

# Q and P(t)

What is the probability of going from i (C?) to j (G?) in time t with rate matrix Q?

$$P(t) = \exp(tQ) = \sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} = I + tQ + \frac{(tQ)^2}{2!} + \frac{(tQ)^3}{3!} + \dots$$

- i.**  $P(0) = I$
- ii.**  $P(\varepsilon)$  close to  $I + \varepsilon Q$  for  $\varepsilon$  small
- iii.**  $P'(0) = Q$ .
- iv.**  $\lim_{t \rightarrow \infty} P(t)$  has the equilibrium frequencies of the 4 nucleotides in each row
- v.** Waiting time in state j,  $T_j$ ,  $P(T_j > t) = e^{-q_{jj}t}$
- vi.**  $QE=0$   $E_{ij}=1$  (all i,j)
- vii.**  $PE=E$
- viii.** If  $AB=BA$ , then  $e^{A+B}=e^A e^B$ .

Expected number of events at equilibrium

$$t \sum_{\text{nucleotides}} -q_{ii} \pi_i$$

# Jukes-Cantor (JC69): Total Symmetry

Rate-matrix, R:

T O

		A	C	G	T
FROM	A	$-3*\alpha$	$\alpha$	$\alpha$	$\alpha$
	C	$\alpha$	$-3*\alpha$	$\alpha$	$\alpha$
	G	$\alpha$	$\alpha$	$-3*\alpha$	$\alpha$
	T	$\alpha$	$\alpha$	$\alpha$	$-3*\alpha$

Transition prob. after time t,  $a = \alpha*t$ :

$$P(\text{equal}) = \frac{1}{4}(1 + 3e^{-4a}) \sim 1 - 3a$$

$$P(\text{diff.}) = \frac{1}{4}(1 - 3e^{-4a}) \sim 3a$$

Stationary Distribution: (1,1,1,1)/4.

$$\begin{aligned}
 P &= P(s1) \prod_{i=1}^5 P(s1_i \rightarrow s2_i) = \left(\frac{1}{4}\right)^5 P(T \rightarrow T)P(C \rightarrow G)P(G \rightarrow G)P(G \rightarrow T)P(A \rightarrow T) \\
 &= \left(\frac{1}{4}\right)^5 \left(\frac{1}{4}\right)^5 (1 + 3e^{-4a})^2 (1 - e^{-4a})^3
 \end{aligned}$$

# Principle of Inference: Likelihood

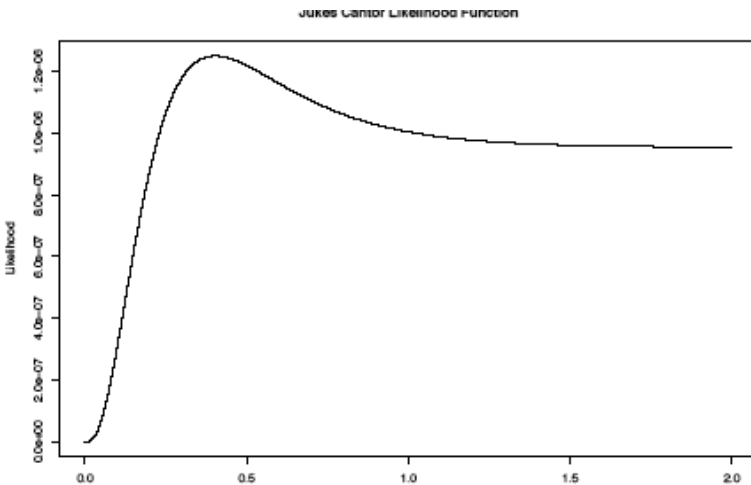
Likelihood function  $L()$  – the probability of data as function of parameters:  $L(\Theta, D)$

LogLikelihood Function –  $l()$ :  $\ln(L(\Theta, D))$

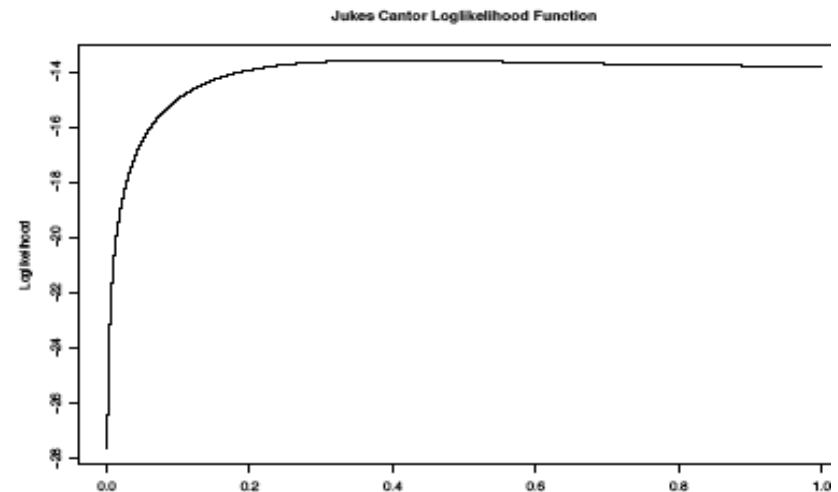
If the data is a series of independent experiments  $L()$  will become a product of Likelihoods of each experiment,  $l()$  will become the sum of LogLikelihoods of each experiment

Consistency :  $\hat{\Theta}(D) \rightarrow \Theta_{true}$  as data increases.

## Likelihood



## LogLikelihood



In Likelihood analysis parameter is not viewed as a random variable.

# From Q to P for Jukes-Cantor

$$\begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix} = \alpha \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix}^i = 4^{i-1} \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix}$$

$$\sum_{i=0}^{\infty} \begin{bmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{bmatrix}^i \frac{t^i}{i!} = 1/4 \left[ I - \sum_{i=1}^{\infty} (-4\alpha t)^i \begin{bmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{bmatrix} \frac{1}{i!} \right] =$$

$$1/4 \left[ I + \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix} e^{-4\alpha t} \right]$$

# Exponentiation/Powering of Matrices

By eigen values:

If  $Q = B\Lambda B^{-1}$  where  $\Lambda = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}$  then  $Q^i = B\Lambda B^{-1}B\Lambda B^{-1} \dots B\Lambda B^{-1} = B\Lambda^i B^{-1}$

and  $\sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} = \sum_{i=0}^{\infty} \frac{(tB\Lambda B^{-1})^i}{i!} = B \left[ \sum_{i=0}^{\infty} \frac{(t\Lambda)^i}{i!} \right] B^{-1} = B \begin{pmatrix} \exp t\lambda_1 & 0 & 0 & 0 \\ 0 & \exp t\lambda_2 & 0 & 0 \\ 0 & 0 & \exp t\lambda_3 & 0 \\ 0 & 0 & 0 & \exp t\lambda_4 \end{pmatrix} B^{-1}$

Finding  $\Lambda$ :  $\det(Q - \lambda I) = 0$

Finding  $B$ :  $(Q - \lambda_i I)b_i = 0$

**JC69:**

$$P(t) = \begin{pmatrix} 1 & 1/4 & 0 & 1 \\ 1 & 1/4 & 0 & -1 \\ 1 & -1/4 & 1 & 0 \\ 1 & -1/4 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \exp -4t\alpha & 0 & 0 \\ 0 & 0 & \exp -4t\alpha & 0 \\ 0 & 0 & 0 & \exp -4t\alpha \end{pmatrix} \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/8 & 1/8 & -1/8 & -1/8 \\ 0 & 0 & 1 & -1 \\ 1 & -1 & 0 & 0 \end{pmatrix}$$

Numerically:

$$\sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} \sim \sum_{i=0}^k \frac{(tQ)^i}{i!} \quad \text{where } k \sim 6-10$$

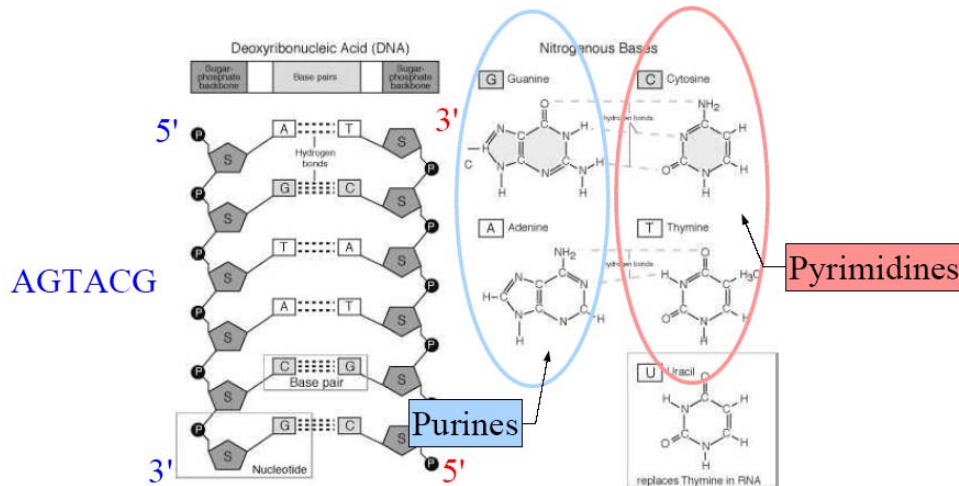
# Kimura 2-parameter model - K80

TO

	A	C	G	T
F A	$-2*\beta-\alpha$	$\beta$	$\alpha$	$\beta$
R C	$\beta$	$-2*\beta-\alpha$	$\beta$	$\alpha$
O G	$\alpha$	$\beta$	$-2*\beta-\alpha$	$\beta$
M T	$\beta$	$\alpha$	$\beta$	$-2*\beta-\alpha$

$a = \alpha * t$        $b = \beta * t$

*P(t)*



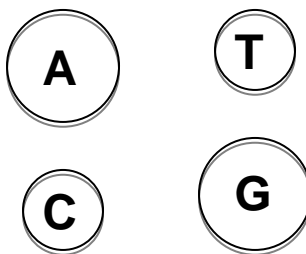
start	$.25(1 - e^{-4b})$
$.25(1 + e^{-4b} + 2e^{-2(a+b)})$	$.25(1 - e^{-4b})$
$.25(1 + e^{-4b} - 2e^{-2(a+b)})$	$.25(1 - e^{-4b})$

# Felsenstein81 & Hasegawa, Kishino & Yano 85

Unequal base composition: (Felsenstein, 1981 F81)

$$Q_{i,j} = C^* \pi_j \quad i \text{ unequal } j$$

Rates to frequent nucleotides are high - ( $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ )

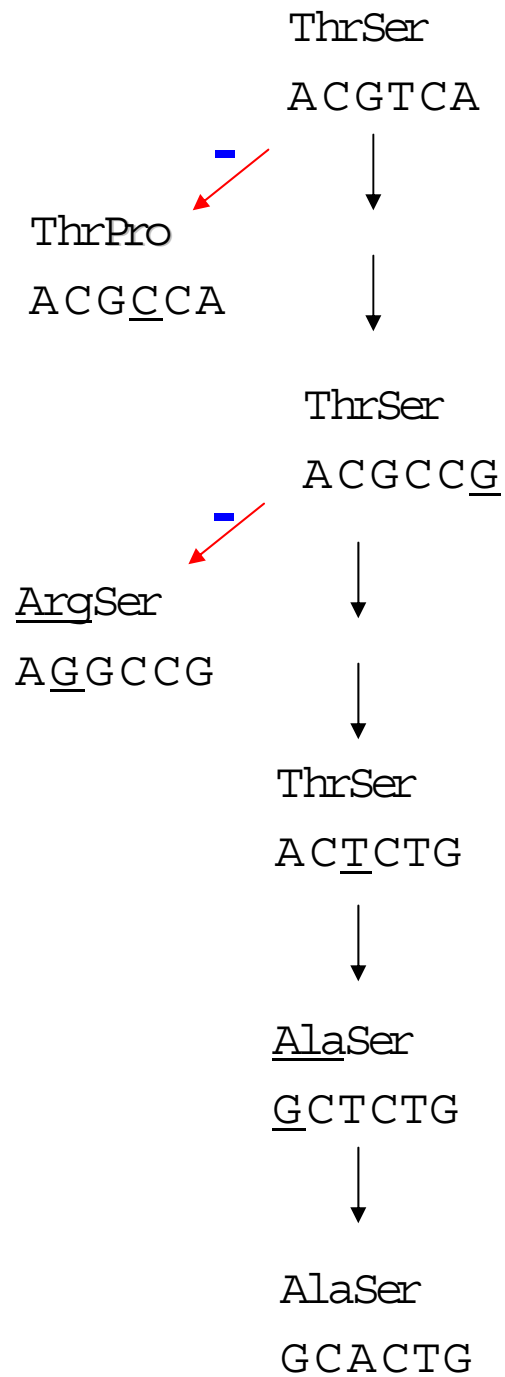
$$Tv/Tr = (\pi_T \pi_C + \pi_A \pi_G) / [(\pi_T + \pi_C)(\pi_A + \pi_G)]$$


Tv/Tr & composition bias (Hasegawa, Kishino & Yano, 1985 HKY85)

$$Q_{i,j} = \begin{cases} (\alpha/\beta)^* C^* \pi_j & i \rightarrow j \text{ a transition} \\ C^* \pi_j & i \rightarrow j \text{ a transversion} \end{cases}$$

$$Tv/Tr = (\alpha/\beta) (\pi_T \pi_C + \pi_A \pi_G) / [(\pi_T + \pi_C)(\pi_A + \pi_G)]$$

# Measuring Selection



Certain events have functional consequences and will be selected out. The strength and localization of this selection is of great interest.

The selection criteria could in principle be anything, but the selection against amino acid changes is without comparison the most important

# The Genetic Code

3 classes of sites:

4

2-2

1-1-1-1

i. 

TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA	!!!	TGA	!!!
TTG	Leu	TCG	Ser	TAG	!!!	TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCT	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

4 (3<sup>rd</sup>)

1-1-1-1 (3<sup>rd</sup>)

ii. T→A (2<sup>nd</sup>)

## Problems:

i. Not all fit into those categories.

ii. Change in one site can change the status of another.

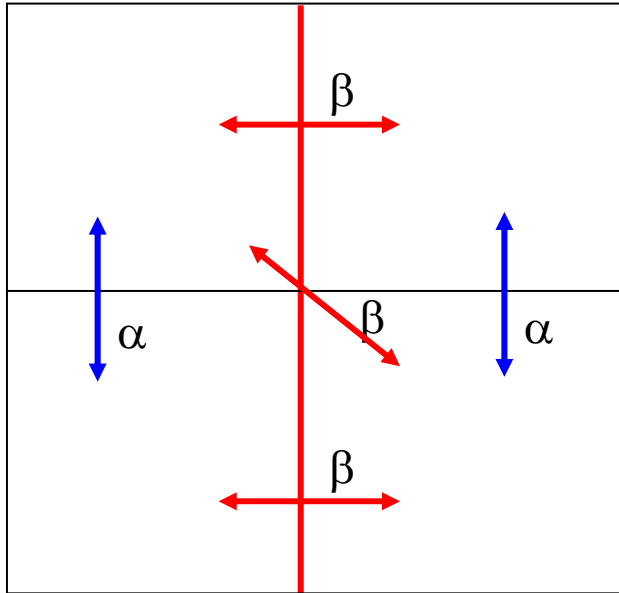
# Possible events if the genetic code remade from Li,1997

Possible number of substitutions:  $61 \text{ (codons)} \times 3 \text{ (positions)} \times 3 \text{ (alternative nucleotides)}$ .

Substitutions	Number	Percent
Total in all codons	549	100
Synonymous	134	25
Nonsynonymous	415	75
Missense	392	71
Nonsense	23	4

# Kimura's 2 parameter model & Li's Model.

Rates:



Probabilities:

start	
$.25(1 + e^{-4b} + 2e^{-2(a+b)})$	$.25(1 - e^{-4b})$
$.25(1 + e^{-4b} - 2e^{-2(a+b)})$	$.25(1 - e^{-4b})$

Selection on the 3 kinds of sites  $(a,b) \rightarrow (?,?)$

1-1-1-1  $(f^* \alpha, f^* \beta)$

2-2  $(\alpha, f^* \beta)$

4  $(\alpha, \beta)$

# alpha-globin from rabbit and mouse.

Ser Thr Glu Met Cys Leu Met Gly Gly  
 TCA ACT GAG ATG TGT TTA ATG GGG GGA  
 \* \* \* \* \* \* \* \*  
 TCG ACA GGG ATA TAT CTA ATG GGT ATA  
 Ser Thr Gly Ile Tyr Leu Met Gly Ile

Sites	Total	Conserved	Transitions	Transversions
1-1-1-1	274	246 (.8978)	12(.0438)	16(.0584)
2-2	77	51 (.6623)	21 (.2727)	5(.0649)
4	78	47 (.6026)	16 (.2051)	15(.1923)

$$Z(\alpha, \beta t) = .50[1 + \exp(-2\alpha t) - 2\exp(-t(\alpha + \beta))] \quad \text{transition}$$

$$Y(\alpha, \beta t) = .25[1 - \exp(-2\beta t)] \quad \text{transversion}$$

$$X(\alpha, \beta t) = .25[1 + \exp(-2\alpha t) + 2\exp(-t(\alpha + \beta))] \quad \text{identity}$$

L(observations, a, b, f) =

$$C(429, 274, 77, 78) * \{X(a*f, b*f)^{246} * Y(a*f, b*f)^{12} * Z(a*f, b*f)^{16}\} * \{X(a, b*f)^{51} * Y(a, b*f)^{21} * Z(a, b*f)^{5}\} * \{X(a, b)^{47} * Y(a, b)^{16} * Z(a, b)^{15}\}$$

where a = at and b = bt.

Estimated Parameters: a = 0.3003 b = 0.1871 2\*b = 0.3742 (a + 2\*b) = 0.6745 f = 0.1663

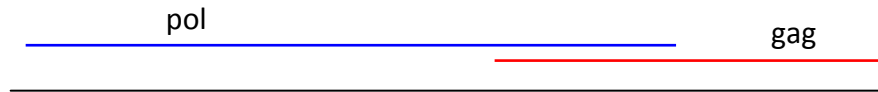
	Transitions	Transversions
1-1-1-1	a*f = 0.0500	2*b*f = 0.0622
2-2	a = 0.3004	2*b*f = 0.0622
4	a = 0.3004	2*b = 0.3741

Expected number of: replacement substitutions 35.49      synonymous 75.93  
 Replacement sites : 246 + (0.3742/0.6744)\*77 = 314.72  
 Silent sites : 429 - 314.72 = 114.28       $K_s = .6644$   $K_a = .1127$

# Extension to Overlapping Regions

Hein & Stoevlbaek, 95

<b>2<sup>nd</sup></b> \ <b>1<sup>st</sup></b>	1-1-1-1	2-2	4
1-1-1-1 sites	$(f_1 f_2 a, f_1 f_2 b)$	$(f_2 a, f_1 f_2 b)$	$(f_2 a, f_2 b)$
2-2	$(f_1 a, f_1 f_2 b)$	$(f_2 a, f_1 f_2 b)$	$(a, f_2 b)$
4	$(f_1 a, f_1 b)$	$(a, f_1 b)$	$(a, b)$

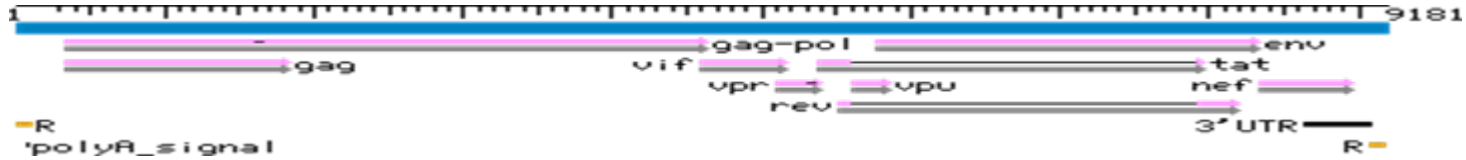


## Example: Gag & Pol from HIV

<b>Pol</b> \ <b>Gag</b>	1-1-1-1	2-2	4
1-1-1-1 sites	64	31	34
2-2	40	7	0
4	27	2	0

**MLE:**  $a = .084$   $b = .024$   $a + 2b = .133$   $f_{gag} = .403$   $f_{pol} = .229$

# HIV1 Analysis



## Hasegawa, Kisino & Yano Substitution Model Parameters:

$\alpha^*$	$\beta^*$	$\pi_A$	$\pi_C$	$\pi_G$	$\pi_T$
0.350	0.105	0.361	0.181	0.236	0.222
0.015	0.005	0.004	0.003	0.003	

## Selection Factors

GAG	0.385	(s.d. 0.030)
POL	0.220	(s.d. 0.017)
VIF	0.407	(s.d. 0.035)
VPR	0.494	(s.d. 0.044)
TAT	1.229	(s.d. 0.104)
REV	0.596	(s.d. 0.052)
VPU	0.902	(s.d. 0.079)
ENV	0.889	(s.d. 0.051)
NEF	0.928	(s.d. 0.073)

**Estimated Distance per Site: 0.194**

# Statistical Test of Models

(Goldman,1990)

Data: 3 sequences of length L

ACGTTGCAA ...

AGCTTTTGA ...

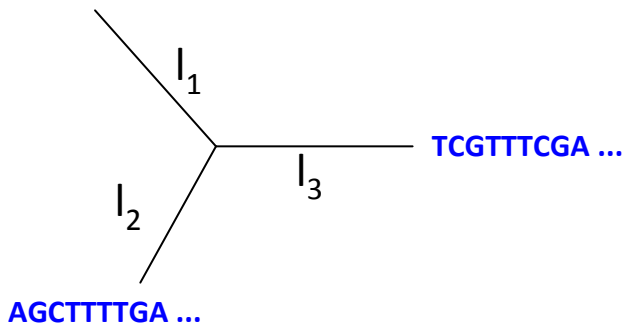
TCGTTTCGA ...

**A.** Likelihood (free multinomial model 63 free parameters)

$$L_1 = p_{AAA}^{\#AAA*} \dots p_{AAC}^{\#AAC*} \dots p_{TTT}^{\#TTT} \quad \text{where } p_{N_1N_2N_3} = \#(N_1N_2N_3)/L$$

**B.** Jukes-Cantor and unknown branch lengths

ACGTTGCAA ...



$$L_2 = p_{AAA}(l_1', l_2', l_3')^{\#AAA*} \dots p_{TTT}(l_1', l_2', l_3')^{\#TTT}$$

Test statistics: I.  $\sum (\text{expected-observed})^2/\text{expected}$  or II:  $-2 \ln Q = 2(\ln L_1 - \ln L_2)$

JC69 Jukes-Cantor: 3 parameters =>  $\chi^2$  60 d.of freedom

Problems: i. To few observations pr. pattern.

ii. Many competing hypothesis.

Parametric bootstrap:

i. Maximum likelihood to estimate the parameters.

ii. Simulate with estimated model.

iii. Make simulated distribution of  $-2 \ln Q$ .

iv. Where is real  $-2 \ln Q$  in this distribution?

# Rate variation between sites:iid each site

The rate at each position is drawn independently from a distribution, typically a  $\Gamma$  (or lognormal) distribution.  $G(a,b)$  has density  $x^{\beta-1}e^{-\alpha x}/\Gamma(\beta)$ , where  $\alpha$  is called scale parameter and  $\beta$  form parameter.

Let  $L(p_i, \Theta, t)$  be the likelihood for observing the  $i$ 'th pattern,  $t$  all time lengths,  $\Theta$  the parameters describing the process parameters and  $f(r_i)$  the continuous distribution of rate(s). Then 
$$L = \prod \int L(p_i, \Theta, r_i) f(r_i) dr_i$$

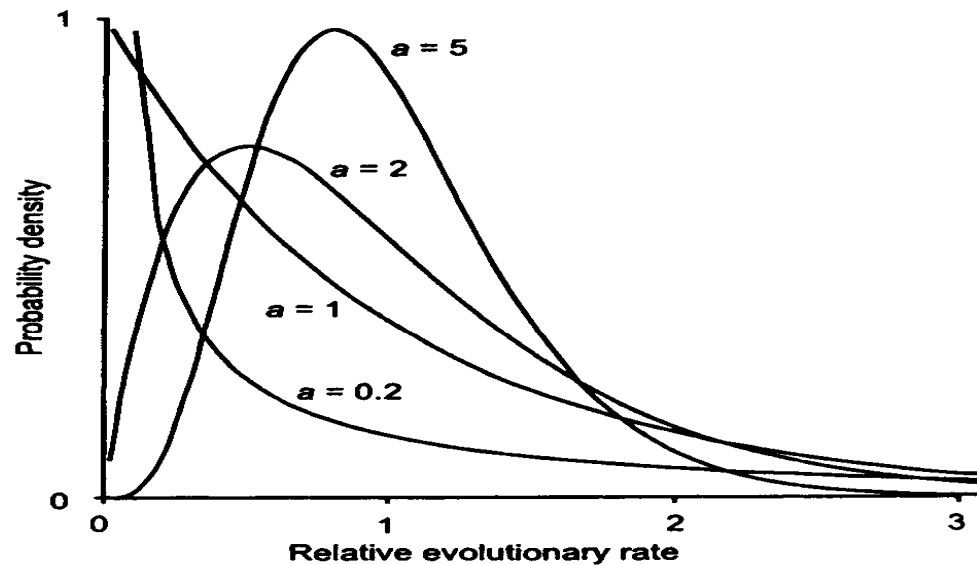


FIGURE 2.3. Gamma distributions of substitution rates among sites for different gamma parameters ( $a$ 's).