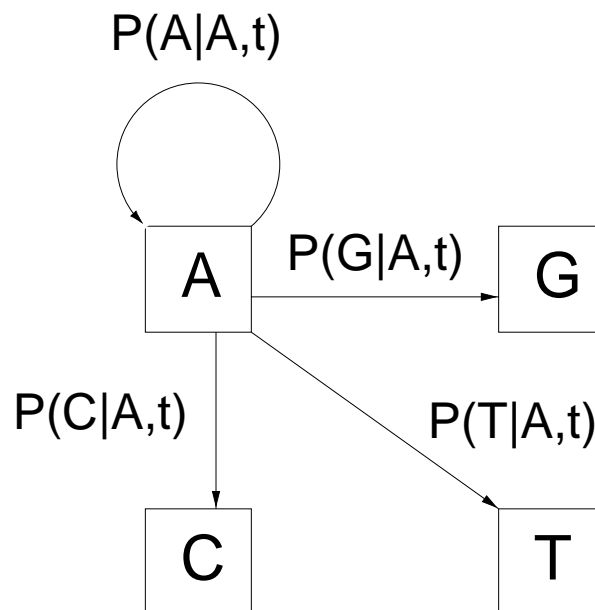
- to evaluate the quality of the tree
- to evaluate competing hypothesis



Evolution

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 C A T C G T T C ...



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

Probabilistic Models 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** :

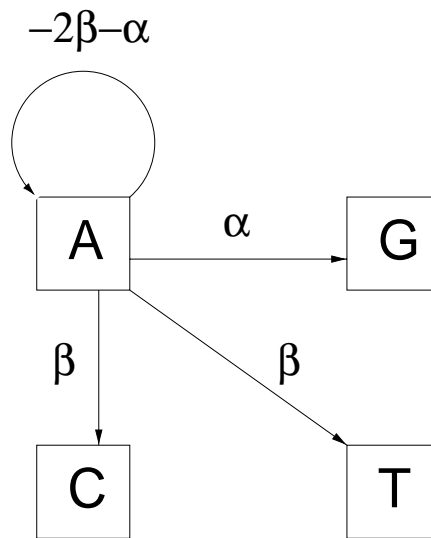
$$P(y_{t+\Delta t}|y_t) = P(y_{\Delta t}|y_0)$$

- 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

$$\begin{aligned}\mathbf{P}(0) &= \mathbf{I} & \mathbf{P}(dt) - \mathbf{P}(0) &= \mathbf{R}dt \\ \mathbf{P}(t + dt) &= \mathbf{P}(dt)\mathbf{P}(t) = (\mathbf{I} + \mathbf{R}dt)\mathbf{P}(t) \\ \frac{d\mathbf{P}}{dt} &= \mathbf{R}\mathbf{P} \implies \mathbf{P}(t) = e^{\mathbf{R}t}\end{aligned}$$



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

Transition Probabilities

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

$$f(t) = \frac{1}{4}(1 - e^{-4\beta t})$$

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

$$d(t) = 1 - 2f(t) - g(t)$$

Molecular time: $w := 4\beta t$

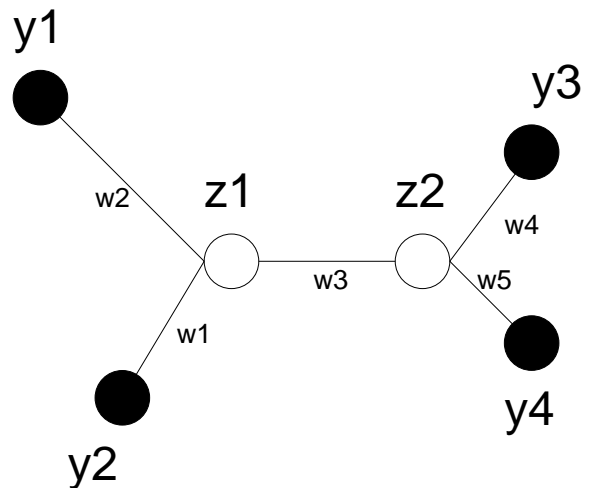
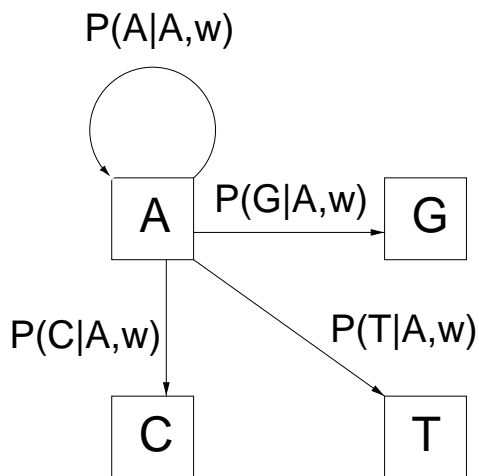
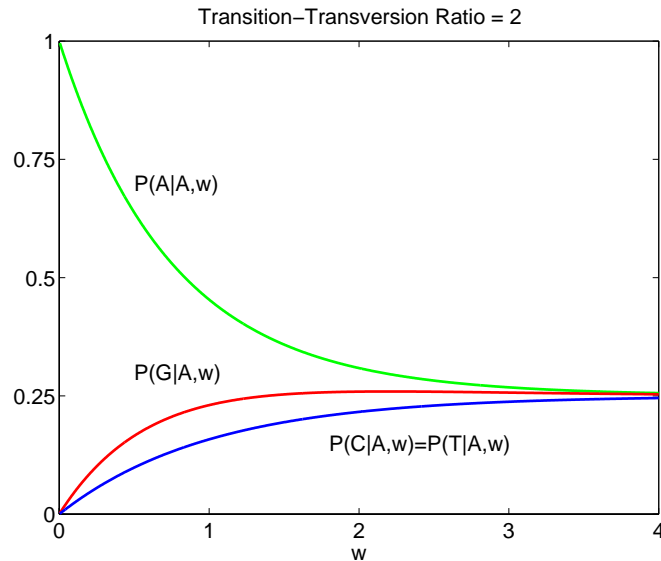
$$f(w) = \frac{1}{4}(1 - e^{-4w})$$

$$g(w) = \frac{1}{4}(1 + e^{-4w} - 2e^{-2(\tau+1)w})$$

$$d(w) = 1 - 2f(w) - g(w)$$

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

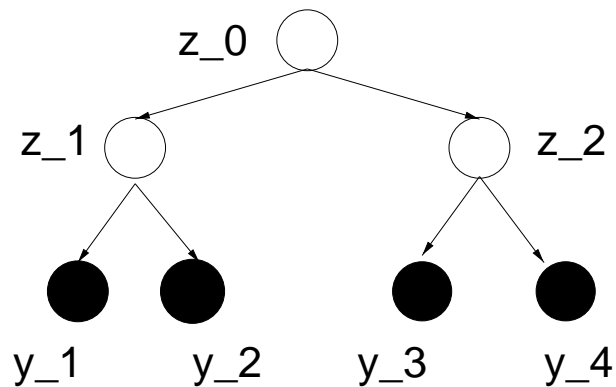
A Probabilistic Model of Evolution



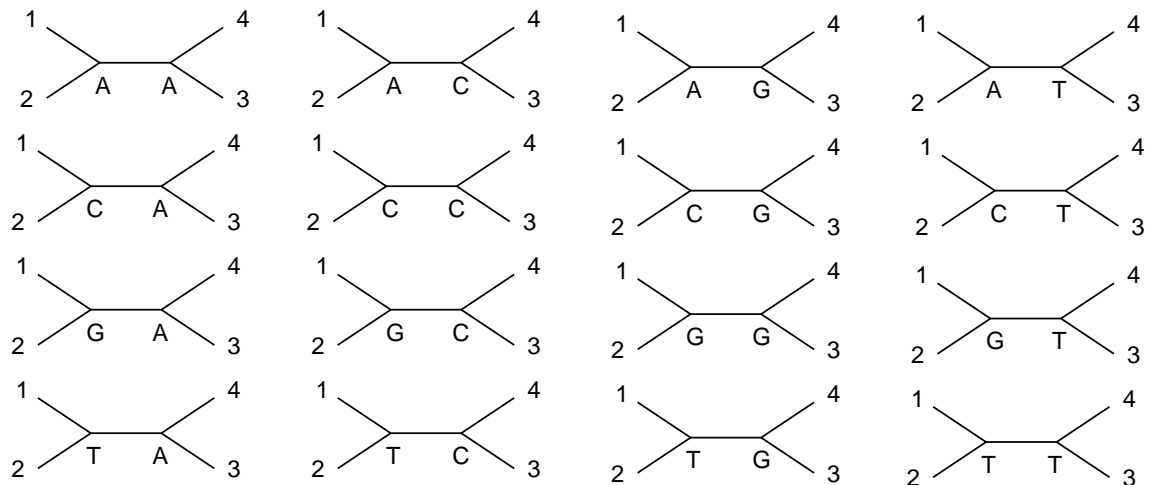
$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) = P(y_1 | z_1, w_2) P(y_2 | z_1, w_1) P(z_2 | z_1, w_3) P(y_3 | z_2, w_4) P(y_4 | z_2, w_5)$$

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

Tree Likelihood: Factorisation and Marginalisation



$$P(y_1, y_2, y_3, y_4, z_0, z_1, z_2) = P(z_0) P(z_1|z_0) P(z_2|z_0) P(y_1|z_1) P(y_2|z_1) P(y_3|z_2) P(y_4|z_2)$$



$$P(y_1, y_2, y_3, y_4) = \sum_{z_0, z_1, z_2} P(y_1, y_2, y_3, y_4, z_0, z_1, z_2)$$

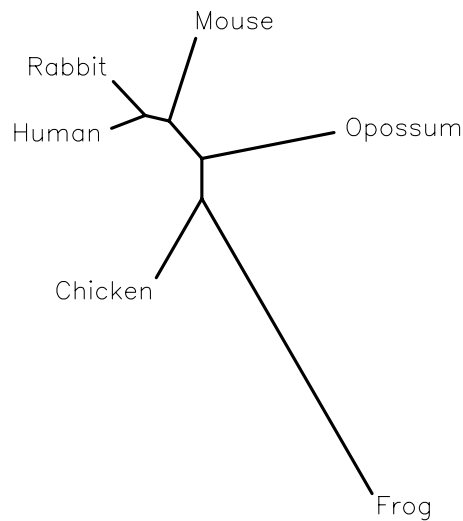
Probabilistic Approach to Phylogeny

Frog	GT C GCGGGTCAAAC TTTCCGTCTCGCG
Chicken	AG C ATCGTTCTATTTTACCGGCTCCCG
Human	TG T ATCGCTCAAGATTGCCATCGCGCG
Rabbit	TG T GTCGCTCAAGATTGCCATCGCGCG
Mouse	TG T CGTGGTCTAGATTGCCATCGCGCG
Opossum	TG T ATCGCTCTAGTTTGCCAGCTCCCG

$$\mathbf{D} = (\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N)$$

$$P(\mathbf{D}|\mathbf{w}, S) = \prod_{t=1}^N P(\mathbf{y}_t|\mathbf{w}, S)$$

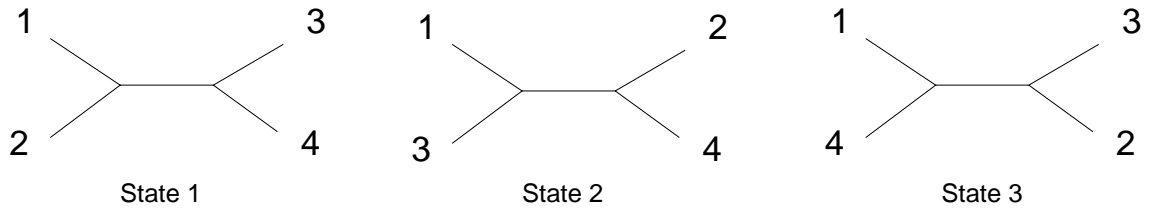
Optimise topology S and branch lengths \mathbf{w} with maximum likelihood



Maximum Likelihood

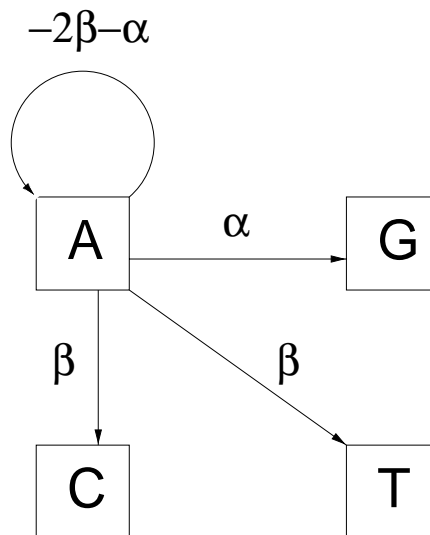
Maximise the likelihood of $L(S, \mathbf{w}, \mathbf{R}) = \ln P(D|S, \mathbf{w}, \mathbf{R})$

- Tree topology S



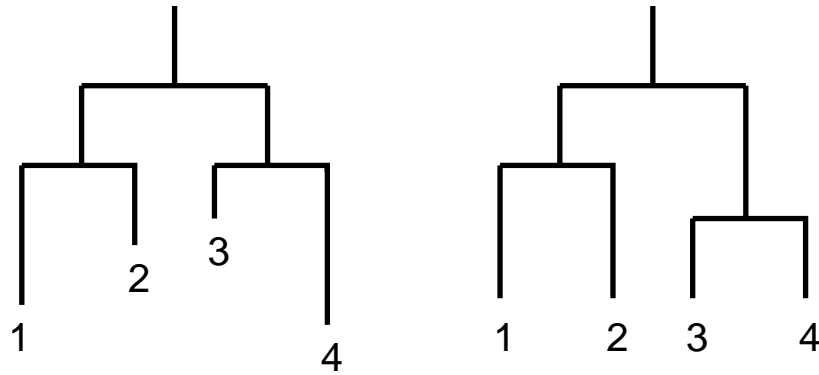
- Branch lengths $\mathbf{w} = (w_1, w_2, w_3, w_4, w_5)$

- Evolutionary parameters: Rate matrix \mathbf{R}



Hypothesis Testing

Nested models

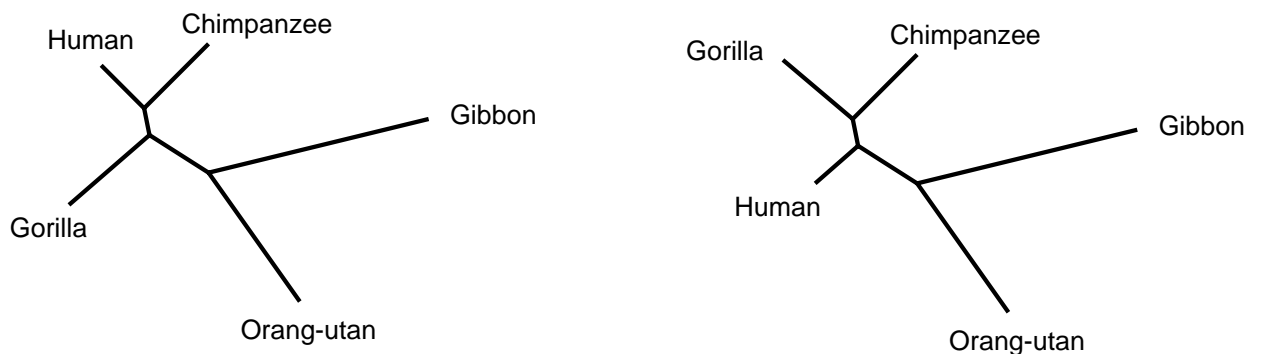


H_0 : ultrametric tree (molecular clock)

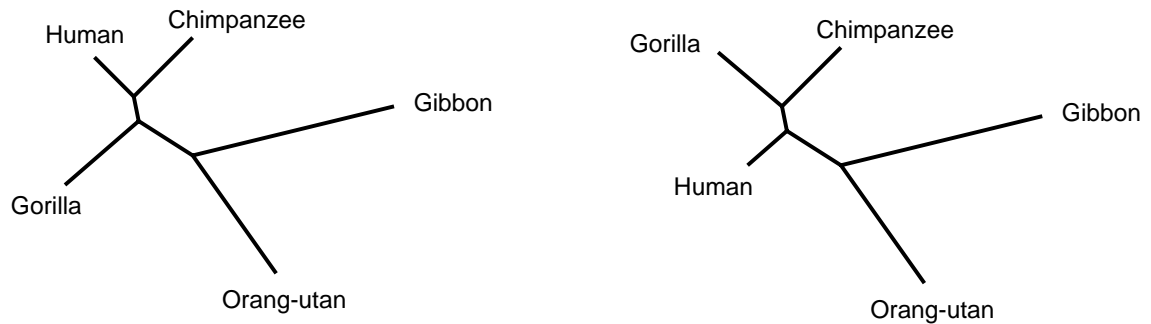
($K - 1$ constraints, $K =$ number of leaf nodes).

Likelihood ratio test: $2(L_1 - L_0) \sim \chi^2_{(K-2)}$

Non-nested models



Bootstrapping

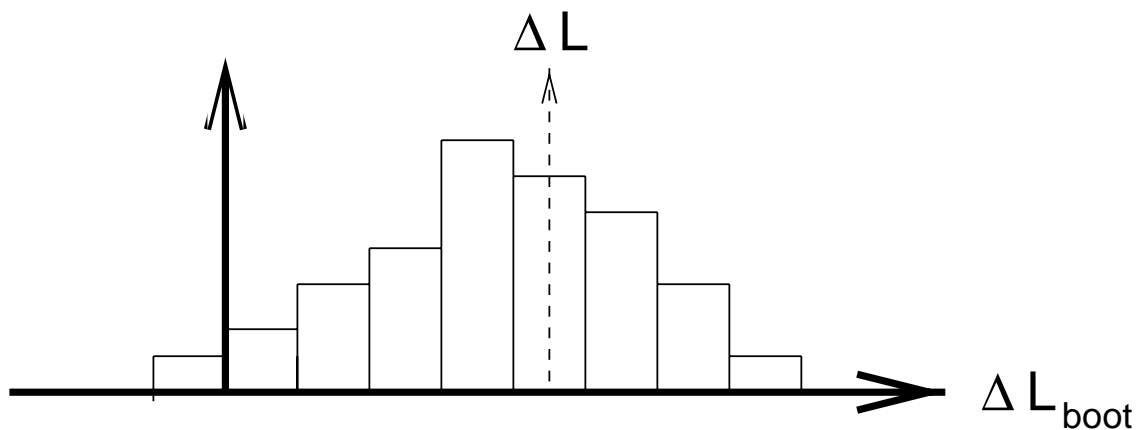


$$\begin{aligned} L_A &= \ln P(D|S_A, \hat{\mathbf{w}}_A) \\ L_B &= \ln P(D|S_B, \hat{\mathbf{w}}_B) \end{aligned} \longrightarrow \Delta L = L_A - L_B \neq 0?$$

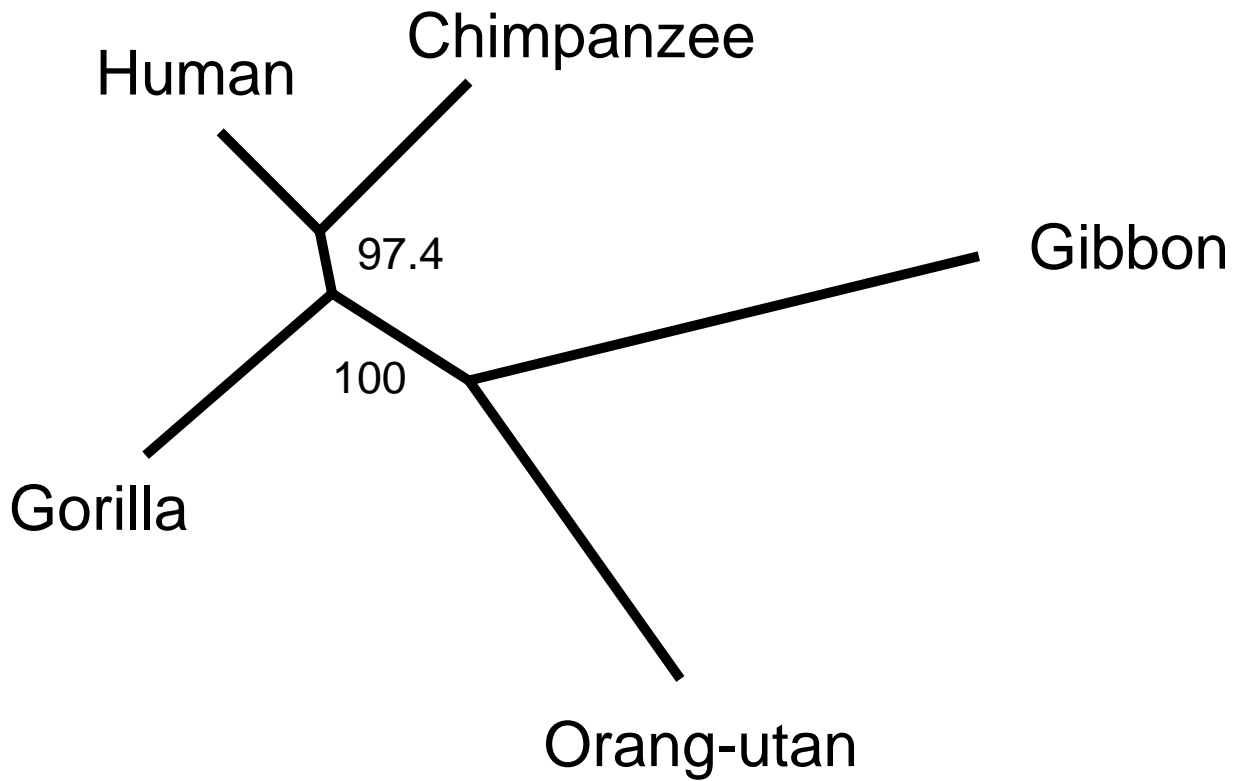
Resample with replacement from D

$$\begin{aligned} D &= \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} \longrightarrow \begin{aligned} D_1 &= \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2, \mathbf{x}_4\} \\ D_2 &= \{\mathbf{x}_2, \mathbf{x}_2, \mathbf{x}_1, \mathbf{x}_1\} \\ &\vdots \\ D_B &= \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_4, \mathbf{x}_1\} \end{aligned} \end{aligned}$$

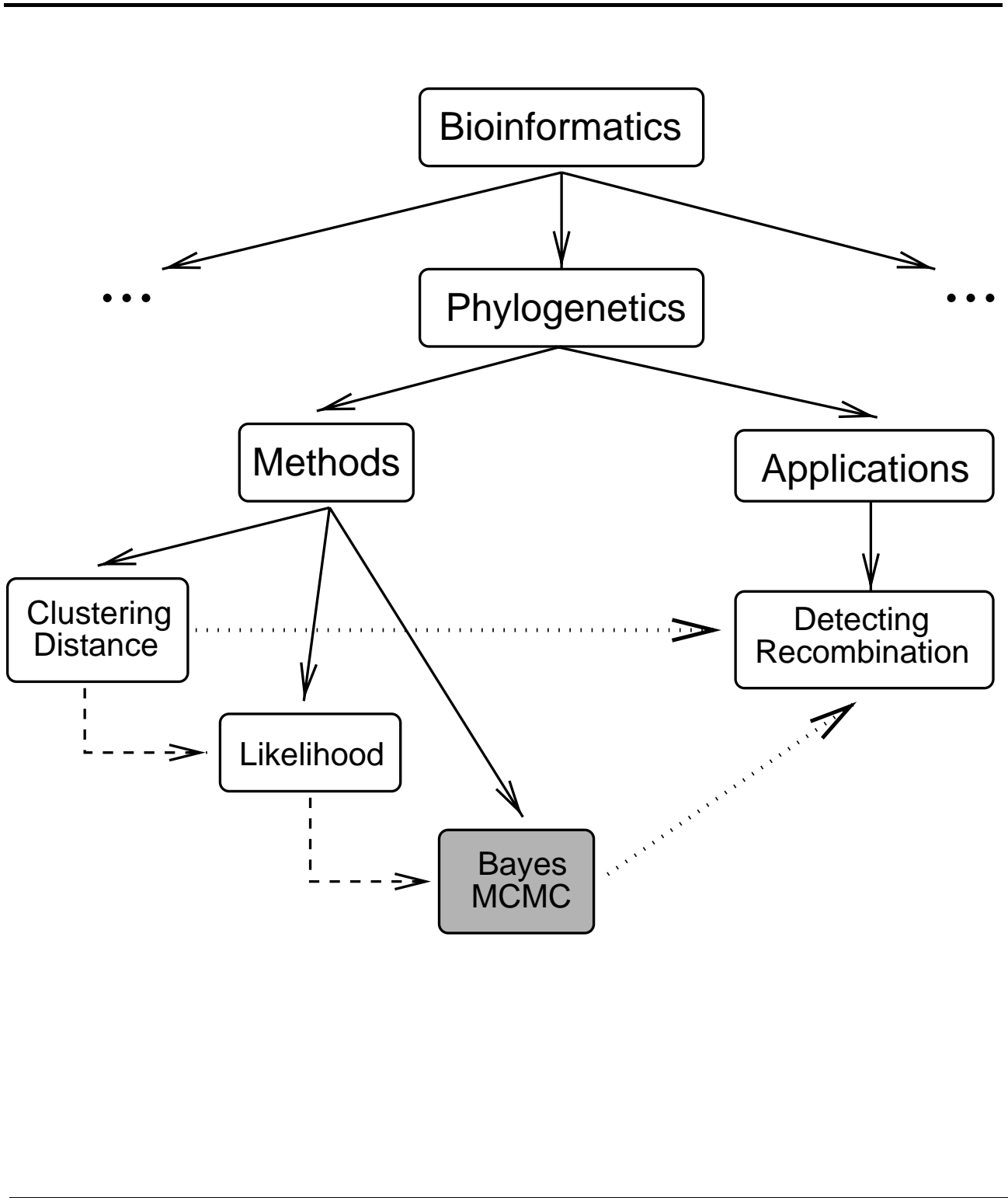
Bootstrap distribution $\{\Delta L_b\}_{b=1}^B$



Monophyletic Groups



Clade	Probability
(Human Chimp)	0.974
(Human Chimp Gorilla)	1.0



Bayesian Model Selection

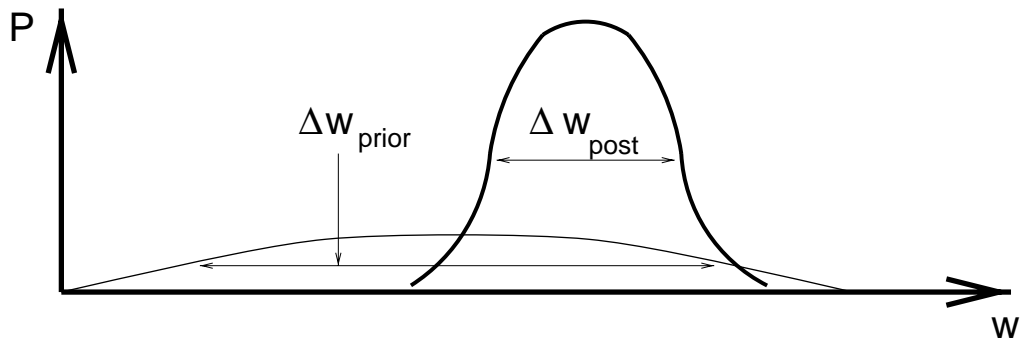
$$\begin{aligned} L_A &= \ln P(D|S_A, \hat{\mathbf{w}}_A) \\ L_B &= \ln P(D|S_B, \hat{\mathbf{w}}_B) \end{aligned} \longrightarrow \Delta L = L_A - L_B \neq 0?$$

Uniform prior: $P(S_i) = \text{const} \rightarrow$ Bayes Factor

$$\frac{P(S_A|D)}{P(S_B|D)} = \frac{P(D|S_A)P(S_A)}{P(D|S_B)P(S_B)} = \frac{P(D|S_A)}{P(D|S_B)}$$

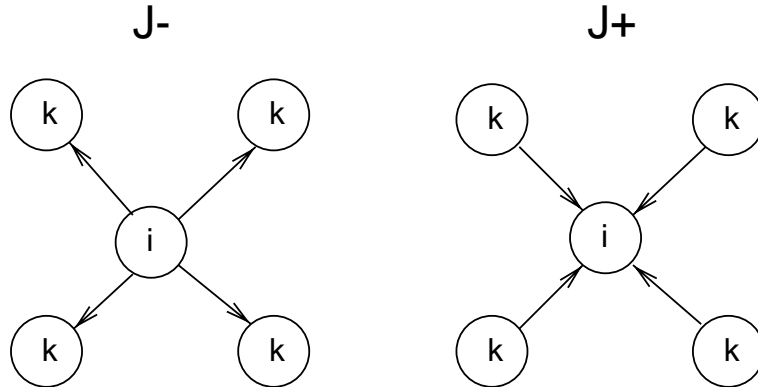
$$\begin{aligned} P(D|S) &= \int P(D, \mathbf{w}|S) = \int P(D|\mathbf{w}, S)P(\mathbf{w}|S)d\mathbf{w} \\ &\approx P(D|\hat{\mathbf{w}}, S) \frac{\Delta \mathbf{w}_{post}}{\Delta \mathbf{w}_{prior}} \end{aligned}$$

$$\ln P(D|S) = \ln P(D|\hat{\mathbf{w}}, S) + \ln \left(\frac{\Delta \mathbf{w}_{post}}{\Delta \mathbf{w}_{prior}} \right)$$



Metropolis-Hastings Algorithm

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



$$0 = \frac{d}{dt}P(\theta_i|D) = J_+ - J_- = \frac{\sum_k T(\theta_i|\theta_k)P(\theta_k|D)}{\sum_k T(\theta_k|\theta_i)P(\theta_i|D)}$$

Detailed balance:

$$T(\theta_i|\theta_k)P(\theta_k|D) = T(\theta_k|\theta_i)P(\theta_i|D) \Rightarrow \frac{d}{dt}P(\theta_i|D) = 0$$

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

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:

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

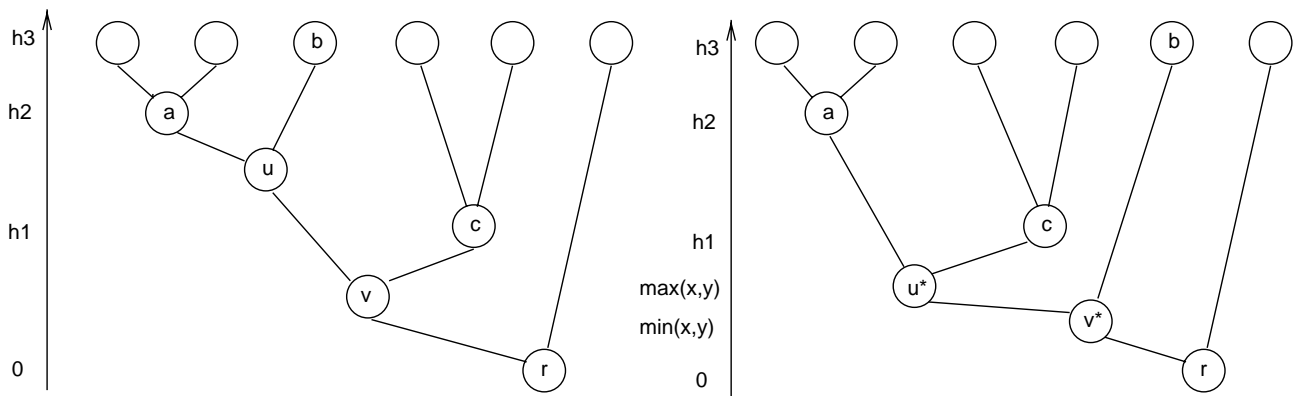
$$a(\theta_k|\theta_i) = \min \left\{ \frac{P(D|\theta_k)P(\theta_k)q(\theta_i|\theta_k)}{P(D|\theta_i)P(\theta_i)q(\theta_k|\theta_i)}, 1 \right\}$$

Algorithm

After equilibration, sample $\{\theta_i\}$.

$$\int f(\theta)P(\theta|D) = \sum_i f(\theta_i)$$

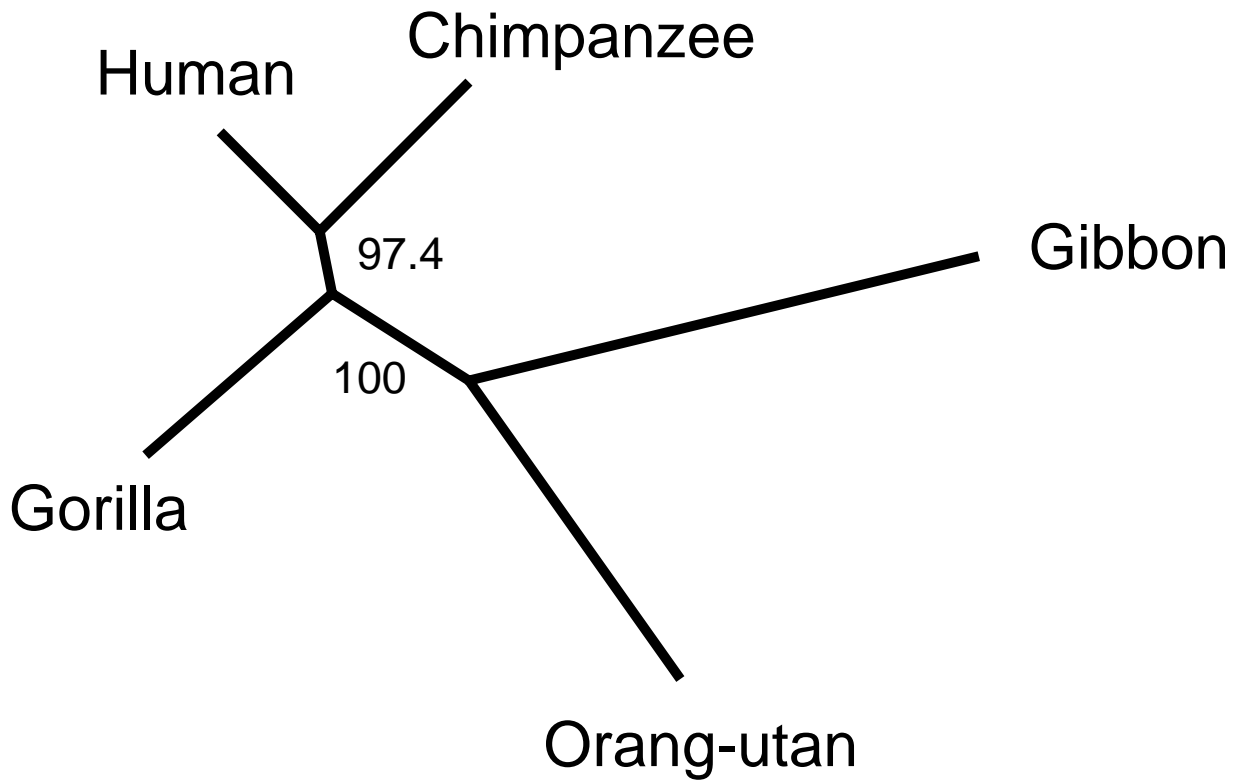
Moves in Tree Space and Hastings Ratio



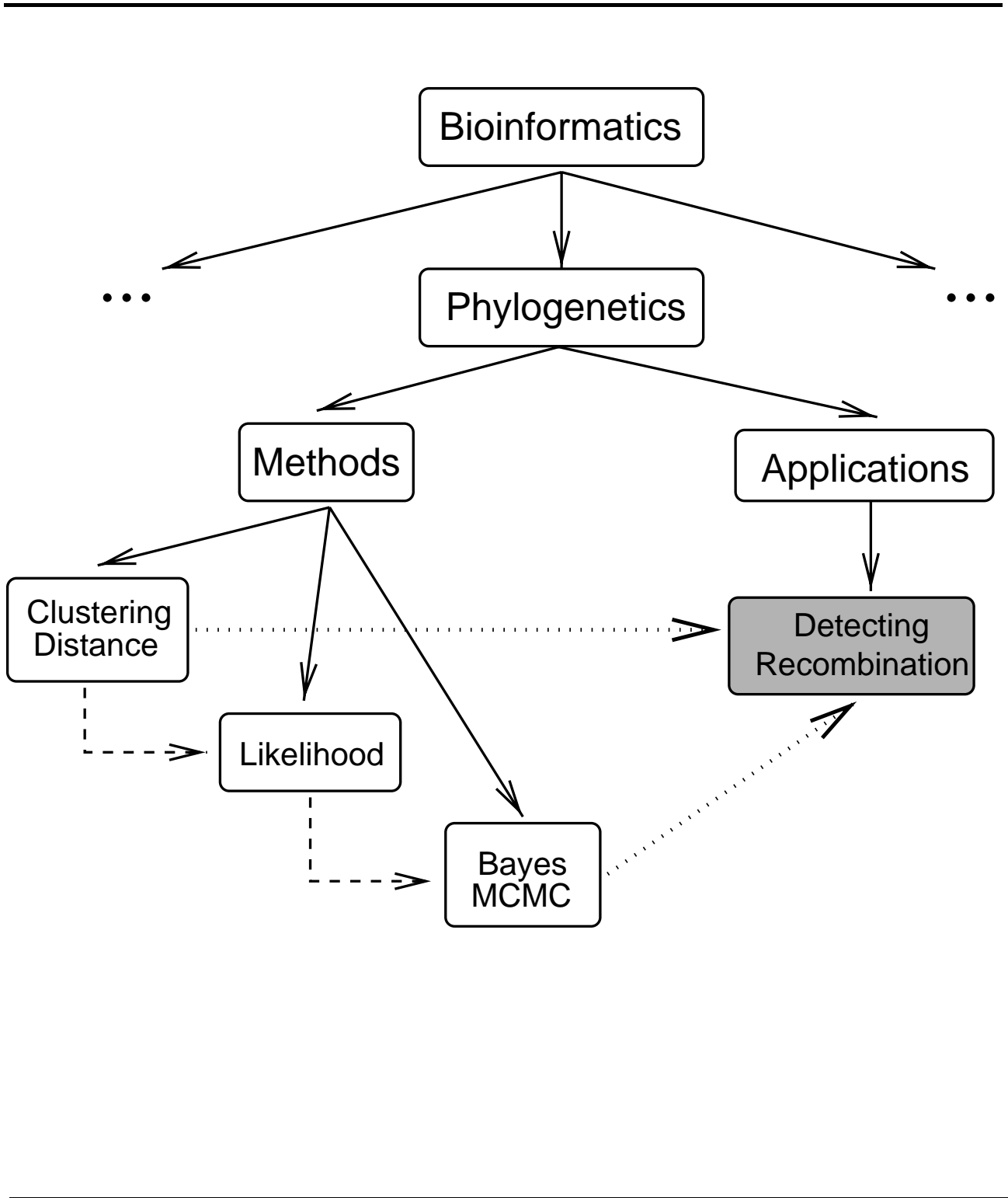
- x uniformly chosen at random from $[0, h_1]$
- y uniformly chosen at random from $[0, h_2]$
- New nodes u^* and v^* such that
 - $\text{dist}(r, u^*) = \max(x, y)$
 - $\text{dist}(r, v^*) = \min(x, y)$
- $\max(x, y) > h_1 \implies$ Leave topology unchanged
- $\max(x, y) < h_1 \implies$ Choose between three possible topologies

Current tree	$\max(x, y)$	Hastings ratio
$\text{dist}(v, u) > \text{dist}(v, c)$	$< h_1$	3
$\text{dist}(v, u) < \text{dist}(v, c)$	$> h_1$	1/3
otherwise		1

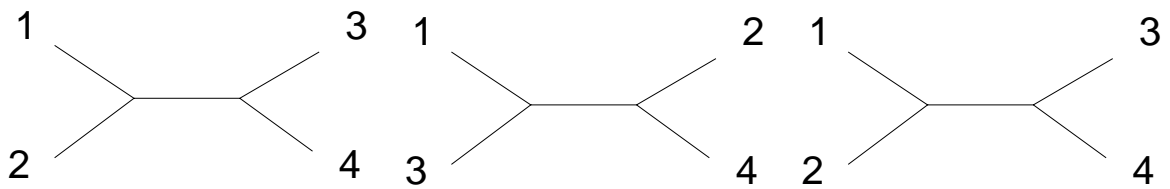
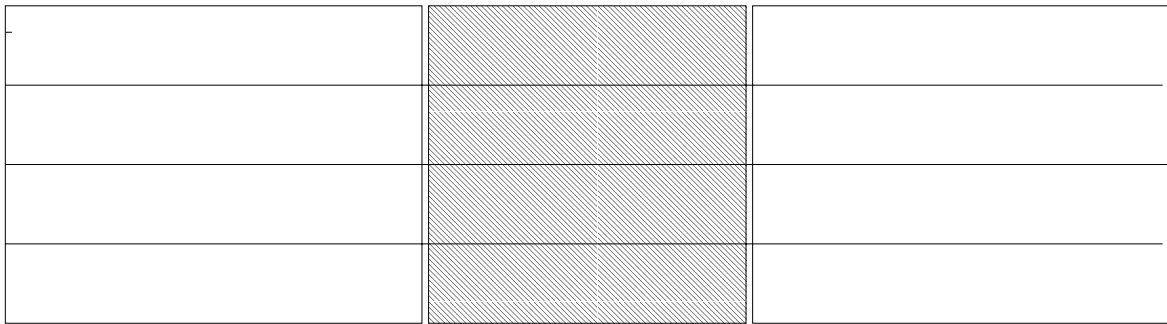
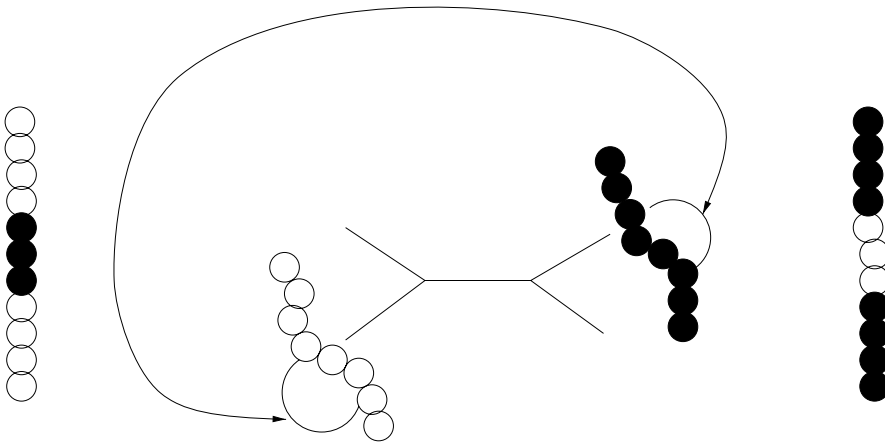
Monophyletic Groups



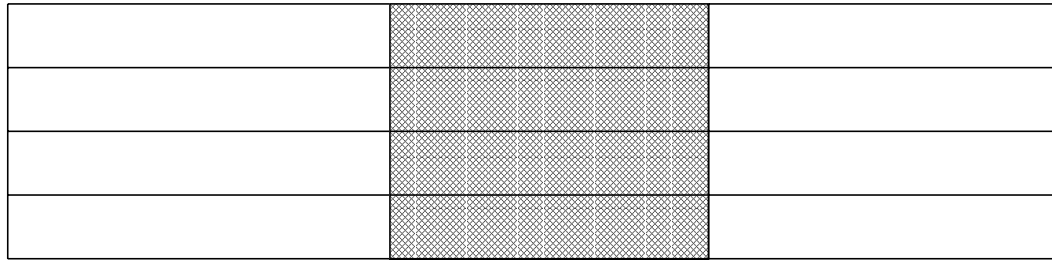
Clade	Probability
(Human Chimp)	0.974
(Human Chimp Gorilla)	1.0



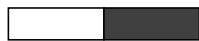
Recombination



TOPAL



DSS small



DSS large



DSS small

DSS large



DSS small

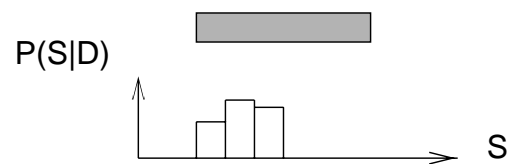
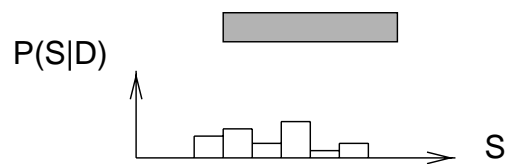
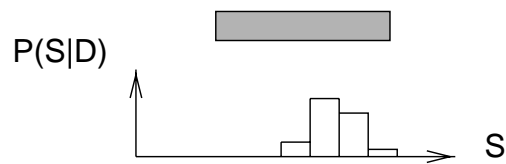
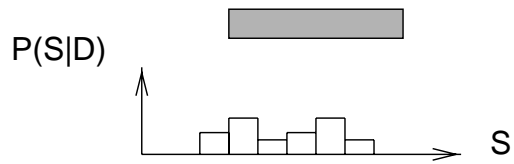
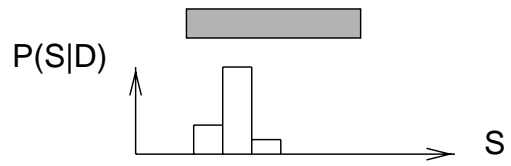
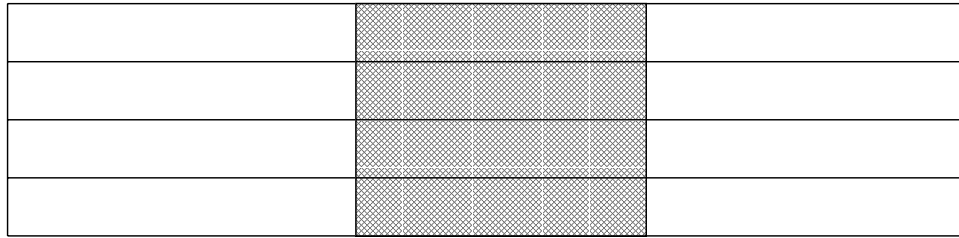


$$SoS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \quad DSS = |SoS_{left} - SoS_{right}|$$

i, k	labels for taxa
\hat{d}_{ik}	distances obtained from Neighbour Joining
d_{ik}	true distances

$$SoS = \sum_i \sum_k (d_{ik} - \hat{d}_{ik})^2 \rightarrow \sum_i \sum_k \left(\frac{d_{ik} - \hat{d}_{ik}}{d_{ik}} \right)^2$$

Detection of Recombination with MCMC



Detection of Recombination with MCMC

$$P_k(t) := P(k|\mathbf{D}_t) = \int P(k, \mathbf{w}|\mathbf{D}_t)d\mathbf{w}$$

$$P(k, \mathbf{w}|\mathbf{D}_t) \approx \frac{1}{N} \sum_{i=1}^N \delta_{k,k_{ti}} \delta(\mathbf{w} - \mathbf{w}_{ti})$$

$$P_k(t) = \frac{1}{N} \sum_{i=1}^N \delta_{k,k_{ti}} = \frac{N_k(t)}{N}$$

Entropy

$$H(t) = - \sum_k P_k(t) \ln P_k(t) \quad 0 \leq H(t) \leq \ln K$$

Divergence measure in probability space

$$KL(P, Q) = \sum_k P_k \ln \left(\frac{P_k}{Q_k} \right)$$

$$KL(Q, P) = \sum_k Q_k \ln \left(\frac{Q_k}{P_k} \right)$$

$$Q_k = \frac{1}{N} \sum_{t=1}^N P_k(t)$$

$$\tilde{Q}_k = \delta_{k,k^*}, \quad k^* = \operatorname{argmax}\{Q_k\}$$

Detection of Recombination with MCMC

$$KL(P, Q) = \sum_k P_k \ln \left(\frac{P_k}{Q_k} \right)$$

$$Q_k = \frac{1}{N} \sum_{t=1}^N P_k(t)$$

$$KL_1(t) = - \sum_k P_k(t) \ln Q_k - H(t)$$

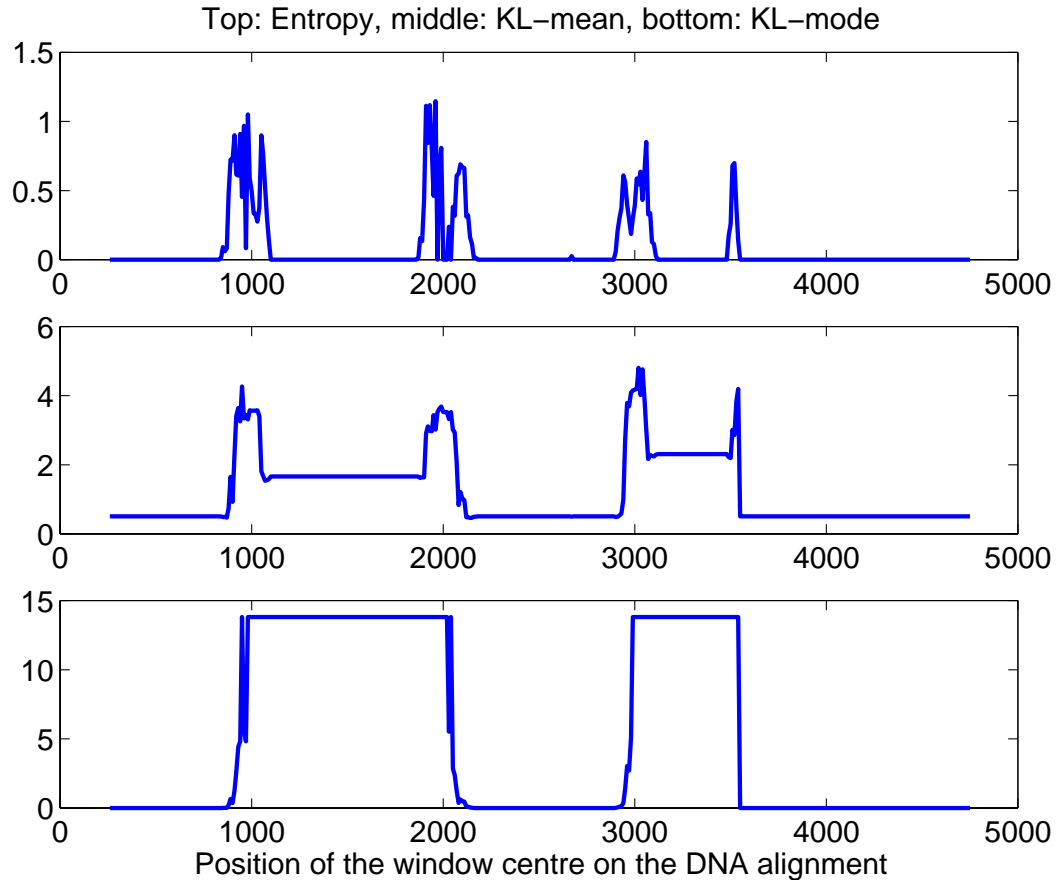
$$KL(\tilde{Q}, P) = \sum_k \tilde{Q}_k \ln \left(\frac{\tilde{Q}_k}{P_k} \right)$$

$$\tilde{Q}_k = \delta_{k, k^*}, \quad k^* = \operatorname{argmax}\{Q_k\}$$

$$KL_2(t) = - \ln P_{k^*}(t)$$

Detection of Recombination with MCMC

- Synthetic population of 8 strains
- DNA sequence alignment of 5000 bp
- Two recombinations events:
 - 1st recombination → 1000-2000 bp
 - 2nd recombination → 3000-3500 bp
- Window size = 500 bp



Summary

- Bioinformatics
- Phylogenetics
- Distance methods and clustering
- Maximum likelihood and bootstrapping
- A Bayesian approach and MCMC
- Recombination

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