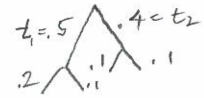
The continuous-time formulation:

parameters T^ρ, $\vec{p}_\rho$ but replace 2) with

2a) a rate matrix $Q = \begin{pmatrix} q_{AA} & \cdots & q_{AT} \\ \vdots & \ddots & \vdots \\ q_{TA} & \cdots & q_{TT} \end{pmatrix}$

with non-negative off diagonal entries and row sums 0

2b) for each edge e in the tree, a non-negative branch length t_e.



The Markov transition matrix $M_e = e^{Qt_e} = \left(e^{Q}\right)^{t_e}$

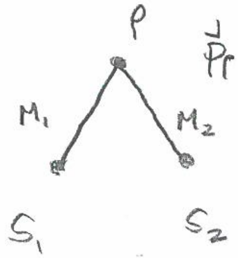
An important assumption in the continuous time formulation is a Common Rate Matrix $\underline{Q}$ for all edges of the tree.

III) Computing the expected frequency arrays.

i.e. the P(leaves show pattern $i_1 i_2 \ldots i_n$ at leaves)

→ Detour to page 10.5

Eg. 2-edge tree: Parameters $(T, \vec{p}_\rho, \{M_1, M_2\})$



Let P be the expected joint frequency array

P is 4×4 with entries

$P(i,j) = P(S_1 = i, S_2 = j)$ = expected value of seeing pattern $\begin{smallmatrix} i \\ j \end{smallmatrix}$

$S_1 : i$
$S_2 : j$

The joint distribution $P(i,j)$ for a 1-edge tree

$$\begin{array}{c} s_0 \\ \downarrow \vec{p}_0, M \\ s_1 \end{array}$$

$$= \quad P = \operatorname{diag}(\vec{p}_0)\, M$$

Since

$$\begin{pmatrix} p_A & 0 & 0 & 0 \\ 0 & p_G & 0 & 0 \\ 0 & 0 & p_C & 0 \\ 0 & 0 & 0 & p_T \end{pmatrix} \begin{pmatrix} P(S_1 = A \mid S_0 = A) & P(S_1 = G \mid S_0 = A) & P(S_1 = C \mid S_0 = A) & P(S_1 = T \mid S_0 = A) \\ P(S_1 = A \mid S_0 = G) & & \ddots & \\ P(S_1 = A \mid S_0 = C) & & & \\ P(S_1 = A \mid S_0 = T) & & & \end{pmatrix}$$

$$= \begin{pmatrix} P(S_0 = A, S_1 = A) & P(S_0 = A, S_1 = G) & P(S_0 = A, S_1 = C) & P(S_0 = A, S_1 = T) \\ P(S_0 = G, S_1 = A) & & & \\ P(S_0 = C, S_1 = A) & & \ddots & \\ P(S_0 = T, S_1 = A) & & & \\ & & & P(S_0 = T, S_1 = T) \end{pmatrix}$$

$\uparrow$

P!

Two review items from linear algebra:

Matrix Transpose $\quad M^T$

$\vec{v} \neq \vec{0}$

Eigenvalues and Eigenvectors: Suppose M is an $m \times n$ matrix, λ, $\vec{v}$ is $n \times 1$ vector, and

$$M\vec{v} = \lambda\vec{v}$$

$\uparrow$ eigenvector $\qquad \uparrow$ eigenvalue

Left eigenvectors ...

$$\vec{v}M = \lambda\vec{v}$$

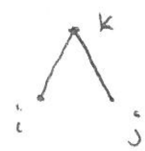
To do this, sum over all possible states at the root ρ

$\rho \quad \vec{p} = (p_A, p_G, p_C, p_T)$

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

The probability of pattern $\begin{smallmatrix} i \\ j \end{smallmatrix}$ at the leaves is

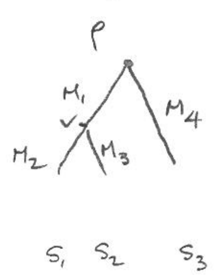
$$P(S_1 = i, S_2 = j) = P(i,j) = \sum_{k=1}^{4} \underbrace{p_k \, M_1(k,i) \, M_2(k,j)}$$



Exercise for Math Students:

Show $P = M_1^T \operatorname{diag}(\vec{p}_\rho) M_2$ for this 2-edge tree.

Eg. 3-edge tree: Parameters $\left(T, \vec{p}_\rho, \{M_1, M_2, M_3\}\right)$



must sum over states at ρ and v

$$P(A,A,G) = P(1,1,2) =$$

$$\sum_{j=1}^{4} \sum_{i=1}^{4} \vec{p}(i) \, M_1(i,j) \, M_2(j,A) \, M_3(j,A) \, M_4(i,G)$$

$\uparrow \qquad \uparrow$

$v \qquad \rho$

degree 5 polynomial with 16 summands

IV) Specific Markov models on trees used in phylogenetics

The JUKES-CANTOR model   1 parameter model

- $\vec{p}_r = (.25 \ \ .25 \ \ .25 \ \ .25)$ uniform root distribution
  all bases equally likely

- rate matrix $Q = \begin{pmatrix} -\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & -\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & -\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & -\alpha \end{pmatrix}$

  $\alpha > 0$

  $\alpha \equiv$ total rate at which a specific base is changing to any of the other 3

$\alpha/3$ = off-diagonal $\Rightarrow$ constant rate for all conversions

- for a tree branch of length t

$$M(t) = e^{Qt} = \begin{pmatrix} 1-a & a/3 & a/3 & a/3 \\ a/3 & 1-a & a/3 & a/3 \\ a/3 & a/3 & 1-a & a/3 \\ a/3 & a/3 & a/3 & 1-a \end{pmatrix}$$

where $a = a(t) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}\alpha t}\right)$

The matrix exponential is actually easy to compute since the matrix of eigenvectors is a <u>Hademard matrix</u> and the eigenvalues of Q are $0, -\frac{4}{3}\alpha, -\frac{4}{3}\alpha, -\frac{4}{3}\alpha$

Finally, note that the JC model has a STABLE BASE DISTRIBUTION

$(.25 \ \ .25 \ \ .25 \ \ .25)M = (.25 \ \ .25 \ \ .25 \ \ .25) \Rightarrow$ Should see uniform distribution of states in all sequences.

A stable base distribution $\vec{p}$ for M is a eigenvector of M with eigenvalue $\lambda = 1$.

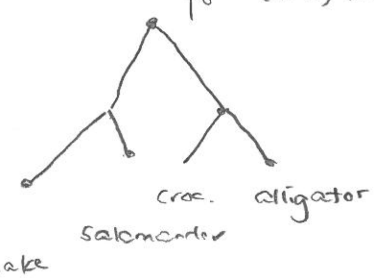
$$\vec{p}M = 1 M$$
$$\uparrow$$
$$\lambda$$

Detour: Review eigenvalues and eigenvectors possibly.

In particular, when the root distribution $\vec{p}_0 = \vec{p}$ = stable base distribution, then at <u>all</u> vertices of tree, the state distribution is $\vec{p}_v = \vec{p}$. This is called a <u>STATIONARY MODEL</u>.

Eg. Jukes-Cantor JC

$\vec{p}_0 = (.25, .25, .25, .25)$



$\vec{p}_{alligator} = (\qquad\qquad)$

$\vec{p}_{salamander} = (\qquad\qquad)$

etc.

The KIMURA MODELS $\quad \vec{p}_0 = (.25 \quad .25 \quad .25 \quad .25)$

K2P $\equiv$ Kimura 2-parameter model

A, G, C, T order

$$Q = \begin{pmatrix} * & \beta & \gamma & \gamma \\ \beta & * & \gamma & \gamma \\ \gamma & \gamma & * & \beta \\ \gamma & \gamma & \beta & * \end{pmatrix}$$

$\beta \equiv$ transition rate

$2\gamma \equiv$ transversion rate

$\uparrow$

2 types. Eg $A \to C$, $A \to T$ both equally likely

$$M_{K2P} = e^{Qt} = \begin{pmatrix} * & b & c & c \\ b & * & c & c \\ c & c & * & b \\ c & c & b & * \end{pmatrix}$$

where $b = \frac{1}{4}\left(1 - 2e^{-2(\beta+\gamma)t} + e^{-4\gamma t}\right)$

$c = \frac{1}{4}\left(1 - e^{-4\gamma t}\right)$

$* = 1 - b - 2c$ since row sums = 1.

Kimura 3 parameter model: K3ST

$\vec{p}_0 = (.25 \quad .25 \quad .25 \quad .25)$ uniform

$$Q = \begin{pmatrix} * & \beta & \gamma & \delta \\ \beta & * & \delta & \gamma \\ \gamma & \delta & * & \beta \\ \delta & \gamma & \beta & * \end{pmatrix}$$ rate matrix and

associated Markov matrices

$$M(t) = e^{Qt} = \begin{pmatrix} * & b & c & d \\ b & * & d & c \\ c & d & * & b \\ d & c & b & * \end{pmatrix}$$

Both Kimura models are stationary, i.e. the distribution of As, Gs, Cs, Ts should be uniform at <u>all</u> nodes of tree including leaves.

TIME REVERSIBLE MODELS including the

GENERAL TIME REVERSIBLE model. (GTR)

Idea: Ancestor S_0

$$S_0 \downarrow S_1$$

Model parameters $(\vec{p}_0, M_{0\to 1})$

Descendant S_1

If the direction of the edge is reversed, then a TIME REVERSIBLE model uses the <u>exact</u> <u>same</u> <u>parameters</u>.

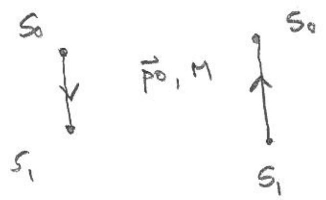
$$S_1 \downarrow S_0 \qquad \vec{p}_0, M_{0\to 1}$$

Clearly, a model can only be time reversible if it has a stable base distribution.

Mathematically, a model is time reversible if the joint distribution,

i.e. the expected pattern frequency array does not change if you

change the direction of an edge



If $P = P(i,j) = P(S_0 = i, S_1 = j)$, then

$$\underset{S_0 \to S_1}{P = \operatorname{diag}(p_0) M} = \left[\operatorname{diag}(p_0) M \right]^T = P^T$$

Defn: Process is time reversible, if

$$\operatorname{diag}(p_0) M = M^T \operatorname{diag}(p_0)$$

or in the continuous time formulation

$$\operatorname{diag}(p_0) Q = Q^T \operatorname{diag}(p_0)$$

Given a tree T^ρ, the GENERAL TIME REVERSIBLE model has numerical parameters

1) root distribution $p_\rho = (p_A, p_G, p_C, p_T)$

2) six rate parameters $\alpha, \beta, \gamma, \delta, \epsilon, \eta$

3) branch lengths t_e for each edge in T^ρ.

With these parameters, the common time reversible rate matrix is

$$Q = \begin{pmatrix} * & p_G\alpha & p_C\beta & p_T\gamma \\ p_A\alpha & * & p_C\delta & p_T\epsilon \\ p_A\beta & p_G\gamma & * & p_T\mu \\ p_A\gamma & p_G\epsilon & p_C\mu & * \end{pmatrix}$$

$*$ makes row sum $= 0$

and $M_e = e^{Q t_e}$ for each edge.

GTR: 1) has $\vec{p}_0$ eigenvector $\equiv$ stable base distribution

2) most commonly used model in phylogenetic analyses

Other models

F81 $\equiv$ arbitrary $\vec{p}_0$ $\alpha = \beta = \gamma = \epsilon = \delta = \mu = 1$ "JC without uniform root distribution"

HKY $\equiv$ arbitrary $\vec{p}_0$ $\beta = \gamma = \delta = \epsilon = 1$ $\alpha = \mu = K$

"K2ST' without uniform root distribution"