BINF6201/8201

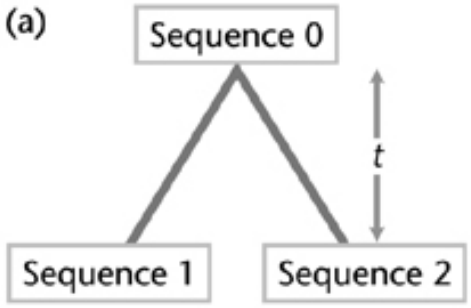
Evolutionary Change in DNA sequences

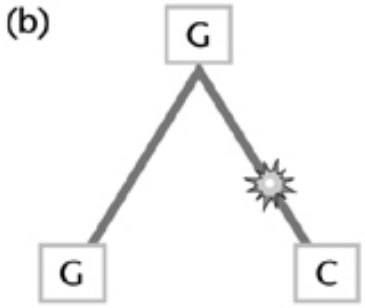
09-20-2016

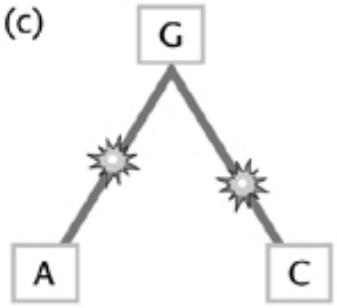
Nucleotide substitutions in a DNA sequence

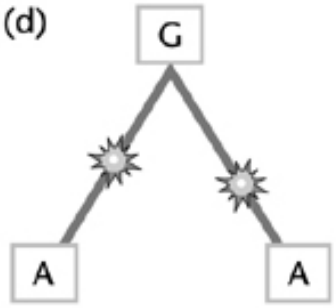
- At the molecular level, the basic process in the evolution of a DNA sequence is the change in nucleotides with time, in which one nucleotide substitutes another through fixation.
- The number of substitutions per site between two sequences since their divergence from the ancestor gene, ***d***, forms the base to reconstruct their evolutionary history and to estimate their rate of evolution.
- d is usually larger than the number of observed different nucleotides per site between the two sequences, D, i.e., $d \geq D$.

Given $D = 3/10$, what is d ?





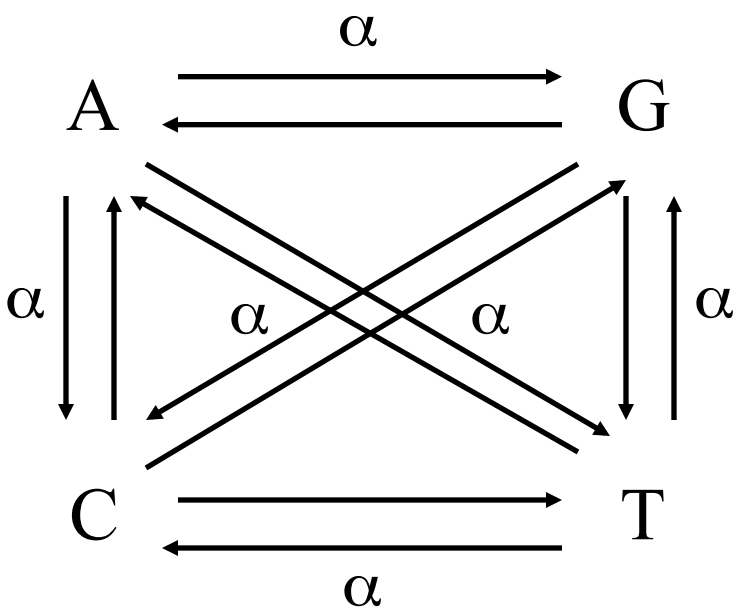




Nucleotide substitution in a DNA sequence

- Because we cannot repeat the evolutionary history to observe the nucleotide substitution process, we rely on developing mathematical models that account for the substitution process.
- The earliest DNA substitution model is the **Jukes-Cantor one-parameter (J-C)** model (1969), which assumes that each nucleotide has equal probability or rate to be substituted by any of the other three in a fixed period of time.

J-C model



The J-C model for nucleotide substitution

- Assume that the nucleotide at a certain site in DNA sequence is A at time 0, then, what is the probability that this site will be still occupied by A at time t, $p_{AA}(t)$?
- Let's first consider the probability of having A in this site after one single unit time, $p_{AA}(1)$. In this case, $p_{AA}(1) = 1 - 3\alpha$.
- Now let's compute $p_{AA}(2)$ by considering the two possible scenarios for this site to still be occupied by A:

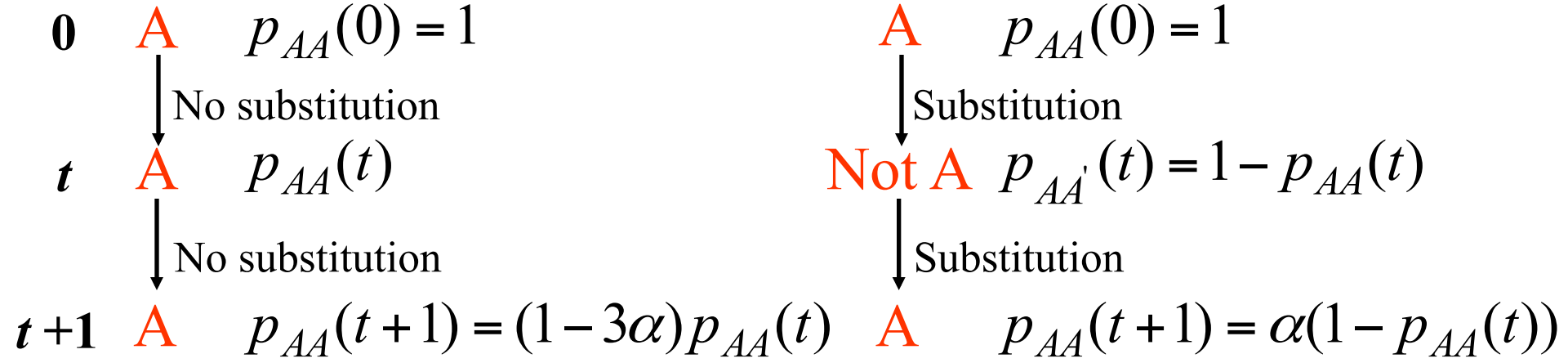
	Scenario 1		**Scenario 2**	
$t = 0$	A	$p_{AA}(0)$	A	$p_{AA}(0)$
	↓ No substitution		↓ Substitution	
$t = 1$	A	$p_{AA}(1)$	Not A	$p_{AA'}(1) = 1 - p_{AA}(1)$
	↓ No substitution		↓ Substitution	
$t = 2$	A	$p_{AA}(2) = (1-3\alpha)p_{AA}(1)$	A	$p_{AA}(2) = \alpha(1 - p_{AA}(1))$

The J-C model for nucleotide substitution

- Therefore the probability of having A at time 2 is the sum of this two probabilities:

$$p_{AA}(2) = (1-3\alpha)p_{AA}(1) + \alpha(1-p_{AA}(1)).$$

- In general, if the initial nucleotide at a site is A, then after a period of time t, it can be any nucleotide on this site.



- The probability of having A at this site at time $t + 1$ is,

$$p_{AA}(t+1) = (1-3\alpha)p_{AA}(t) + \alpha(1-p_{AA}(t)).$$

- Rearranging this equation, we have:

$$p_{AA}(t+1) = p_{AA}(t) - 3\alpha p_{AA}(t) + \alpha - \alpha p_{AA}(t),$$

$$p_{AA}(t+1) - p_{AA}(t) = -3\alpha p_{AA}(t) + \alpha - \alpha p_{AA}(t) = -4\alpha p_{AA}(t) + \alpha.$$

The J-C model for nucleotide substitution

- Treating this as a continuous-time model, we have,

$$\Delta p_{AA}(t) = -4\alpha p_{AA}(t) + \alpha$$

$$\frac{dp_{AA}(t)}{dt} = -4\alpha p_{AA}(t) + \alpha$$

$$= 4\alpha(\frac{1}{4} - p_{AA}(t)),$$

$$\frac{dp_{AA}(t)}{(1/4 - p_{AA}(t))} = 4\alpha dt,$$

$$\int \frac{dp_{AA}(t)}{(p_{AA}(t) - 1/4)} = -\int 4\alpha dt,$$

$$\ln|p_{AA}(t) - \frac{1}{4}| = -4\alpha t + C',$$

$$p_{AA}(t) - \frac{1}{4} = Ce^{-4\alpha t},$$

$$p_{AA}(t) = \frac{1}{4} + Ce^{-4\alpha t}.$$

When t = 0, $p_{AA}(t) = p_{AA}(0)$, therefore,

$$p_{AA}(0) = \frac{1}{4} + Ce^{0},$$

$$C = p_{AA}(0) - \frac{1}{4}.$$

Therefore,

$$p_{AA}(t) = \frac{1}{4} + (p_{AA}(0) - \frac{1}{4})e^{-4\alpha t}.$$

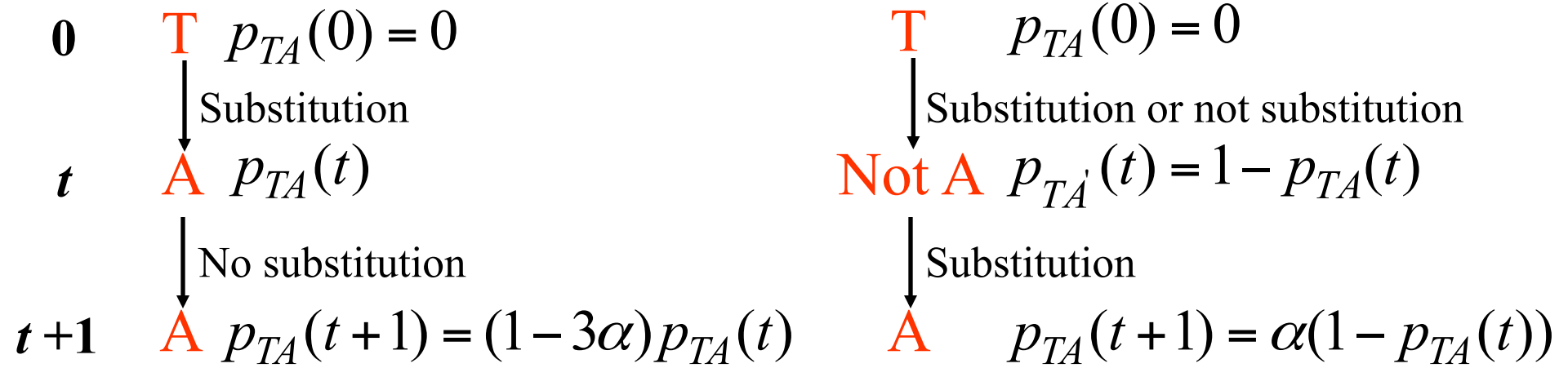
The J-C model for nucleotide substitution

- Due to the equivalence of the four nucleotides in the J-C model, this equation holds for any initial nucleotide N, and any final nucleotide X, therefore, we have,

$$p_{NX}(t+1) = (1-3\alpha)p_{NX}(t) + \alpha(1-p_{NX}(t)) = -4\alpha p_{NX}(t) + \alpha \text{ and}$$

$$p_{NX}(t) = \frac{1}{4} + (p_{NX}(0) - \frac{1}{4})e^{-4\alpha t}.$$

Example for P_{TA} (t +1):



$$p_{TA}(t+1) = (1-3\alpha)p_{TA}(t) + \alpha(1-p_{AA}(t)),$$

$$p_{TA}(t+1) - p_{TA}(t) = -3\alpha p_{TA}(t) + \alpha - \alpha p_{TA}(t) = -4\alpha p_{TA}(t) + \alpha.$$

The J-C model for nucleotide substitution

- If we start with A, then $p_{AA}(0) = 1$, and we have,

$$p_{AA}(t) = \frac{1}{4} + (p_{AA}(0) - \frac{1}{4})e^{-4\alpha t} = \frac{1}{4} + (1 - \frac{1}{4})e^{-4\alpha t},$$

$$p_{AA}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}.$$

- If we start with T, then $p_{TA}(0) = 0$, and we have,

$$p_{TA}(t) = \frac{1}{4} + (p_{TA}(0) - \frac{1}{4})e^{-4\alpha t} = \frac{1}{4} + (0 - \frac{1}{4})e^{-4\alpha t},$$

$$p_{TA}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}.$$

The J-C model for nucleotide substitution

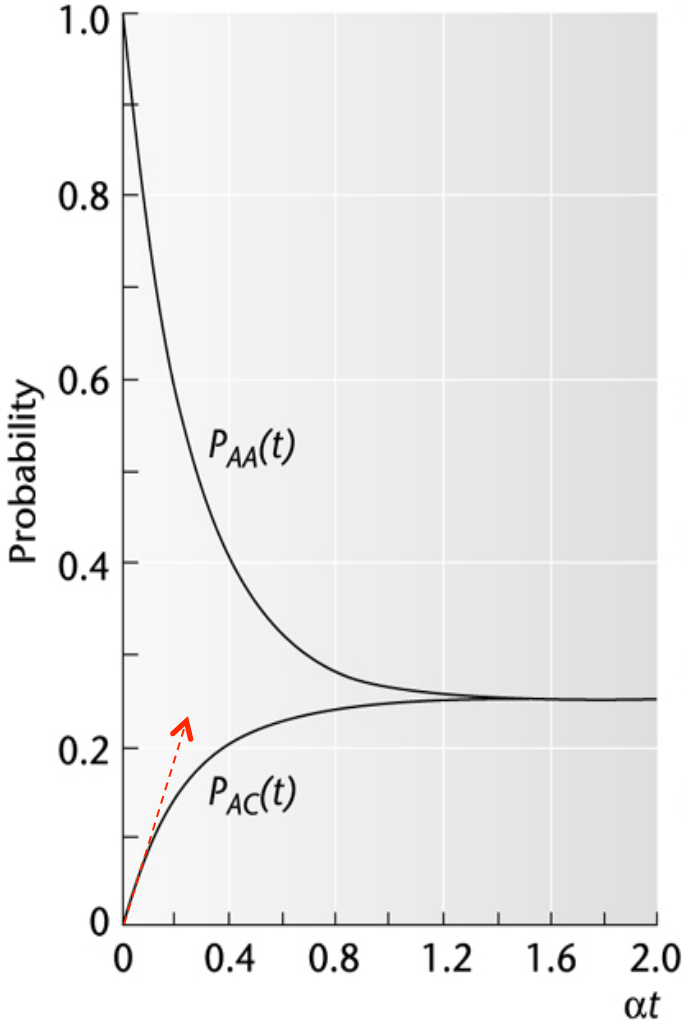
- Generally, we can write the probability that a nucleotide will remain unchanged after t units of time as,

$$p_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t},$$

and the probability that the a nucleotide will change to a different one after t units of time as

$$p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}.$$

- When t approaches to infinite, both p_{ii} and p_{ij} approach to ¼ .

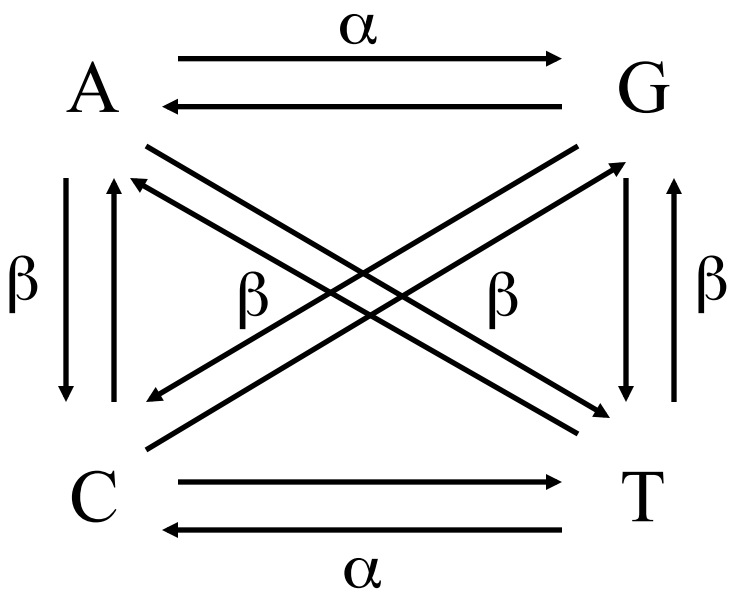


- If the original sequence consists of the same nucleotide, e.g., a string of A's, then p_{ij}(t) is also the frequency of j at time t in the sequence. Thus the equilibrium frequency of each of the four nucleotides is 1/4.
- When t is small, change in p_{ik} is linear, but it becomes nonlinear as t increases because of back substitutions.

Kimura's two-parameter model for nucleotide substitution

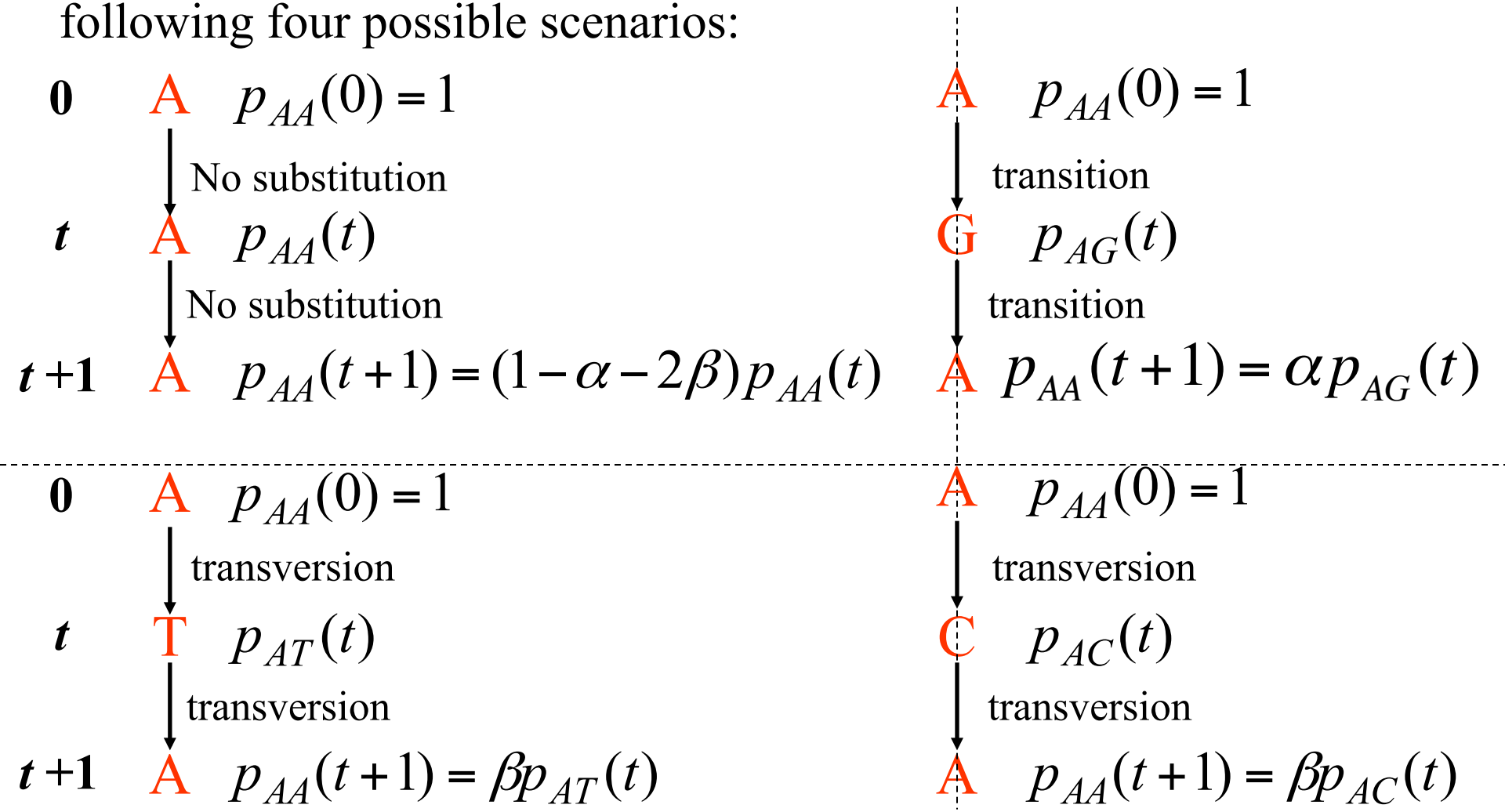
- In reality, a nucleotide is more likely to undergo a transitional substitution than a transversional substitution.
- To account for such difference between transitional and transversional substitution rates, Kimura (1983) proposed a two-parameter (K2P) model, in which the rate (probability) for a nucleotide to undergo a transitional substitution is α, and a transversional substitution, β.

Kimura's two-parameter (K2P) model



Kimura's two-parameter model

- If the initial nucleotide at a site is A, after a period of time t, it can be any nucleotide on this site. To compute the probability of having A remain at the site at time t+1, $P_{AA}(t+1)$, we need to consider the following four possible scenarios:



Kimura's two-parameter model

- Therefore, the final probability is the sum of these four probabilities:

$$p_{AA}(t+1) = (1-\alpha-2\beta)p_{AA}(t) + \alpha p_{AG}(t) + \beta p_{AC}(t) + \beta p_{AT}(t),$$

$$p_{AA}(t+1) - p_{AA}(t) = -(\alpha+2\beta)p_{AA}(t) + \alpha p_{AG}(t) + \beta p_{AC}(t) + \beta p_{AT}(t).$$

- Treating this as a continuous-time model, we have,

$$\frac{dp_{AA}(t)}{dt} = -(\alpha+2\beta)p_{AA}(t) + \alpha p_{AG}(t) + \beta p_{AC}(t) + \beta p_{AT}(t).$$

- Similarly, we can derive the following equations for an initial A at a site to become a G, C and T at time $t+1$:

$$p_{AG}(t+1) = \alpha p_{AA}(t) + (1-\alpha-2\beta)p_{AG}(t) + \beta p_{AC}(t) + \beta p_{AT}(t),$$

$$p_{AC}(t+1) = \beta p_{AA}(t) + \beta p_{AG}(t) + (1-\alpha-2\beta)p_{AC}(t) + \alpha p_{AT}(t),$$

$$p_{AT}(t+1) = \beta p_{AA}(t) + \beta p_{AG}(t) + \alpha p_{AC}(t) + (1-\alpha-2\beta)p_{AT}(t).$$

Or

$$\frac{dp_{AG}}{dt} = \alpha p_{AA}(t) - (\alpha+2\beta)p_{AG}(t) + \beta p_{AC}(t) + \beta p_{AT}(t),$$

$$\frac{dp_{AC}}{dt} = \beta p_{AA}(t) + \beta p_{AG}(t) - (\alpha+2\beta)p_{AC}(t) + \alpha p_{AT}(t),$$

$$\frac{dp_{AT}}{dt} = \beta p_{AA}(t) + \beta p_{AG}(t) + \alpha p_{AC}(t) - (\alpha+2\beta)p_{AT}(t).$$

Substitution rate matrix

- We can write these four equations in a matrix form,

$$\left[p_{AA}(t+1)\;\; p_{AG}(t+1)\;\; p_{AC}(t+1)\;\; p_{AT}(t+1)\right] = p_{A\cdot}(t)\mathrm{M}$$

$$= \left[p_{AA}(t)\;\; p_{AG}(t)\;\; p_{AC}(t)\;\; p_{AT}(t)\right] \begin{bmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta) \end{bmatrix}$$

- By extending our analysis to other cases, we have,

$$\begin{bmatrix} p_{AA}(t+1) & p_{AG}(t+1) & p_{AC}(t+1) & p_{AT}(t+1) \\ p_{GA}(t+1) & p_{GG}(t+1) & p_{GC}(t+1) & p_{GT}(t+1) \\ p_{CA}(t+1) & p_{CG}(t+1) & p_{CC}(t+1) & p_{CT}(t+1) \\ p_{TA}(t+1) & p_{TG}(t+1) & p_{TC}(t+1) & p_{TT}(t+1) \end{bmatrix}$$

$$= \begin{bmatrix} p_{AA}(t) & p_{AG}(t) & p_{AC}(t) & p_{AT}(t) \\ p_{GA}(t) & p_{GG}(t) & p_{GC}(t) & p_{GT}(t) \\ p_{CA}(t) & p_{CG}(t) & p_{CC}(t) & p_{CT}(t) \\ p_{TA}(t) & p_{TG}(t) & p_{TC}(t) & p_{TT}(t) \end{bmatrix} \begin{bmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta) \end{bmatrix}$$

Substitution rate matrix

- The matrix,

$$M = \begin{array}{c} \begin{matrix} A & G & C & T \end{matrix} \\ \begin{bmatrix} 1-\alpha-2\beta & \alpha & \beta & \beta \\ \alpha & 1-\alpha-2\beta & \beta & \beta \\ \beta & \beta & 1-\alpha-2\beta & \alpha \\ \beta & \beta & \alpha & 1-\alpha-2\beta \end{bmatrix} \end{array} \begin{matrix} A \\ G \\ C \\ T \end{matrix}$$

  is called **substitution rate matrix**, whose item m_{ij} is the probability to change nucleotide i to j. The sum over each row and each column should be 1.

- This matrix is also called a **Markov chain state transition matrix**, because it defines a **Markov chain process**.

- The substitution rate matrix for the J-C model is,

$$M = \begin{array}{c} \begin{matrix} A & G & C & T \end{matrix} \\ \begin{bmatrix} 1-3\alpha & \alpha & \alpha & \alpha \\ \alpha & 1-3\alpha & \alpha & \alpha \\ \alpha & \alpha & 1-3\alpha & \alpha \\ \alpha & \alpha & \alpha & 1-3\alpha \end{bmatrix} \end{array} \begin{matrix} A \\ G \\ C \\ T \end{matrix}$$

Kimura's two-parameter

- In general, given a substitution rate matrix M, the probability that a initial nucleotide *i* at time *t* will be substituted by *j* at time *t* + 1 is given by vector dot product,

$$p_{ij}(t+1) = \sum_{k \in \{A,C,G,T\}} p_{ik}(t) m_{kj}$$

- By solving the equations for A to remain unchanged, or change to G, C or T in the K2P model, we can find the probability that an initial A at a site

$$\frac{dp_{AA}(t)}{dt} = -(\alpha + 2\beta)p_{AA}(t) + \alpha p_{AG}(t) + \beta p_{AC}(t) + \beta p_{AT}(t)$$

$$\frac{dp_{AG}(t)}{dt} = \alpha p_{AA}(t) - (\alpha + 2\beta)p_{AG}(t) + \beta p_{AC}(t) + \beta p_{AT}(t)$$

$$\frac{dp_{AC}(t)}{dt} = \beta p_{AA}(t) + \beta p_{AG}(t) - (\alpha + 2\beta)p_{AC}(t) + \alpha p_{AT}(t)$$

$$\frac{dp_{AT}(t)}{dt} = \beta p_{AA}(t) + \beta p_{AG}(t) + \alpha p_{AC}(t) - (\alpha + 2\beta)p_{AT}(t)$$

1. remains unchanged at *t*, $p_{AA}(t)$,
2. undergoes a transitional substitution at *t*, $p_{AG}(t)$, and
3. undergoes a transversional substitution at *t*, p_{AC}(t) or $p_{AT}(t)$.

Kimura's two-parameter

- The probability that the nucleotide remains unchanged should be the same for all nucleotides, i.e., $p_{AA}= p_{GG}=p_{CC} = p_{TT}$. Let's denote this probability by $X(t)$, then,

$$X(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} + \frac{1}{2}e^{-2(\alpha+\beta)t}.$$

- The probability that the nucleotide undergoes a transitional substitution should be the same for all nucleotides, i.e., $p_{AG}= p_{GA}=p_{CT} = p_{TC}$. Let's denote this probability by $Y(t)$, then,

$$Y(t) = \frac{1}{4} + \frac{1}{4}e^{-4\beta t} - \frac{1}{2}e^{-2(\alpha+\beta)t}.$$

- The probability that the nucleotide undergoes a specific type of transversional substitution should be the same for all nucleotides, i.e., $p_{AC}= p_{CA}=p_{AT} = p_{TA}= p_{GC}= p_{CG}=p_{GT} = p_{TG}$. Let's denote this probability by $Z(t)$, then,

$$Z(t) = \frac{1}{4} - \frac{1}{4}e^{-4\beta t}.$$

- Clearly, as there are two types transversional substitution, we have,

$$X(t) + Y(t) + 2Z(t) = 1.$$

Comparison of the J-C and K2P models

- In all these cases, when $t \to \infty$, $X(t) \to 1/4$, $Y(t) \to 1/4$ and $Z(t) \to 1/4$.
- Therefore, as in the J-C model, the equilibrium value for the frequency of each nucleotide in the P2K model is also 1/4.

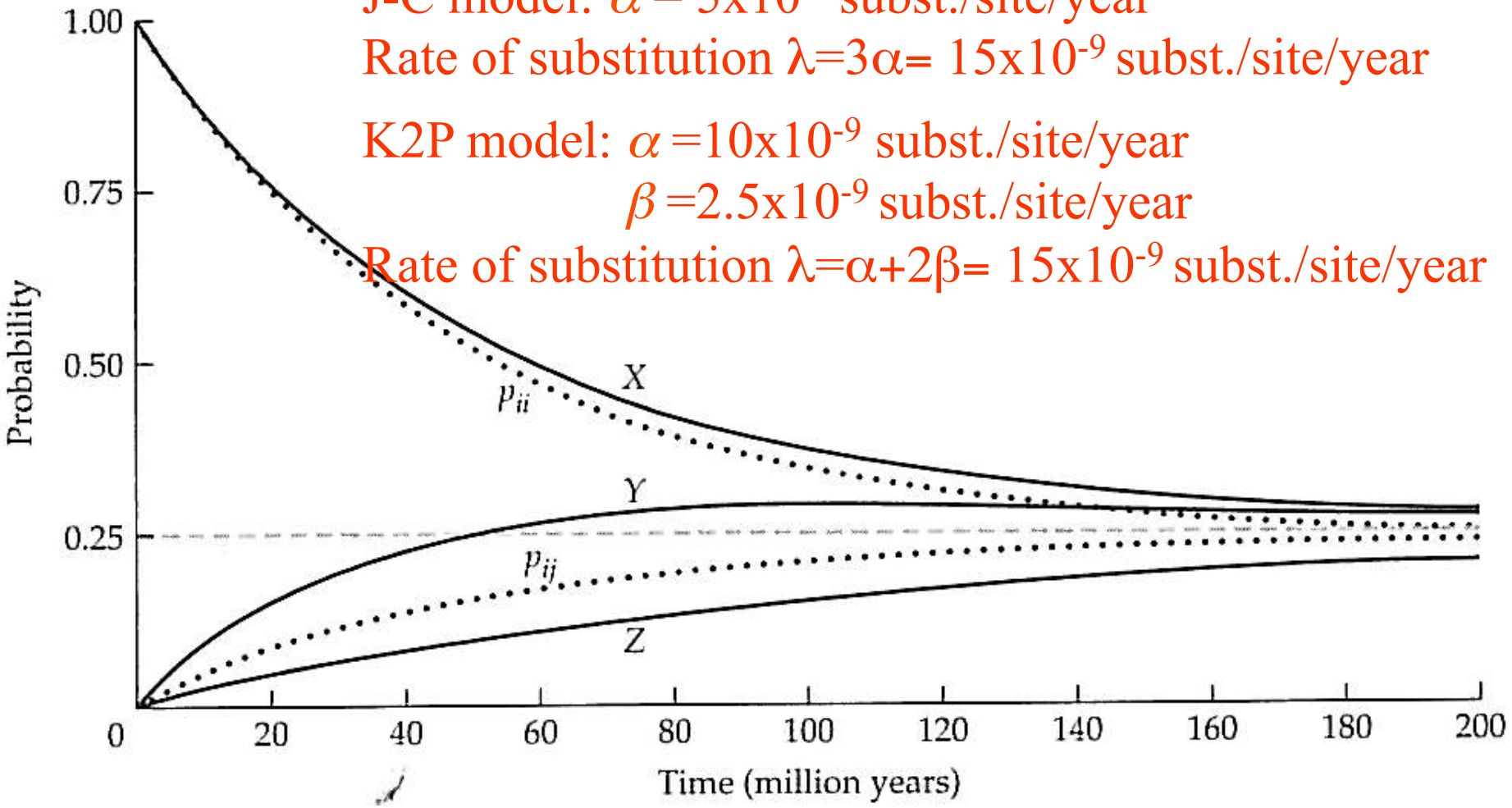
J-C model: $\alpha = 5\text{x}10^{-9}$ subst./site/year

Rate of substitution $\lambda = 3\alpha = 15\text{x}10^{-9}$ subst./site/year

K2P model: $\alpha = 10\text{x}10^{-9}$ subst./site/year

$\beta = 2.5\text{x}10^{-9}$ subst./site/year

Rate of substitution $\lambda = \alpha + 2\beta = 15\text{x}10^{-9}$ subst./site/year



Other models for nucleotide substitution

- To account for the different substitution rates among different nucleotides, models that contain more parameters have been developed.
- The choice of models depends on the problem to solve and the dataset to use.
- If the dataset is too small, the use of a more complex model may not necessarily improve the analysis.

TABLE 3.1 **Models of nucleotide substitution**

O\S[a]	A	T	C	G
a. Two-parameter model (Kimura 1980)				
A	$1-\alpha-2\beta$	β	β	α
T	β	$1-\alpha-2\beta$	α	β
C	β	α	$1-\alpha-2\beta$	β
G	α	β	β	$1-\alpha-2\beta$
b. Four-parameter model (Blaisdell 1985)				
A	$1-\alpha-2\gamma$	γ	γ	α
T	δ	$1-\alpha-2\delta$	α	δ
C	δ	β	$1-\beta-2\delta$	δ
G	β	γ	γ	$1-\beta-2\gamma$
c. Six-parameter model (Kimura 1981a)				
A	$1-2\alpha-\gamma$	γ	α	α
T	δ	$1-2\alpha-\delta$	α	α
C	β	β	$1-2\beta-\varepsilon$	ε
G	β	β	ξ	$1-2\beta-\xi$
d. Nine-parameter model				
A	$1-g_T\beta_1-g_C\gamma_1-g_G\alpha_1$	$g_T\beta_1$	$g_C\gamma_1$	$g_G\alpha_1$
T	$g_A\beta_1$	$1-g_A\beta_1-g_C\alpha_2-g_G\gamma_2$	$g_C\alpha_2$	$g_G\gamma_2$
C	$g_A\gamma_1$	$g_T\alpha_2$	$1-g_A\gamma_1-g_T\alpha_2-g_G\beta_2$	$g_G\beta_2$
G	$g_A\alpha_1$	$g_T\gamma_2$	$g_C\beta_2$	$1-g_A\alpha_1-g_T\gamma_2-g_C\beta_2$
e. General model				
A	$1-\alpha_{12}-\alpha_{13}-\alpha_{14}$	α_{12}	α_{13}	α_{14}
T	α_{21}	$1-\alpha_{21}-\alpha_{23}-\alpha_{24}$	α_{23}	α_{24}
C	α_{31}	α_{32}	$1-\alpha_{31}-\alpha_{32}-\alpha_{34}$	α_{34}
G	α_{41}	α_{42}	α_{43}	$1-\alpha_{41}-\alpha_{42}-\alpha_{43}$

[a]O, Original nucleotide; S, substitute nucleotide.

Time reversibility of substitution rate models

- A substitution process is called **time reversible** if the probability of starting nucleotide *i* and changing to *j* in a time interval is the same as the probability of starting from nucleotide *j* and going back to *i* in the same time duration. Mathematically, time reversibility requires,

$$p_{ij}(t)\tilde{p}_i = p_{ji}(t)\tilde{p}_j$$

  for all *i*, *j* and *t,* where $\tilde{p}_i$ and $\tilde{p}_j$ are the equilibrium frequency of nucleotides *i* and *j*, respectively.

- For $t = 1$, this equation becomes,

$$m_{ij}\tilde{p}_i = m_{ji}\tilde{p}_j$$

- For both the J-C and K2P models, we have,

$$\tilde{p}_A = \tilde{p}_G = \tilde{p}_C = \tilde{p}_T = 1/4 \quad \text{and} \quad m_{ij} = m_{ji}.$$

  Therefore, time reversibility holds for both the models.

- Time reversibility simplifies the theoretical study of nucleotide sequence evolution.