

A Brief Tutorial on Phylogenetics

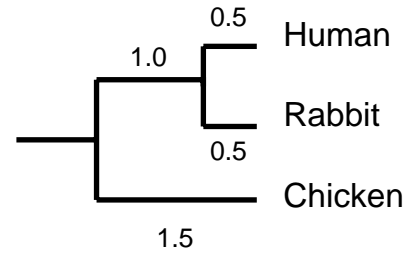
Dirk Husmeier
 Biomathematics and Statistics Scotland
 at the Scottish Crop Research Institute
 Invergowrie, Dundee DD2 5DA, UK
 Email: dirk@bioss.ac.uk
 http://www.bioss.ac.uk/~dirk

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



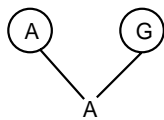
slide-1

slide-2

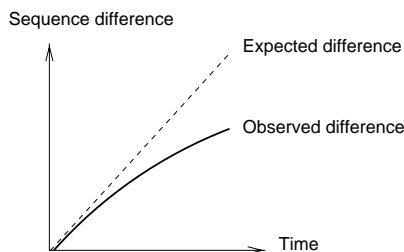
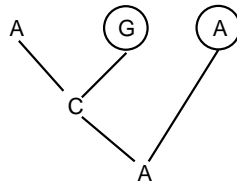
Genetic Distance

- 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



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

slide-3

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.

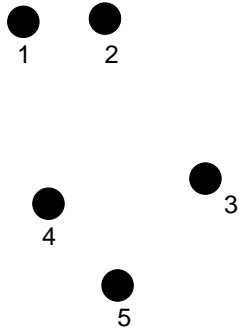
Iteration

- 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 .

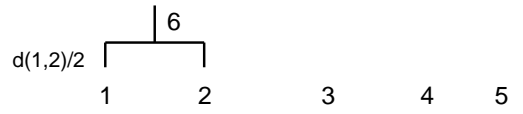
Termination

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

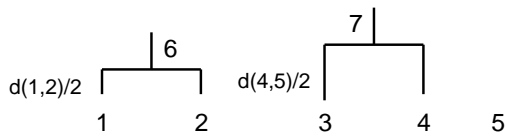
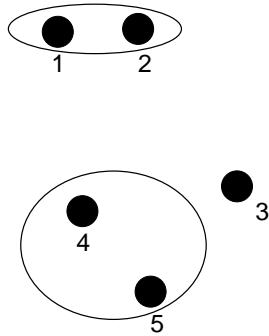
slide-4



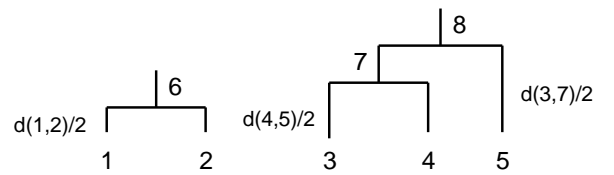
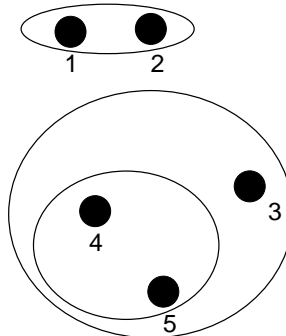
1 2 3 4 5

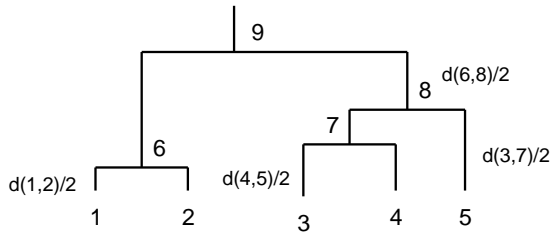
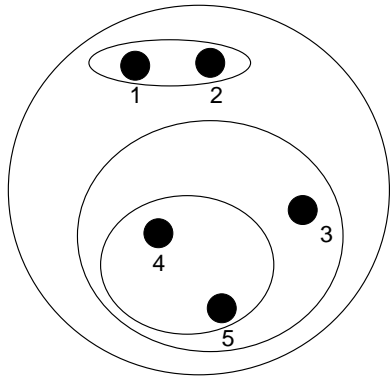


Inferring Phylogeny by Clustering: UPGMA

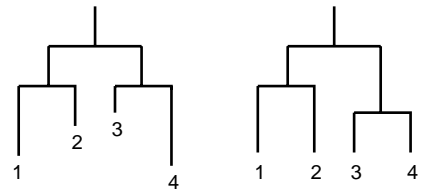


Inferring Phylogeny by Clustering: UPGMA

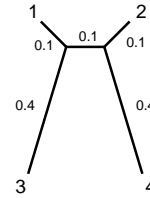




slide-9



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



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

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

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

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

slide-10

Inferring Phylogeny by Clustering: Neighbour Joining

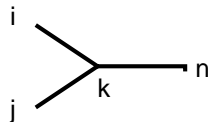
Tree metric

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

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

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

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



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

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

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

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

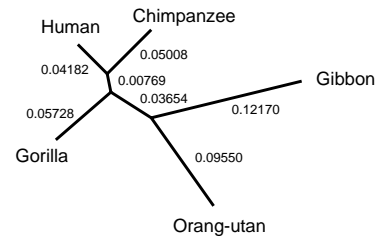
Iteration

- Find pair of node (i, j) that minimise D_{ij} .
- Replace (i, j) by new node k with new distances:
 $d_{kn} = \frac{1}{2}(d_{in} + d_{jn} - d_{ij})$

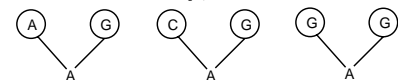
slide-11

Application of Neighbour Joining

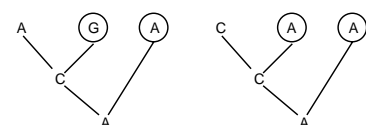
	Human	Chimp	Gorilla	Orang-utan	Gibbon
Human	-	0.0919	0.1083	0.1790	0.2057
Chimp	0.0919	-	0.1134	0.1940	0.2168
Gorilla	0.1068	0.1151	-	0.1882	0.2170
Orang-utan	0.1816	0.1898	0.1893	-	0.2172
Gibbon	0.2078	0.2160	0.2155	0.2172	-



Single substitution: 1 change, 1 difference
 Coincidental substitution: 2 changes, 1 difference
 Parallel substitution: 2 changes, no difference

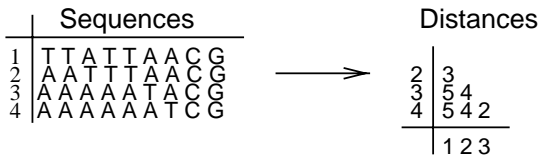


Multiple substitution: 2 changes, 1 difference
 Back substitution: 2 changes, no difference



slide-12

• Loss of Information



• Uninterpretable branch lengths

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

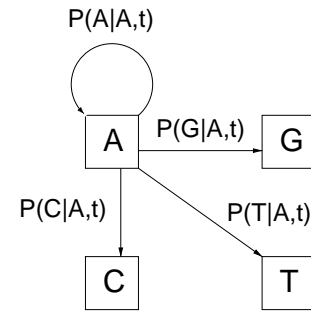
• The method does not optimize an objective function

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

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

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

Transition Rates

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

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

• Process is Markov :

$$P(y_{t+\Delta t} | y_t, y_{t-\Delta t}, \dots) = P(y_{t+\Delta t} | y_t)$$

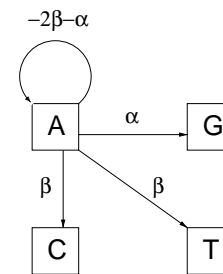
• The Markov process is homogenous :

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

• The Markov process is the same for all positions

• Substitutions at different positions are independent of each other:

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



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

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

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

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

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

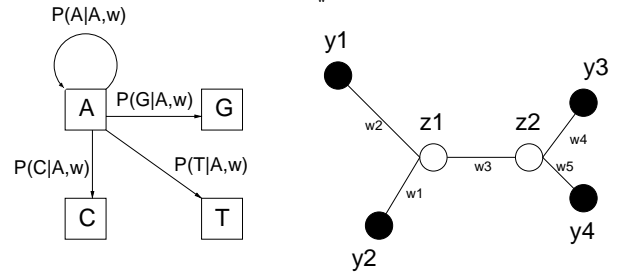
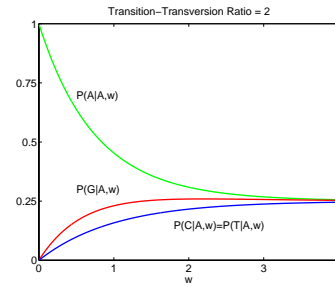
Molecular time: $w := 4\beta t$

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

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

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

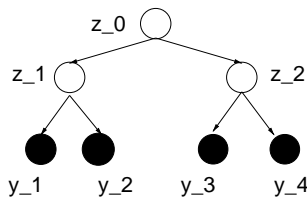
Transition-transversion ratio: $\tau = \frac{\alpha}{\beta}$



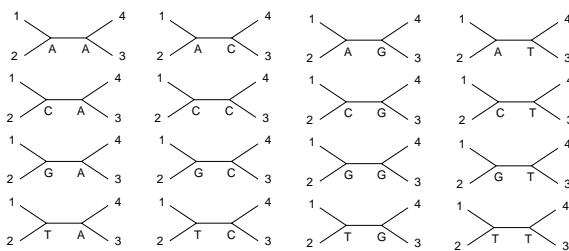
$$P(y_1, y_2, y_3, y_4, z_1, z_2 | \mathbf{w}) = P(y_1 | z_1, w_2) P(y_2 | z_1, w_1) P(z_2 | z_1, w_3) P(y_3 | z_2, w_4) P(y_4 | z_2, w_5)$$

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

Tree Likelihood: Factorisation and Marginalisation



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



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

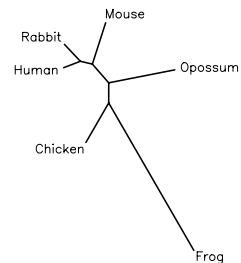
Probabilistic Approach to Phylogeny

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

$$D = (y_1, y_2, \dots, y_N)$$

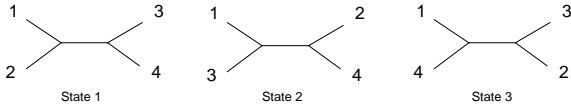
$$P(D | \mathbf{w}, S) = \prod_{t=1}^N P(y_t | \mathbf{w}, S)$$

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



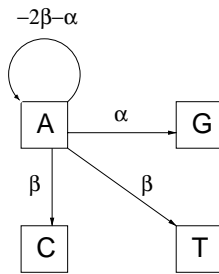
Maximise the likelihood of $L(S, \mathbf{w}, \mathbf{R}) = \ln P(D|S, \mathbf{w}, \mathbf{R})$

- Tree topology S



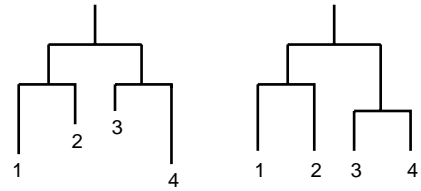
- Branch lengths $\mathbf{w} = (w_1, w_2, w_3, w_4, w_5)$

- Evolutionary parameters: Rate matrix \mathbf{R}



slide-21

Nested models

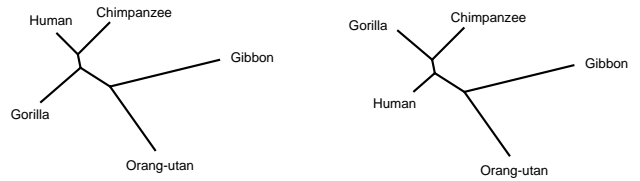


H_0 : ultrametric tree (molecular clock)

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

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

Non-nested models



slide-22

Bootstrapping



$$L_A = \ln P(D|S_A, \hat{\mathbf{w}}_A) \rightarrow \Delta L = L_A - L_B \neq 0?$$

$$L_B = \ln P(D|S_B, \hat{\mathbf{w}}_B)$$

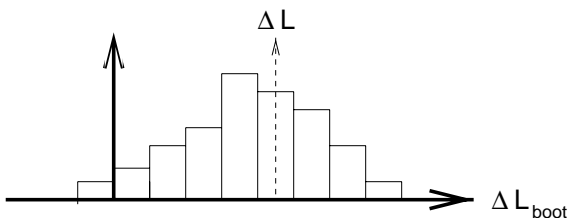
Resample with replacement from D

$$D = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\} \rightarrow D_1 = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2, \mathbf{x}_4\}$$

$$\vdots$$

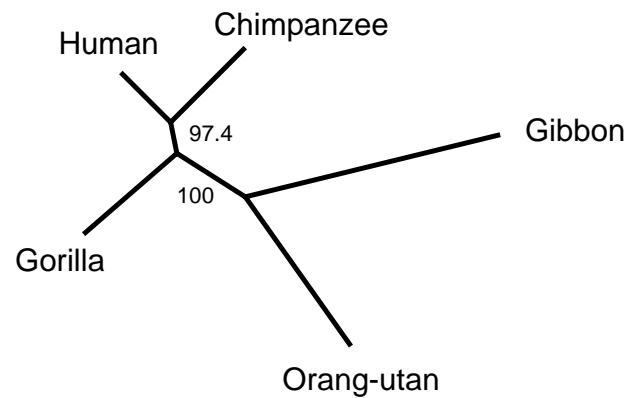
$$D_B = \{\mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_4, \mathbf{x}_1\}$$

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



slide-23

Monophyletic Groups



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

slide-24