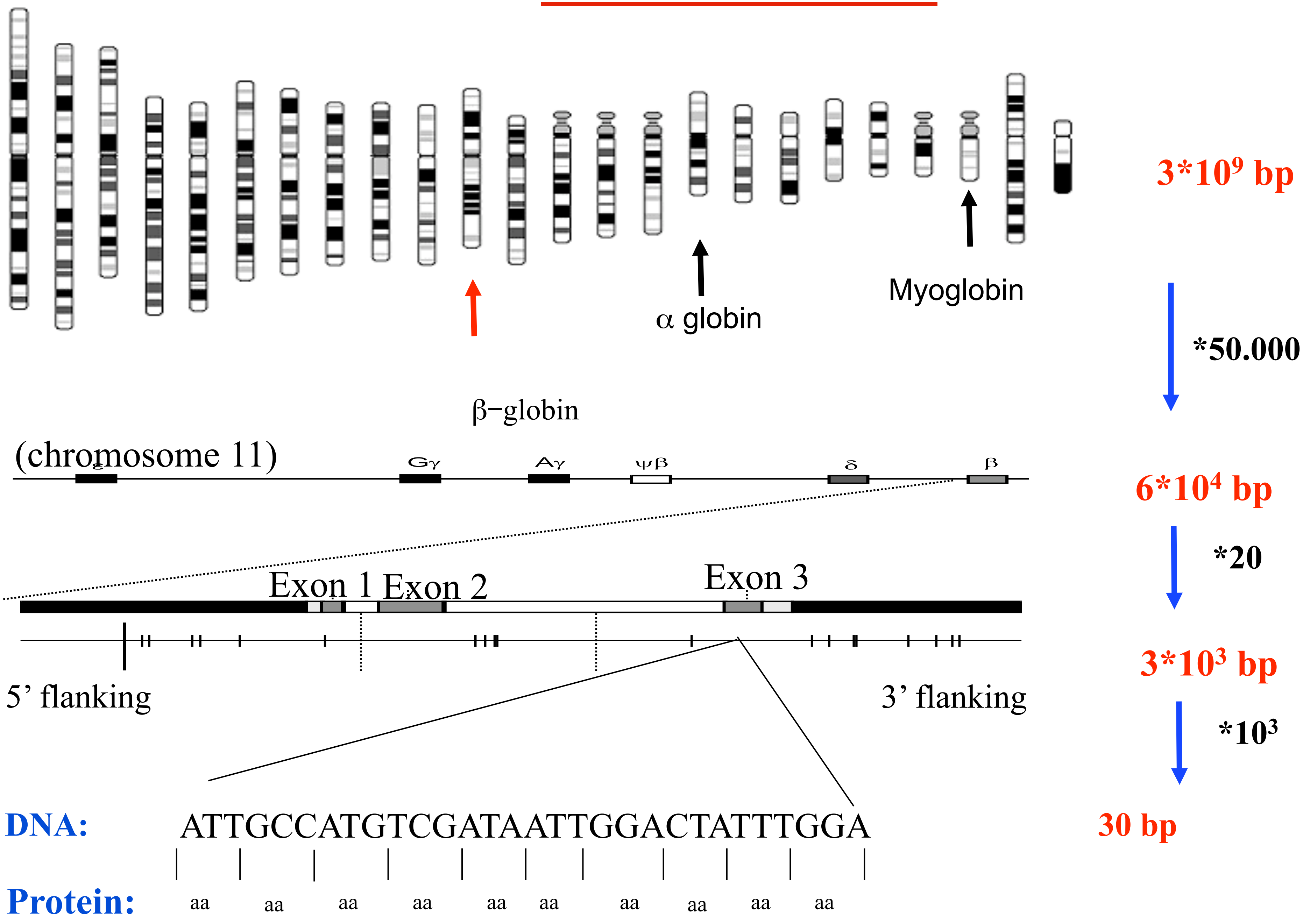
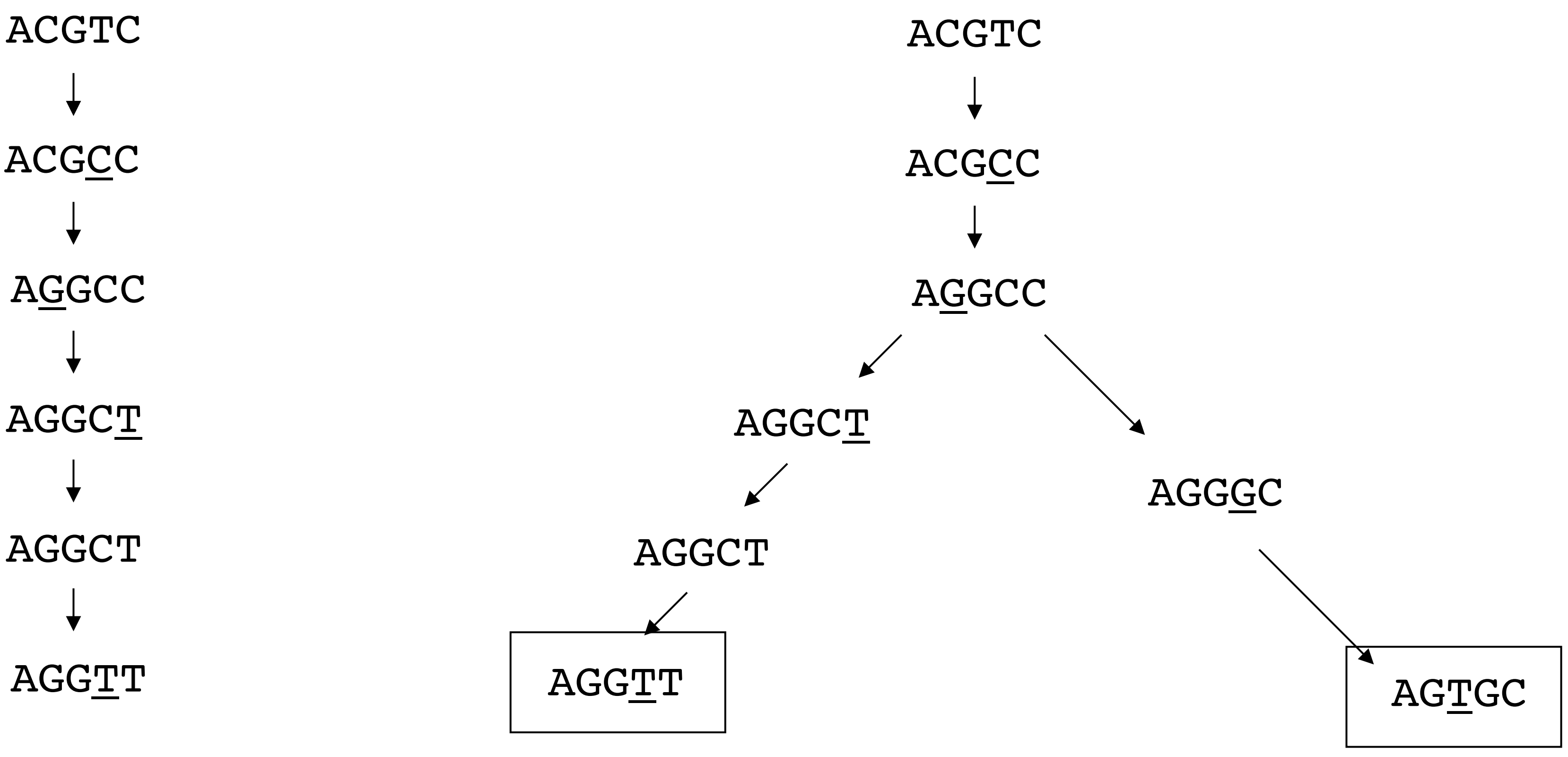
The Human Genome
(Harding & Sanger)



Central Problems: History cannot be observed, only end products.

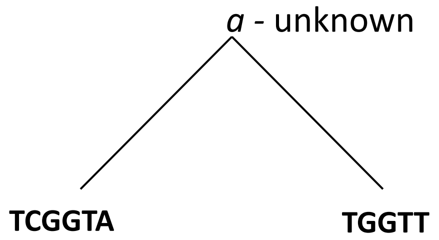


Even if History could be observed, the underlying process couldn't !!

Simplifying Assumptions I

Data: s1=TCGGTA,s2=TGGTT

Biological setup



Probability of Data

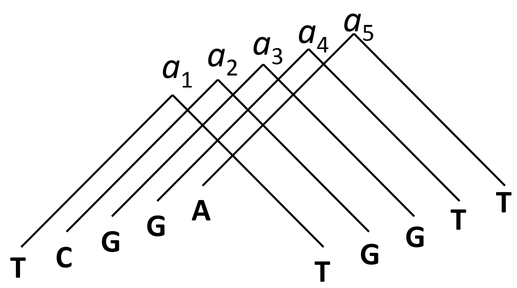
$$P = \sum_{a} P(a) * P(a \to \text{TCGGTA}) P(a \to \text{TGGTT})$$

1) Only substitutions.

```
s1    TCGGTA   →   s1    TCGGA
s2    TGGT-T       s2    TGGTT
```

$$P = \sum_{a} P(a) * P(a \to \text{TCGGA}) P(a \to \text{TGGTT})$$

2) Processes in different positions of the molecule are independent, so the probability for the whole alignment will be the product of the probabilities of the individual patterns.



$$P = \prod_{i=1}^{5} \sum_{a_i} P_i(a_i) * P_i(a_i \to \text{s1}_i) P_i(a_i \to \text{s2}_i)$$

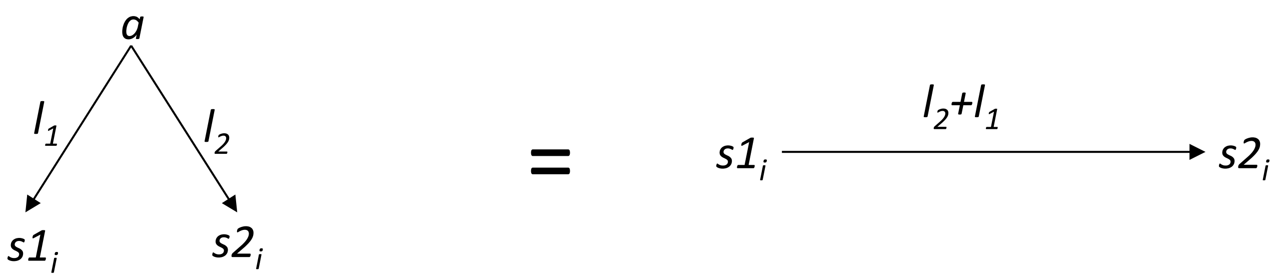
Simplifying Assumptions II

3) The evolutionary process is the same in all positions

$$P = \prod_{i=1}^{5} \sum_{a} P(a_i) * P(a_i \rightarrow s1_i) P(a_i \rightarrow s2_i)$$

4) Time reversibility: Virtually all models of sequence evolution are time reversible. I.e. $\pi_i P_{i,j}(t) = \pi_j P_{j,i}(t)$, where π_i is the stationary distribution of i and $P_t(i\text{->}j)$ the probability that state i has changed into state j after t time. This implies that

$$\sum_{a} P(a) * P_{l_1}(a_i \rightarrow s1_i) P_{l_2}(a_i \rightarrow s2_i) \quad = \quad P(s1_i) * P_{l_1 + l_2}(s1_i \rightarrow s2_i)$$



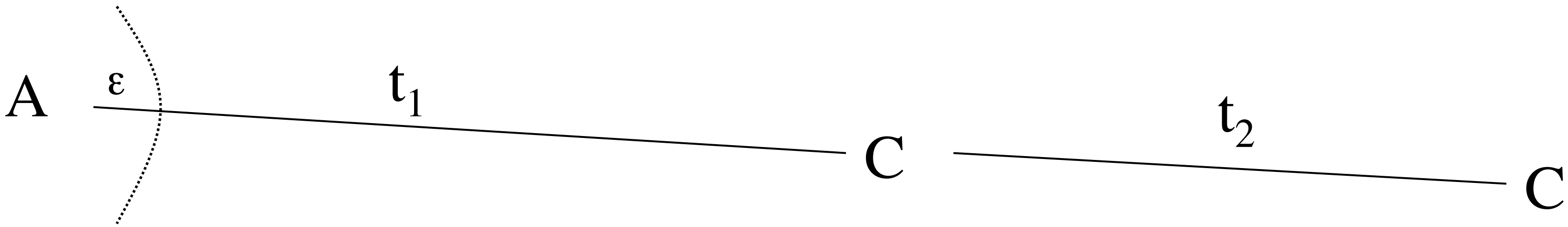
$$P = \prod_{i=1}^{5} P(s1_i) P(s1_i \rightarrow s2_i)$$

Simplifying assumptions III

5) The nucleotide at any position evolves following a continuous time Markov Chain.

$P_{i,j}(t)$ continuous time markov chain on the state space {A,C,G,T}.

$$\lim_{\varepsilon \to 0} \frac{P_{i,j}(\varepsilon)}{\varepsilon} = q_{ij} \qquad \lim_{\varepsilon \to 0} \frac{P_{i,i}(\varepsilon) - 1}{\varepsilon} = -q_{ii}$$



Q - rate matrix:

		TO			
		A	**C**	**G**	**T**
F	**A**	$-(q_{A,C}+q_{A,G}+q_{A,T})$	$q_{A,C}$	$q_{A,G}$	$q_{A,T}$
R	**C**	$q_{C,A}$	$-(q_{C,A}+q_{C,G}+q_{C,T})$	$q_{C,G}$	$q_{C,T}$
O	**G**	$q_{G,A}$	$q_{G,C}$	$-(q_{G,A}+q_{G,C}+q_{G,T})$	$q_{G,T}$
M	**T**	$q_{T,A}$	$q_{T,C}$	$q_{T,G}$	$-(q_{T,A}+q_{T,C}+q_{T,G})$

6) The rate matrix, **Q**, for the continuous time Markov Chain is the same at all times (and often all positions). However, it is possible to let the rate of events, r_i, vary from site to site, then the term for passed time, t, will be substituted by r_i*t.

Q and P(t)

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

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

i. P(0) = I

ii. P(ε) close to I+εQ for ε small

iii. P'(0) = Q.

iv. lim 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

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

Jukes-Cantor (JC69): Total Symmetry

Rate-matrix, R:

T O

		A	C	G	T
F	A	$-3*\alpha$	α	α	α
R	C	α	$-3*\alpha$	α	α
O	G	α	α	$-3*\alpha$	α
M	T	α	α	α	$-3*\alpha$

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

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

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

$$P = P(s1)\prod_{i=1}^{5} P(s1_i \to s2_i) = (\frac{1}{4})^5 \mathrm{P(T \to T)P(C \to G)P(G \to G)P(G \to T)P(A \to T)}$$

$$= (\frac{1}{4})^5 (\frac{1}{4})^5 (1 + 3e^{-4a})^2 (1 - e^{-4a})^3$$

Principle of Inference: Likelihood

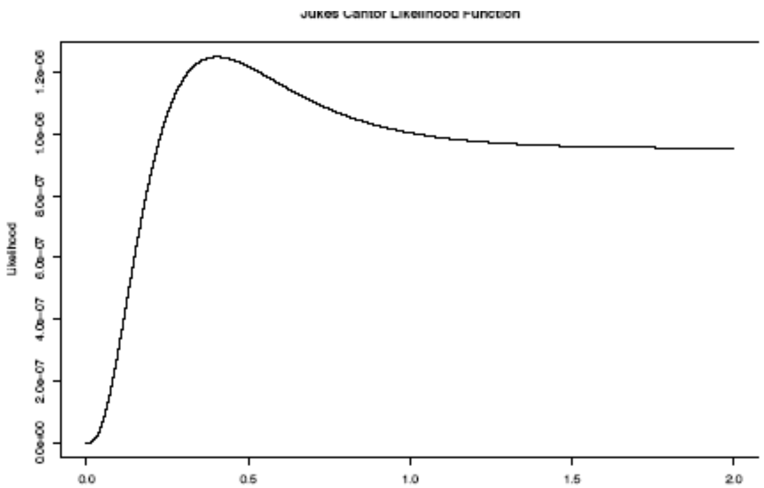
Likelihood function L() – the probability of data as function of parameters: L(Θ,D)

LogLikelihood Function – l(): ln(L(Θ,D))

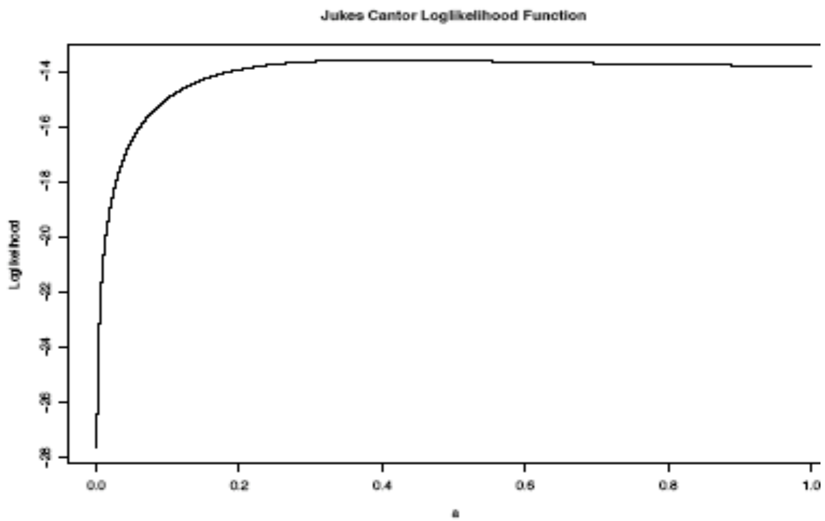
If the data is a series of independent experiments L() will become a product of Likelihoods of each experiment, l() will become the sum of LogLikelihoods of each experiment

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

Likelihood



LogLikelihood



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

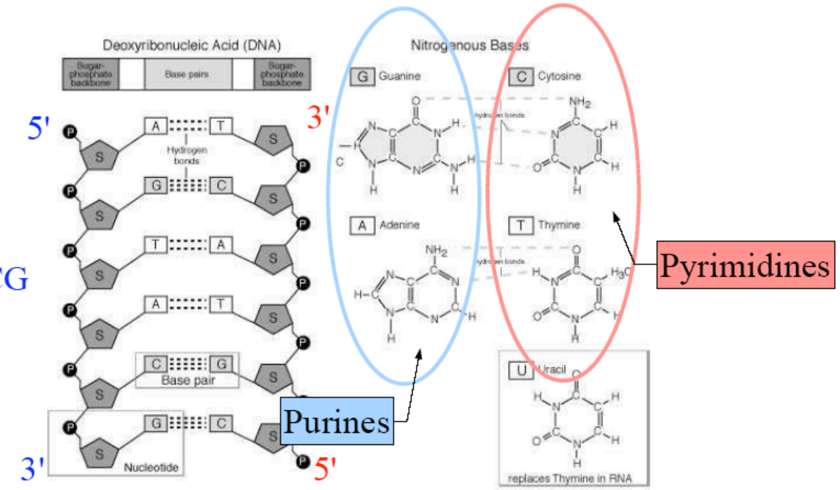
Kimura 2-parameter model - K80

Q:

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

a = α*t **b = β*t**

AGTACG



P(t)

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

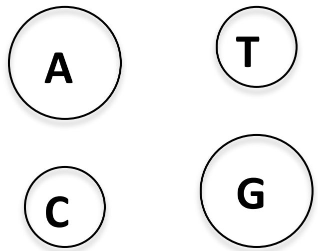
Felsenstein81 & Hasegawa, Kishino & Yano 85

Unequal base composition: (Felsenstein, 1981 F81)

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

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

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



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

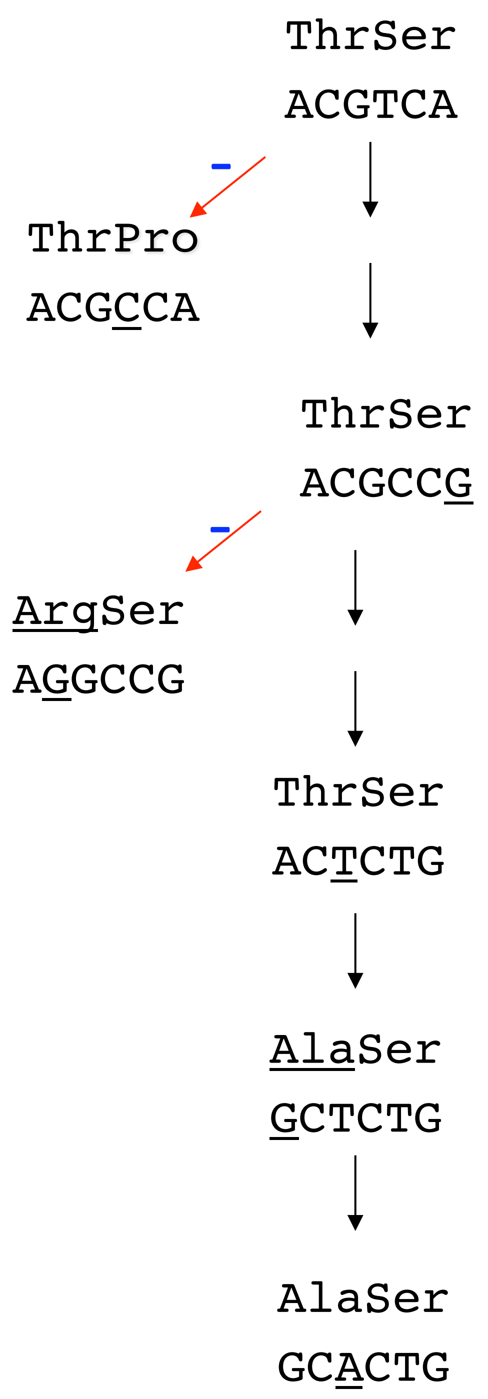
$$Q_{i,j} = \begin{cases} (\alpha/\beta)^* C^* \pi_j & \text{i- >j a transition} \\ C^* \pi_j & \text{i- >j a transversion} \end{cases}$$

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

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 (3rd)

1-1-1-1 (3rd)

ii. T→A (2nd)

Problems:

i. Not all fit into those categories.

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

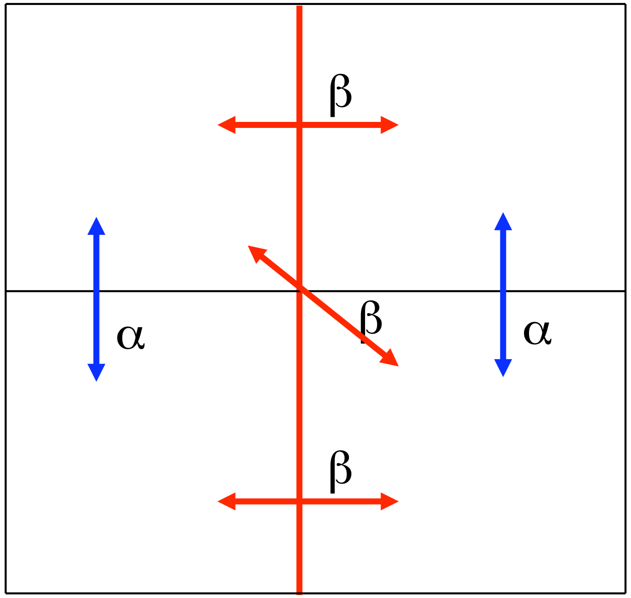
Possible events if the genetic code remade from Li,1997

Possible number of substitutions: 61 (codons)*3 (positions)*3 (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)→(?,?)

1-1-1-1	**(f*α,f*β)**
2-2	**(α,f*β)**
4	**(α, β)**

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)

$Z(\alpha t,\beta t) = .50[1+\exp(-2\alpha t) - 2\exp(-t(\alpha+\beta)]$ transition
$Y(\alpha t,\beta t) = .25[1-\exp(-2\beta t)]$ transversion
$X(\alpha t,\beta t) = .25[1+\exp(-2\alpha t) + 2\exp(-t(\alpha+\beta)]$ identity

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

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 $K_s = .6644$ $K_a = .1127$

Summary of Models and Phylogenies

Models

- The need for models for models of sequence evolution
- The assumptions of models
- The basic models
- selection and coding regions

Phylogenies

- Enumeration of Phylogenies
- The Basic Principles
    - Distance
    - Parsimony
    - Likelihood

Some Definitions

State space – a set often corresponding of possible observations ie {A,C,G,T}.

A random variable, X can take values in the state space with probabilities ie P{X=A} = ¼. The value taken often indicated by small letters - x

Stochastic Process is a set of time labeled stochastic variables X_t
ie $P\{X_0=A, X_1=C, .., X_5=G\} = .00122$

Time can be discrete or continuous, in our context it will almost always mean natural numbers, N {0,1,2,3,4..}, or an interval on the real line, R.

Markov Property: $P\{X_i | X_{i-1},\ldots,X_0\} = P\{X_i | X_{i-1}\}$
ie $P\{X_i, X_{i-1},\ldots,X_0\} = P\{X_0\}P\{X_1 | X_0\}..P\{X_i | X_{i-1}\}$

Time Homogeneity – the process is the same for all t.

From Q to P for Jukes-Cantor

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

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

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

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

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 $Q^i = B\Lambda B^{-1} B\Lambda B^{-1} ... B\Lambda B^{-1} = B\Lambda^i B^{-1}$

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

Finding Λ: $det\ (Q-\lambda I)=0$ *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: $$\sum_{i=0}^{\infty} \frac{(tQ)^i}{i!} \sim \sum_{i=0}^{k} \frac{(tQ)^i}{i!}$$ *where k ~6-10*