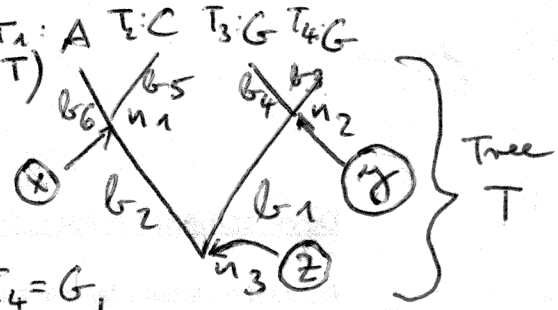
$$P(n_i \mid T) = P(T_1 = A, T_2 = C, T_3 = G, T_4 = G \mid T)$$

n_i: site i

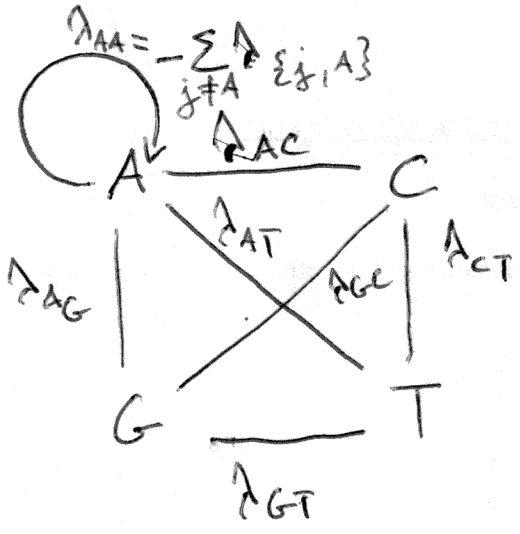


$$= \sum_{x \in X} \sum_{y \in X} \sum_{z \in X} P(T_1 = A, T_2 = C, T_3 = G, T_4 = G, n_3 = z, n_2 = y, n_3 = x \mid T) \qquad x \in \{A, C, G, T\}$$

$$= \sum_{x \in X} \sum_{y \in X} \sum_{z \in X} P(n_3 = z)\, P(z \to x)\, P(z \to y)\, P(x \to A)\, P(x \to C)\, P(y \to G)\, P(y \to x)$$

↑ factor in the b_i into all $P(\to)$

How can we get $P(x \to C \mid b_5)$



	A	C	G	T
A	*			
C	λ_{AC}	*		
G	λ_{AG}	λ_{CG}	*	
T	λ_{AT}	λ_{CT}	λ_{GT}	*

SYMMETRIC

rate matrix R

equilibrium frequencies

$$\vec{\pi} = (\pi_A, \pi_C, \pi_G, \pi_T) \qquad \sum_{x \in X} \pi_x \overset{!}{=} 1$$

$-\sum \lambda_{ii} \overset{!}{=} 1 \Rightarrow$ only 5 free parameters in the subst. matrix

$\vec{\pi}$ yields 3 additional free parameters

(2)

I JC: Jukes - Cantor model

$$JC: \begin{pmatrix} * & & & \\ a & * & & \\ a & a & * & \\ a & a & a & * \end{pmatrix} \quad \vec{\pi} = (0.25, 0.25, 0.25, 0.25)$$

$\pi_i = \pi_j \ \forall \ i, j$

II Felsenstein 81 model: $\pi_i \neq \pi_j$

III Kimura 2 parameter model

$\pi_i = \pi_j$

$$\begin{pmatrix} * & & & \\ \beta & * & & \\ \alpha & \beta & * & \\ \beta & \alpha & \beta & * \end{pmatrix}$$

IV HKY85 $\pi_i \neq \pi_j$ and Kimura 2 parameter model matrix

V GTR: General Time Reversible model

$$\begin{pmatrix} * & & & \\ \alpha & * & & \\ \beta & \gamma & * & \\ \delta & \varepsilon & \zeta & * \end{pmatrix}$$

time reversibility requires entries of a $\underline{Q}$ matrix

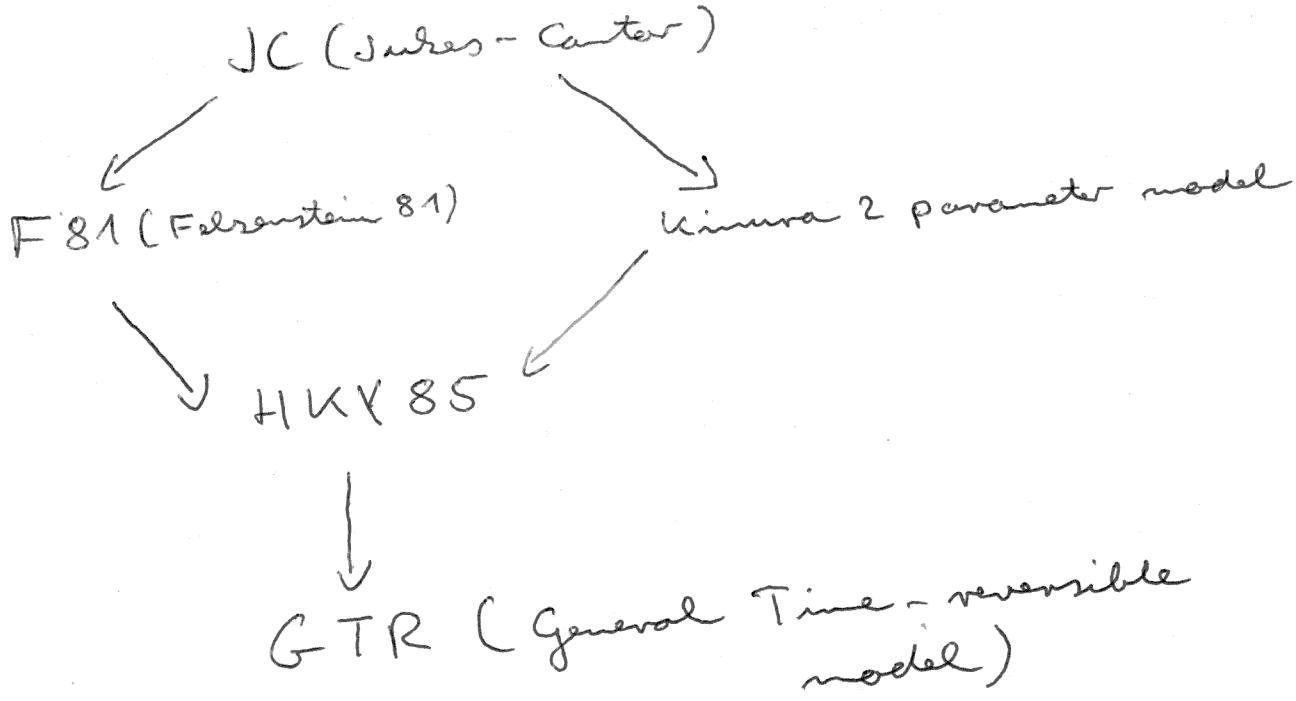
$$\pi_i \cdot q_{ij} = \pi_j \, q_{ji}$$

Q matrice: $Q = \text{diag}(\vec{\pi}) \cdot R$

$$\begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_G & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix}$$

2a

Model hierarchy and nested models



(3)

Q matrice of GTR model

$$Q = \begin{pmatrix} * & \pi_C \lambda_{AC} & \cdots \\ \pi_A \lambda_{AC} & * & \cdots \\ \pi_A \lambda_{AG} & \cdots & * \\ \pi_A \lambda_{AT} & \cdots & * \end{pmatrix}$$

$\pi_A(\pi_C \lambda_{AC}) = \pi_C(\pi_A \lambda_{AC})$

time reversibility holds

Likelihood ratio test

$$LHR \quad D = -2 \ln\left(\frac{LH(M_1)}{LH(M_2)}\right)$$

Model 1

Model 2

difference in # params. of $M_1 M_2$

compare to X^2 distribution

Only use this for nested models

AKAIKE INFORMATION CRITERION

$$AIC_i = -2 \ln(LH(M_i)) + 2 \cdot (p_i)$$

number of free parameters in model M_i

Models do not need to be nested

Model i

$$BIC_i = -2 \ln(LH(M_i)) + p_i \cdot \ln(n)$$

size of the data

sites or

sites * # taxa

Back to Likelihood calculations

$$LH = \sum\sum\sum p(x)\, P(x \to y \mid b_1, M)$$

M: model (Q-matrise)

b_1: branch length

JC model $\lambda_{AC} = \lambda_{AT} = \ldots = \lambda_{GT} = \alpha \cdot \frac{1}{3}$

λ_{AA}

$\alpha \longrightarrow \frac{4}{3}\alpha$

$\frac{1}{3}\alpha \longrightarrow C$
$\longrightarrow G$
$\longrightarrow T$

$\frac{1}{3}\alpha \to C$
$\to G$
$\to T$
$\to A$

Poisson distribution for calculating $P(t)$ from Q

$$\text{k events in time t} = \frac{(at)^k}{k!} e^{-\frac{4}{3}at}$$

$k := 0 \to$ no event

$$e^{-\frac{4}{3}at} = P(\text{no event})$$

$$P(\text{none event}) = 1 - e^{-\frac{4}{3}at}$$

for large t $P(\text{none event}) = \frac{1}{4}$

$$P(A \to C \mid t) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}at}\right)$$

Calculating $P(t)$ for a Q matrise of a GTR model → see slides

in general we compute: $P(t) = e^{Qt}$

(5)

$P_{ij}(t) = e^{Qt}$ $\qquad Q = U \Lambda U^{-1}$

↑ remember: $\text{diag}(\vec{\pi}) \cdot R$

TRICK me Q'

$Q' := \text{diag}\left(\sqrt{\pi}\right)^{-1} Q \, \text{diag}\left(\sqrt{\pi}\right)$

$Q' = U' \Lambda U'^{T}$

↑ diagonal of Eigenvalues

$\Leftrightarrow \text{diag}\left(\sqrt{\pi}\right)^{-1} Q \, \text{diag}\left(\sqrt{\pi}\right) = U' \Lambda U'^{T}$

$\Leftrightarrow Q = \underbrace{\left(\text{diag}\left(\sqrt{\pi}\right) U'\right)}_{= U} \Lambda \underbrace{\left(U'^{T} \text{diag}\left(\sqrt{\pi}\right)^{-1}\right)}_{= U^{-1}}$

$\Leftrightarrow Q = U \Lambda U^{-1}$

$$P(t) = e^{Qt} = \sum_{k=0}^{\infty} \frac{1}{k!} (Qt)^k = \sum_{k=0}^{\infty} \frac{1}{k!} \underbrace{(U \Lambda U^{-1})^k}_{U \Lambda \underbrace{U^{-1} U}_{\text{Identity Matrice}} \Lambda \underbrace{U^{-1} U}_{\text{Identity Matrice}} \Lambda} \cdot t^k$$

$$= U \Lambda^k U^{-1}$$

$$= U \left(\sum_{k=0}^{\infty} \frac{1}{k!} \Lambda^k t^k \right) U^{-1}$$

remember: $\Lambda = \text{diag}(\lambda_i)$

hence $P(t) = U e^{\text{diag}(\lambda_i) \cdot t} U^{-1}$

Protein substitution matrices

Dayhoff et. al.: PAM matrices

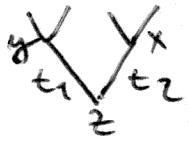
counting approach

Use only closely related sequences such that $A \to T \to C$ does not happen!!

$$\begin{pmatrix} * & 3 & 1 & 0 & \cdots \\ \cdot & * & 2 & 2 & \cdots \\ \cdot & \cdot & * & 2 & \cdots \\ \cdot & \cdot & \cdot & * & \cdots \end{pmatrix}$$ ← each substitution is counted as one change

---

MAXIMUM LIKELIHOOD BRANCH LENGTH OPTIMIZATION



$$\sum_z \sum_y \sum_x \pi_z \, P(z \to y | t_1) \cdot P(z \to x | t_2) \cdot \underbrace{LH_1(y) \cdot LH_2(x)}$$

$$= \sum_x \sum_y \pi_y \, P(y \to x | t_1 + t_2) \cdot C(x,y)^{C(x,y)}$$

remains constant

$\Rightarrow$ we can move z around the branch connecting x and y and will get the same result

Newton-Raphson procedure, second order Taylor-Approximation

$$\tilde{P}(t) = P(a) + P'(a)(t-a) + \frac{1}{2} P''(a)(t-a)^2 + \varepsilon$$

$$= \underbrace{(\dots)}_{c_0} + \underbrace{\left(P'(a) - \frac{2}{2} P''(a)\, a\right) t}_{c_1} + \underbrace{\left(\frac{1}{2} P''(a)\, t^2\right)}_{c_2}$$

$$= c_0 + c_1 \cdot t + c_2 \cdot t^2$$

minimization ↙

$$P' = c_1 + 2c_2 \cdot t \stackrel{!}{=} 0$$

$$\Leftrightarrow \quad t = -\frac{c_1}{2c_2} = -\frac{P'(a) - P''(a) \cdot a}{2 \cdot \frac{1}{2} P''(a)} = a - \frac{P'(a)}{P''(a)}$$

→ NEXT

(7) Derivatives of the phylogenetic likelihood function:

$$P(t) = \sum_x \sum_y C_{(x,y)} \cdot P(y \to x \mid t)$$

$$= \sum_x \sum_y C_{(x,y)} \cdot U \, e^{\lambda_{xy} t} \cdot U^{-1}$$

$$\Rightarrow$$

$$P'(t) = \sum_x \sum_y C_{(x,y)} \, U \, \lambda_{xy} \, e^{\lambda_{xy} t} \, U^{-1}$$

$$P''(t) = \sum_x \sum_y C_{(x,y)} \, U \, \lambda_{yx}^2 \cdot e^{\lambda_{xy} t} \cdot U^{-1}$$