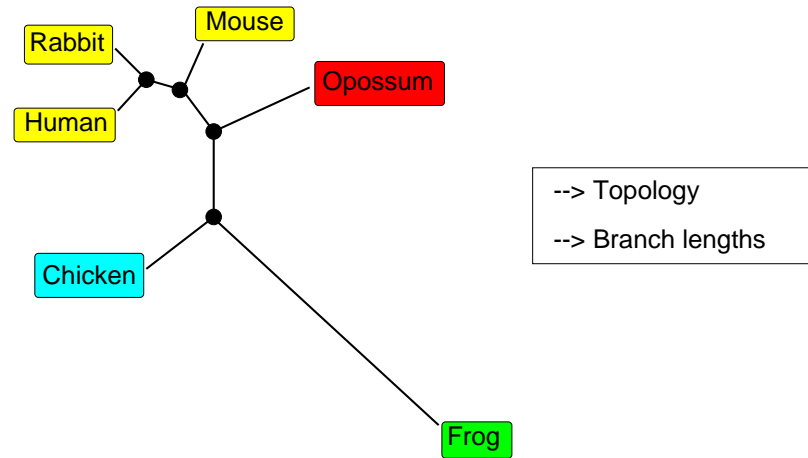


Introduction to Phylogenetics

Part 2

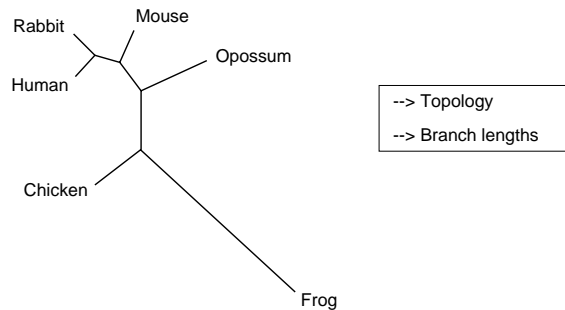
Dirk Husmeier
 Biomathematics and Statistics Scotland
 at the Scottish Crop Research Institute
 Invergowrie, Dundee DD2 5DA, UK
 Email: dirk@bioss.ac.uk
<http://www.bioss.ac.uk/~dirk>

Phylogenetic tree

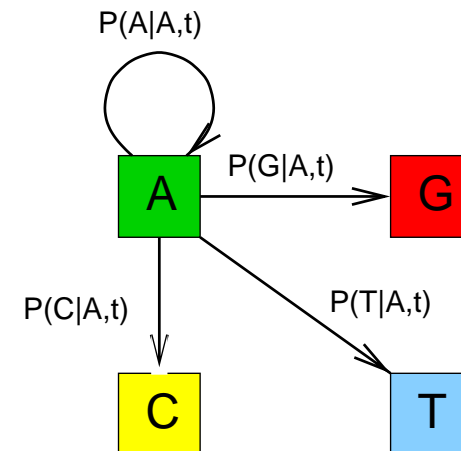


Probabilistic models of evolution

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



Mutation probabilities



Markov model of evolution

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

- 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

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

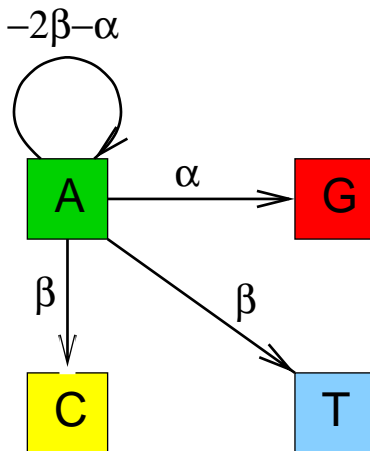
$$\mathbf{P}(dt) - \mathbf{P}(0) = \mathbf{R}dt$$

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

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

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

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: $w = 4\beta t$

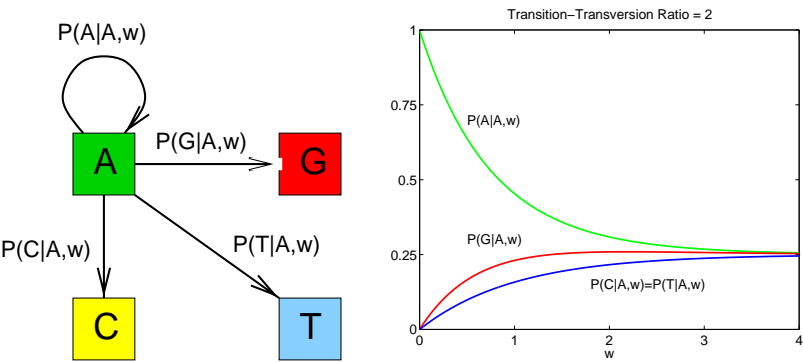
$$f(w) = \frac{1}{4}(1 - e^{-w})$$

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

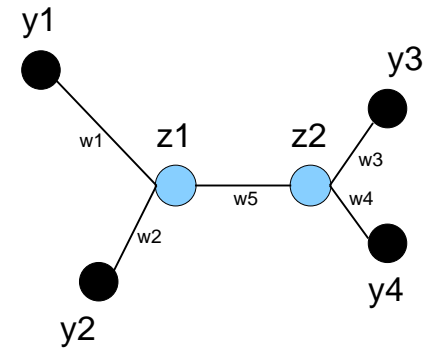
$$d(w) = 1 - 2f(w) - g(w)$$

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

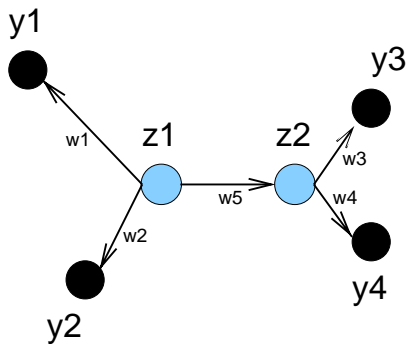
Transition probabilities



Phylogenetic tree as an undirected graph



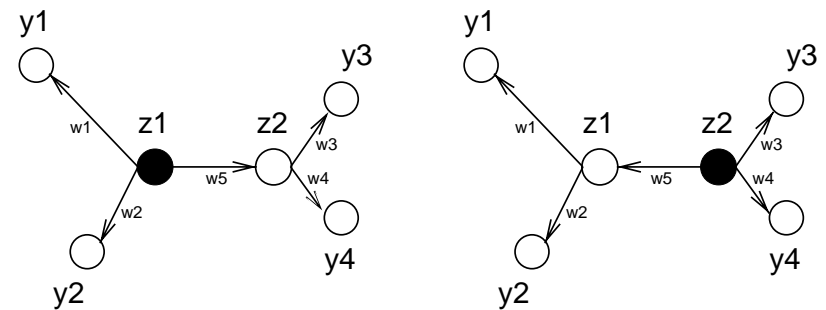
Phylogenetic tree as a directed graph



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

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

Different directed graphs



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.

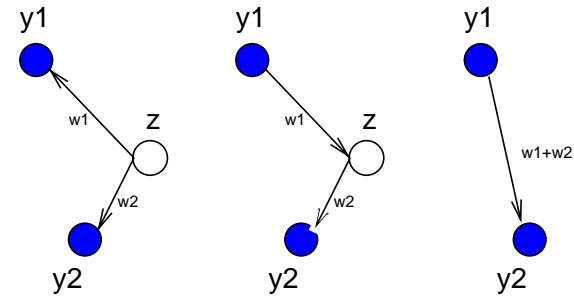
$$P(z_1|z_2, w_5)P(z_2) = P(z_2|z_1, w_5)P(z_1)$$

- 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:

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

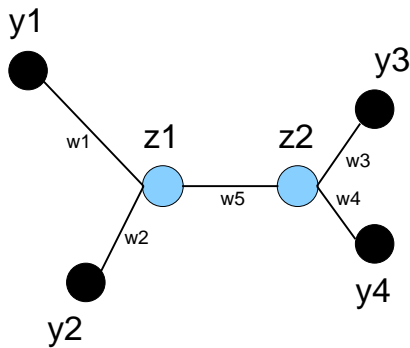


$$\sum_z P(y_2|z, w_2)P(y_1|z, w_1)P(z) = \sum_z P(y_2|z, w_2)P(z|y_1, w_1)P(y_1) = P(y_2|y_1, w_1+w_2)P(y_1)$$

Reversibility \implies left = middle

Multiplicativity \implies middle = right

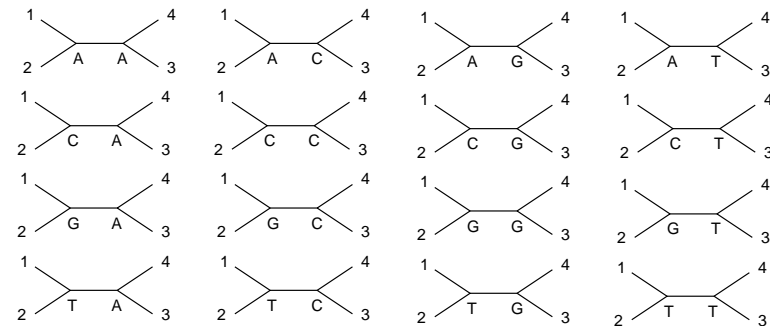
Expansion of the joint probability



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

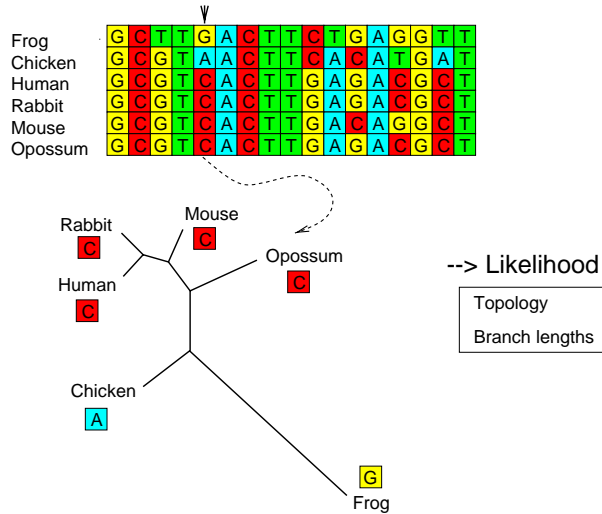
$$= P(y_1|z_1, w_1)P(y_2|z_1, w_2)P(z_2|z_1, w_5)P(y_3|z_2, w_3)P(y_4|z_2, w_4)P(z_1)$$

Marginalisation



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

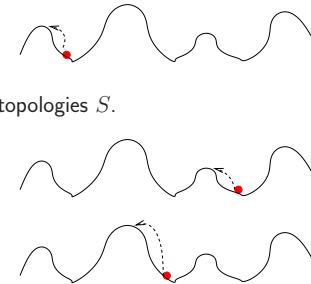
Statistical approach to phylogenetics



Maximum likelihood

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

$$\Delta w \propto \nabla \ln P(D|S, w)$$



- Repeat for different tree topologies S .

NP hard problem

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

