A MODEL OF EVOLUTIONARY BASE SUBSTITUTIONS AND ITS APPLICATION WITH SPECIAL REFERENCE TO RAPID CHANGE OF PSEUDOGENES*

NAOYUKI TAKAHATA AND MOTOO KIMURA

National Institute of Genetics, Mishima, 411 Japan

Manuscript received February 2, 1981
Revised copy received May 18, 1981

ABSTRACT

A model of evolutionary base substitutions that can incorporate different substitutional rates between the four bases and that takes into account unequal composition of bases in DNA sequences is proposed. Using this model, we derived formulae that enable us to estimate the evolutionary distances in terms of the number of nucleotide substitutions through comparative studies of nucleotide sequences. In order to check the validity of various formulae, Monte Carlo experiments were performed. These formulae were applied to analyze data on DNA sequences from diverse organisms. Particular attention was paid to problems concerning a globin pseudogene in the mouse and the time of its origin through duplication. We obtained a result suggesting that the evolutionary rates of substitution in the first and second codon positions of the pseudogene were roughly 10 times faster than those in the normal globin genes; whereas, the rate in the third position remained almost unchanged. Application of our formulae to histone genes H2B and H3 of the sea urchin showed that, in each of these genes, the rate in the third codon position is tremendously higher than that in the second position. All of these observations can easily and consistently be interpreted by the neutral theory of molecular evolution.

RECENT developments in DNA-sequencing techniques (MAXAM and GILBERT 1977; SANGER, NICKLEN and COULSON 1977), together with methods for amplifying gene copies in a bacterial plasmid, have made possible rapid determinations of DNA sequences of genes. Because data on DNA sequences are obtainable only from living organisms, it is necessary to develop mathematical models to estimate the number of evolutionary nucleotide substitutions through comparison of DNA sequences of homologous genes in related species.

Prior to a recent flood of data on nucleotide sequences, there already existed a large body of data on amino acid sequences in diverse organisms, and many mathematical models have been proposed to treat protein evolution in terms of amino acid substitutions, (see, for example, ZUCKERKANDL and PAULING 1965; FITCH and MARGOLIASH 1967; JUKES and CANTOR 1969; OHTA and KIMURA 1971; NEI 1975 for review). However, for comparative studies of nucleotide sequences, different mathematical models have to be employed. Sev-

* Contribution No. 1354 from The National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan.

eral authors have considered the problem of estimating the evolutionary distance of the homologous part of the genome between related species (JUKES and CANTOR 1969; KIMURA and OHTA 1972; MIYATA and YASUNAGA 1980; HOLMQUIST 1980; HOLMQUIST and PEARL 1980).

Recently, KIMURA (1980; 1981) developed three different models that can partially incorporate different substitutional rates between four bases and applied them to analyze data on various nucleotide sequences from several organisms. He showed that a preponderance of synonymous and other silent nucleotide substitutions is a general feature of molecular evolution and that this is consistent with the neutral theory (KIMURA 1968; see KIMURA 1979 for review).

In this paper, we extend the mathematical models of KIMURA (1980, 1981), and we derive appropriate formulae for estimating evolutionary distances in terms of the number of nucleotide substitutions per site. In addition to an analytical treatment, we used simulation methods to check the validity of the formulae and determine their range of applicability. This is necessary because formulae for estimating evolutionary distances generally do not have high resolving power when they are applied to evolutionarily distant organisms; they are accompanied by large error variances. Therefore, we conducted extensive Monte Carlo experiments in which the values of the parameters involved were greatly altered.

We have applied our formulae to analyze actual data on nucleotide sequences. Particular attention was paid to the evolution of the globin pseudogene in the mouse (NISHIOKA, LEDER and LEDER 1980; VANIN *et al.* 1980). Although similar analyses have recently been carried out by KIMURA (1980), PROUDFOOT and MANIATIS (1980) and, in more detail, by MIYATA and YASUNAGA (1981), we re-examined the problem to estimate the time of its origin and to discuss the evolutionary implications.

MODEL AND ANALYSIS

Let us consider a model of base substitutions, as shown in Figure 1, in which the four bases are represented by the letters U, A, C and G in terms of mRNA codes. The rates of base substitutions per unit time (say, year) between the four bases are designated by α, β, γ, δ and ε. For instance, α is the rate of transition-type substitutions from U to C or A to G, while β is the rate for the reverse directions. The rates of transversion-type substitutions are denoted by γ, δ and ε. In comparing two homologous sequences, there are 16 combinations of base pairs at each site. The possible combinations and their relative frequencies (probabilities) are listed in Table 1. They represent the expected relative frequencies of their occurrence in two homologous sequences. For example, S (the sum of S_i's) stands for the probability that the bases at a homologous site are identical and P $(= 2P_1 + 2P_2)$ the probability of their showing transition-type differences, while Q $(= 2Q_1 + 2Q_2)$ and R $(= 2R_1 + 2R_2)$ are the probabilities of transversion-type differences.

We denote the probabilities at time T of the four bases by $\mathrm{U}(T)$, $\mathrm{A}(T)$, $\mathrm{C}(T)$ and $\mathrm{G}(T)$. Starting from a common ancestor $(T = 0)$, we can express these

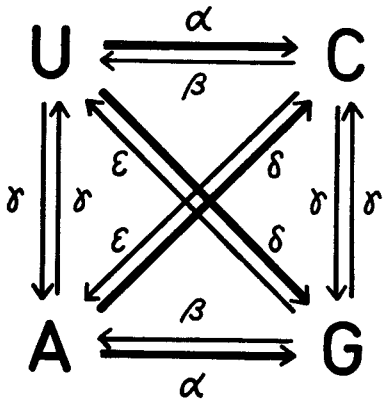

FIGURE 1.—Scheme of base substitutions and their rates per unit time.

probabilities at time $T + \Delta T$, using their probabilities at time T, and the rates of base substitutions, where ΔT stands for a short time interval. Neglecting small quantities involving $(\Delta T)^2$ and higher order terms, we have, for example,

$$\mathrm{U}(T+\Delta T) = \{1-(\alpha+\delta+\gamma)\Delta T\}\mathrm{U}(T) + \beta\Delta T\mathrm{C}(T) + \varepsilon\Delta T\mathrm{C}(T) + \gamma\Delta T\mathrm{A}(T).$$

As seen from Figure 1, the first term in the right-hand side of this equation corresponds to the probability of no change, and the last three terms are the probabilities that U came from the remaining three bases during the short time interval ΔT. The corresponding probabilities for other bases can be obtained in a similar manner, and we get the following set of ordinary differential equations by letting $\Delta T \to 0$;

$$\frac{d\mathrm{U}(T)}{dT} = -(\alpha+\delta+\gamma)\mathrm{U}(T) + \beta\mathrm{C}(T) + \varepsilon\mathrm{G}(T) + \gamma\mathrm{A}(T) ,$$

$$\frac{d\mathrm{A}(T)}{dT} = -(\alpha+\delta+\gamma)\mathrm{A}(T) + \beta\mathrm{G}(T) + \varepsilon\mathrm{C}(T) + \gamma\mathrm{U}(T) ,$$

$$\frac{d\mathrm{C}(T)}{dT} = -(\beta+\varepsilon+\gamma)\mathrm{C}(T) + \alpha\mathrm{U}(T) + \delta\mathrm{A}(T) + \gamma\mathrm{G}(T) ,$$

$$\frac{d\mathrm{G}(T)}{dT} = -(\beta+\varepsilon+\gamma)\mathrm{G}(T) + \alpha\mathrm{A}(T) + \delta\mathrm{U}(T) + \gamma\mathrm{C}(T) .$$

Likewise, we can derive the equations for changes of probabilities of the base

TABLE 1

Types of nucleotide base pairs occurring at homologous sites in two sequences and their probabilities (relative frequencies)

Types	U C A G / U C A G	U C A G / C U G A	U A C G / A U G C	U G A C / G U C A
Probabilities	$\underbrace{S_1 S_2 S_3 S_4}_{S}$	$\underbrace{2P_1\, 2P_2}_{P}$	$\underbrace{2Q_1\, 2Q_2}_{Q}$	$\underbrace{2R_1\, 2R_2}_{R}$

pairs listed in Table 1, although the procedures involved are more complicated. As an example. let us consider the change of base pair UC. Noting that the probability of UC is equal to that of CU, *i.e.*, P_1, we first get the probability of no change occurring in either nucleotide. This is $[1-(\alpha+\gamma+\varepsilon)\Delta T][1-(\beta+\gamma+\delta)\Delta T]$, so that the contribution from this class is $[1-(\alpha+\gamma+\varepsilon)\Delta T]\times[1-(\beta+\gamma+\delta)\Delta T]P_1(T)$. UC is also derived from UU and CC with probabilities $[1-(\alpha+\gamma+\varepsilon)\Delta T]\alpha\Delta T$ and $[1-(\beta+\gamma+\delta)\Delta T]\beta\Delta T$, respectively. Then, the contribution from these classes is $[1-(\alpha+\gamma+\varepsilon)\Delta T]\alpha\Delta TS_1(T)+[1-(\beta+\gamma+\delta)\Delta T]\beta\Delta TS_2(T)$. Additional contributions come from pairs UA and GC with probabilities $[1-(\alpha+\gamma+\varepsilon)\Delta T]\delta\Delta TQ_1(T)$ and $[1-(\beta+\gamma+\delta)\Delta T]\varepsilon\Delta TQ_2(T)$, and also from pairs UG and AC with probabilities $[1-(\alpha+\gamma+\varepsilon)\Delta T]\gamma\Delta TR_1(T)$ and $[1-(\beta+\gamma+\delta)\Delta T]\gamma\Delta TR_2(T)$. Combining all these contributions, and neglecting $(\Delta T)^2$ and higher order terms, as before, we have

$$P_1(T+\Delta T)=[1-(\alpha+\beta+2\gamma+\delta+\varepsilon)\Delta T]P_1(T)+\alpha\Delta TS_1(T)+\beta\Delta TS_2(T)+\delta\Delta TQ_1(T)+\varepsilon\Delta TQ_2(T)+\gamma\Delta T\{R_1(T)+R_2(T)\}.$$

Continuing these calculations for other base pairs and taking the limit $\Delta T\to 0$, we get a complete set of differential equations (equations 1). It is interesting to note that a more convenient derivation of equations (1) is possible if we use the differential equations for $\mathrm{U}(T)$, $\mathrm{A}(T)$, $\mathrm{C}(T)$ and $\mathrm{G}(T)$ and combine them through relationships such as $\frac{dP_1(T)}{dT}=\mathrm{U}(T)\frac{d\mathrm{C}(T)}{dT}+\mathrm{C}(T)\frac{d\mathrm{U}(T)}{dT}$ and $P_1(T)=\mathrm{U}(T)\mathrm{C}(T)$ for the case of the UC pair. This is true because bases in one species change independently from those in other species. We can verify by direct calculation that both derivations give the same set of equations. Thus, we obtain a complete set of differential equations as follows:

$$\left.\begin{aligned}
\frac{dS_1(T)}{dT}&=-2(\alpha+\gamma+\delta)S_1(T)+2\beta P_1(T)+2\gamma Q_1(T)+2\varepsilon R_1(T)\\
\frac{dS_2(T)}{dT}&=-2(\beta+\gamma+\varepsilon)S_2(T)+2\alpha P_1(T)+2\gamma Q_2(T)+2\delta R_2(T)\\
\frac{dS_3(T)}{dT}&=-2(\alpha+\gamma+\delta)S_3(T)+2\beta P_2(T)+2\gamma Q_1(T)+2\varepsilon R_2(T)\\
\frac{dS_4(T)}{dT}&=-2(\beta+\gamma+\varepsilon)S_4(T)+2\alpha P_2(T)+2\gamma Q_2(T)+2\delta R_1(T)\\
\frac{dP_1(T)}{dT}&=-(\alpha+\beta+2\gamma+\delta+\varepsilon)P_1(T)+\alpha S_1(T)+\beta S_2(T)\\
&\quad+\delta Q_1(T)+\varepsilon Q_2(T)+\gamma[R_1(T)+R_2(T)]\\
\frac{dP_2(T)}{dT}&=-(\alpha+\beta+2\gamma+\delta+\varepsilon)P_2(T)+\alpha S_3(T)+\beta S_4(T)\\
&\quad+\delta Q_1(T)+\varepsilon Q_2(T)+\gamma[R_1(T)+R_2(T)]\\
\frac{dQ_1(T)}{dT}&=-2(\alpha+\gamma+\delta)Q_1(T)+\gamma[S_1(T)+S_3(T)]\\
&\quad+\varepsilon[P_1(T)+P_2(T)]+\beta[R_1(T)+R_2(T)]
\end{aligned}\right\}\quad(1)$$

$$\left.\begin{aligned}
\frac{dQ_2(T)}{dT} &= -2(\beta+\gamma+\delta)Q_2(T) + \gamma[S_2(T)+S_4(T)] \\
&\quad + \delta[P_1(T)+P_2(T)] + \alpha[R_1(T)+R_2(T)] \\
\frac{dR_1(T)}{dT} &= -(\alpha+\beta+2\gamma+\delta+\varepsilon)R_1(T) + \delta S_1(T) + \varepsilon S_4(T) \\
&\quad + \gamma[P_1(T)+P_2(T)] + \alpha Q_1(T) + \beta Q_2(T) \\
\frac{dR_2(T)}{dT} &= -(\alpha+\beta+2\gamma+\delta+\varepsilon)R_2(T) + \delta S_3(T) + \varepsilon S_2(T) \\
&\quad + \gamma[P_1(T)+P_2(T)] + \alpha Q_1(T) + \beta Q_2(T).
\end{aligned}\right\}$$

To solve equations (1), we define six variables

$$\left.\begin{aligned}
X_\pm(T) &= S_1(T) + S_3(T) \pm 2Q_1(T) \\
Y_\pm(T) &= S_2(T) + S_4(T) \pm 2Q_2(T) \\
Z_\pm(T) &= P(T) \pm R(T) \ ,
\end{aligned}\right\} \qquad (2)$$

where we take the same sign for the subscripts of X, Y and Z as that in the right-hand side. Then, from (1), we can derive two sets of equations,

$$\frac{d}{dT}\begin{pmatrix} X_+(T) \\ Y_+(T) \\ Z_+(T) \end{pmatrix} = \begin{pmatrix} -2(\alpha+\delta) & 0 & \beta+\varepsilon \\ 0 & -2(\beta+\varepsilon) & \alpha+\delta \\ 2(\alpha+\delta) & 2(\beta+\varepsilon) & -(\alpha+\beta+\delta+\varepsilon) \end{pmatrix}\begin{pmatrix} X_+(T) \\ Y_+(T) \\ Z_+(T) \end{pmatrix} \qquad (3)$$

and

$$\frac{d}{dT}\begin{pmatrix} X_-(T) \\ Y_-(T) \\ Z_-(T) \end{pmatrix} = \begin{pmatrix} -2(\alpha+2\gamma+\delta) & 0 & \beta-\varepsilon \\ 0 & -2(\beta+2\gamma+\varepsilon) & \alpha-\delta \\ 2(\alpha-\delta) & 2(\beta-\varepsilon) & -(\alpha+\beta+4\gamma+\delta+\varepsilon) \end{pmatrix} \times \begin{pmatrix} X_-(T) \\ Y_-(T) \\ Z_-(T) \end{pmatrix} . \qquad (4)$$

In these equations, the transformation matrices have a common form

$$M = \begin{pmatrix} -2c & 0 & b \\ 0 & -2d & a \\ 2a & 2b & -(c+d) \end{pmatrix} . \qquad (5)$$

Note that $a = c$ and $b = d$ hold in (3). As we can easily calculate the eigenvalues and projection operators for the matrix of (5), we can solve the initial value problems of equations (3) and (4). Let λ_i's ($(i = 1, 2$ and $3)$ be the eigenvalues and p_i's be the corresponding projection operators. Then, we have

$$\lambda_1 = -(c+d), \lambda_2 = -(c+d-g) \text{ and } \lambda_3 = -(c+d+g), \qquad (6)$$

and

$$p_1 = \frac{1}{g^2}\begin{pmatrix} 2ab & -2b^2 & -b(d-c) \\ -2a^2 & 2ab & a(d-c) \\ -2a(d-c) & 2b(d-c) & (d-c)^2 \end{pmatrix}$$

$$\left.\begin{aligned} p_2 &= \frac{1}{2g^2}\begin{pmatrix} (d-c)(d-c+g)+2ab & 2b^2 & b(d-c+g) \\ 2a^2 & (d-c)(d-c-g)+2ab & -a(d-c-g) \\ 2a(d-c+g) & 2b(-d+c+g) & 4ab \end{pmatrix} \\ p_3 &= \frac{1}{2g^2}\begin{pmatrix} (d-c)(d-c-g)+2ab & 2b^2 & b(d-c-g) \\ 2a^2 & (d-c)(d-c+g)+2ab & -a(d-c+g) \\ 2a(d-c+g) & 2b(d-c-g) & 4ab \end{pmatrix} \end{aligned}\right\} \quad (7)$$

where $g = \sqrt{(d-c)^2+4ab}$.

By using these formulae, we obtain the solutions $X(T)$ at time T under an arbitrary initial condition of $\mathbf{X}(0)$. Thus,

$$\mathbf{X}(T) = \{e^{\lambda_1 T}P_1 + e^{\lambda_2 T}P_2 + e^{\lambda_3 T}p_3\}\mathbf{X}(0) \quad (8)$$

where $\mathbf{X}(\cdot)$ is a column vector that can be either $(X_+, Y_+, Z_+)^t$ or $(X_-, Y_-, Z_-)^t$, in which the superscript t denotes the transpose.

We assume that the frequency (ω) of U + A does not change with time, and also that $\mathrm{U}(T) = \mathrm{A}(T)$ and $\mathrm{C}(T) = \mathrm{G}(T)$ for all T. Then, $S_i(T)$, $P_i(T)$, $Q_i(T)$ and $R_i(T)$ can all be expressed in terms of $X_\pm(T)$, $Y_\pm(T)$ and $Z_\pm(T)$. The evolutionary rate of base substitutions per unit time is

$$k = (\alpha+\gamma+\delta)\omega + (\beta+\gamma+\varepsilon)(1-\omega)\,, \quad (9)$$

or

$$k = \gamma + 2\omega(1-\omega)(\alpha+\beta+\delta+\varepsilon)\,. \quad (9a)$$

These equations are derived from the consideration that U or A each with the frequency $\omega/2$ changes to the other bases at the rate of $\alpha+\gamma+\delta$, and C or G each with $(1-\omega)/2$ changes at the rate of $\beta+\gamma+\varepsilon$. Note that we have $\omega = (\beta+\varepsilon)/(\alpha+\beta+\delta+\varepsilon)$. Therefore, the expected number of substitutions per site between two species with divergence time T is given by

$$K = 2Tk\,.$$

If we use formula (9a) for k, then

$$K = 2\gamma T + 4\omega(1-\omega)(\alpha+\beta+\delta+\varepsilon)T\,. \quad (10)$$

Before deriving an expression for K in terms of $X_\pm$, $Y_\pm$ and $Z_\pm$, we shall obtain the explicit expression for the eigenvalues and the functional forms of those quantities under the assumption of the steady state of U + A content. As the initial conditions are now

$$\mathbf{X}_\pm(0) = (\omega, 1-\omega, 0)^t \quad (11)$$

for both cases, the solutions for $\mathbf{X}_+(T)$ are expressed in a simple form,

$$\left.\begin{aligned} X_+(T) &= \omega\{\omega + (1-\omega)e^{\lambda_0 T}\} \\ Y_+(T) &= (1-\omega)(1-\omega+\omega e^{\lambda_0 T}) \\ Z_+(T) &= 2\omega(1-\omega)(1-e^{\lambda_0 T}) \end{aligned}\right\} \quad (12)$$

where

$$\lambda_0 = -2(\alpha+\beta+\delta+\varepsilon)\ . \tag{13}$$

On the other hand, the eigenvalues of equation (4) are

$$\begin{aligned}\lambda_1 &= -(\alpha+\beta+\delta+\varepsilon+4\gamma)\\ \lambda_2 &= \lambda_1 + g\\ \lambda_3 &= \lambda_1 - g\end{aligned} \tag{14}$$

where $g^2 = \{\alpha+\delta-(\beta+\varepsilon)\}^2 + 4(\alpha-\delta)(\beta-\varepsilon)$. Using (7) and (14), the solutions for $\mathbf{X}_-(T)$ are, for $g \neq 0$,

$$\left.\begin{aligned}X_-(T) &= \frac{1}{g^2}[2b\{a\omega - b(1-\omega)\}e^{\lambda_1 T} + \{\xi\omega + b^2(1-\omega)\}e^{\lambda_2 T}\\ &\qquad + \{\eta\omega + b^2(1-\omega)\}e^{\lambda_3 T}]\\ Y_-(T) &= \frac{1}{g^2}[-2a\{a\omega - b(1-\omega)\}e^{\lambda_1 T} + \{a^2\omega + \eta(1-\omega)\}e^{\lambda_2 T}\\ &\qquad + \{a^2\omega + \xi(1-\omega)\}e^{\lambda_3 T}]\\ Z_-(T) &= \frac{1}{g^2}[-2(d-c)\{a\omega - b(1-\omega)\}e^{\lambda_1 T} + \{a(d-c+g)\omega\\ &\qquad - b(d-c-g)(1-\omega)\}e^{\lambda_2 T} + \{a(d-c-g)\omega\\ &\qquad - b(d-c+g)(1-\omega)\}e^{\lambda_3 T}]\end{aligned}\right\} \tag{15}$$

in which $a = \alpha - \delta$, $b = \beta - \varepsilon$, $c = \alpha + \delta + 2\gamma$, $d = \beta + \varepsilon + 2\gamma$, $\xi = \frac{1}{2}(d-c)(d-c+g) + ab$ and $\eta = \frac{1}{2}(d-c)(d-c-g) + ab$. In this case, however, it does not seem feasible to derive a simple formula for $2\lambda_1 T = (\lambda_2+\lambda_3)T = -2(\alpha+\beta+\delta+\varepsilon+4\gamma)$ as a function of $\mathbf{X}_-(T)$ and ω.

A great simplification is possible if we assume that $\delta = \theta\alpha$ and $\varepsilon = \theta\beta$, where θ is a constant. Then, equations (15) are much simplified (see 15a below), although equations (12) remain the same. Furthermore, $2\gamma T$ in (10) can be expressed in terms of $X_-(T)$, $Y_-(T)$, $Z_-(T)$ and ω, which can be estimated from observations. The solutions for $\mathbf{X}_-(T)$ become, for $\alpha \neq \beta$ and $\theta \neq 1$,

$$\begin{aligned}X_-(T) &= \frac{\omega}{2g}\{(d-c+g)e^{\lambda_2 T} - (d-c-g)e^{\lambda_3 T}\}\\ Y_-(T) &= \frac{1-\omega}{2g}\{-(d-c-g)e^{\lambda_2 T} + (d-c+g)e^{\lambda_3 T}\}\\ Z_-(T) &= \frac{a\omega + b(1-\omega)}{g}(e^{\lambda_2 T} - e^{\lambda_3 T})\ ,\end{aligned} \tag{15a}$$

where $\omega = \dfrac{\beta}{\alpha+\beta}$. Note that, under the above assumption, if $\alpha = \beta$, the model reduces to the "three-substitution-type" (3ST) model of KIMURA (1981). Using equations (15a), we get

$$X_-(T)Y_-(T) - \left(\frac{Z_-(T)}{2}\right)^2 = \omega(1-\omega)e^{(\lambda_2+\lambda_3)T}\ . \tag{16}$$

Combining this with equations (12) and (13), we get

$$\gamma T = -\frac{1}{8}\ln\left\{\frac{X_-(T)Y_-(T) - \left(\frac{Z_-(T)}{2}\right)^2}{\omega(1-\omega) - \left(\frac{Z_+(T)}{2}\right)}\right\}, \tag{17}$$

and therefore we obtain an appropriate equation for K as follows.

$$K = -\frac{1}{4}\ln\left[\left\{\frac{X_-(T)Y_-(T) - \left(\frac{Z_-(T)}{2}\right)^2}{\omega(1-\omega)}\right\}\left\{1 - \frac{Z_+(T)}{2\omega(1-\omega)}\right\}^{8\omega(1-\omega)-1}\right], \tag{18}$$

or more explicitly

$$K = -\frac{1}{4}\ln\left[\left\{\frac{(S_1+S_3+2Q_1)(S_2+S_4-2Q_2) - \left(\frac{P-R}{2}\right)^2}{\omega(1-\omega)}\right\} \times \left\{1 - \frac{P+R}{2\omega(1-\omega)}\right\}^{8\omega(1-\omega)-1}\right], \tag{18a}$$

where ω is the fraction of the sum of two bases U and A, and S_1, etc., are as defined in Table 1. In addition, we can estimate the unknown parameter θ by using the relationship

$$\frac{1-\theta}{1+\theta} = \frac{\left(\omega - \frac{1}{2}\right)(P-R)}{(1-\omega)(S_1+S_3-2Q_1) - \omega(S_2+S_4-2Q_2)}, \tag{19}$$

while the ratio of β to α can be determined from the assumption that ω does not change with time and that

$$\omega = \frac{\beta}{\alpha+\beta} = S_1 + S_3 + 2Q_1 + \frac{1}{2}(P+R). \tag{20}$$

Formula (19) may be verified by substituting equations (15a) for the right-hand side of equation (19), noting at the same time equations (2). Estimated evolutionary distances (denoted by $\tilde{K}$) for several comparisons are shown in Table 2 together with values of θ and ω. The table also contains the estimates of the standard error of $\tilde{K}$ that were obtained using a procedure similar to the one used by KIMURA (1980, 1981) but assuming that the estimation of ω is not accompanied by sampling error. It is also based on the assumption of a multinomial distribution of the variables in the right-hand side of (18). This seems to give a good estimate in the light of the results of Monte Carlo experiments.

MONTE CARLO EXPERIMENTS

In order to check the validity and the range of applicability of our formula (18), we performed Monte Carlo experiments. The procedures were as follows.

TABLE 2

Evolutionary distances per nucleotide site at the first, second and third codon position estimated by using the 3ST and the present model (TK)

Comparison			Evolutionary distances per nucleotide site $\tilde{K}_1$	$\tilde{K}_2$	$\tilde{K}_3$
	(3ST)		0.300	0.195	0.636
	(TK)		0.299	0.237	0.691
			± 0.044	± 0.037	± 0.160
Chicken β *vs.* Rabbit β					
		ω	0.384	0.637	0.267
		θ	0.489	1.048	— 0.079
	(3ST)		0.265	0.177	0.531
	(TK)		0.269	0.178	0.692
			± 0.038	± 0.032	± 0.121
Human *vs.* Rat pregrowth hormone					
		ω	0.430	0.609	0.258
		θ	0.592	0.135	0.468
	(3ST)		0.182	0.274	0.947
	(TK)		0.178	0.289	∞
			± 0.273	± 0.104	
Human *vs.* Rat insulin C peptide					
		ω	0.081	0.581	0.226
		θ	2.63	0.821	1.00
	(3ST)		0.040	0.000	0.461
	(TK)		0.042	0.000	0.786
			± 0.026	± 0.019	± 0.300
Human *vs.* Rat insulin A + B peptide					
		ω	0.480	0.628	0.275
		θ	0.964	1.00	0.535
	(3ST)		0.160	0.127	0.427
	(TK)		0.171	0.135	0.463
			± 0.031	± 0.030	± 0.080
Rabbit β *vs.* Mouse β					
		ω	0.373	0.644	0.339
		θ	0.954	0.757	0.398
	(3ST)		0.600	0.437	0.903
	(TK)		0.600	0.517	1.29
			± 0.087	± 0.062	± 1.05
Rabbit α *vs.* Rabbit β					
		ω	0.389	0.633	0.234
		θ	0.071	1.09	0.290
	(3ST)		0.124	0.115	0.544
	(TK)		0.126	0.131	0.774
			± 0.029	± 0.027	± 0.205

TABLE 2—Continued

Comparison			Evolutionary distances per nucleotide site $\tilde{K}_1$	$\tilde{K}_2$	$\tilde{K}_3$
Mouse α-1 *vs.* Rabbit α (Exons 1, 2 and 3)					
		ω	0.390	0.575	0.231
		θ	0.539	0.966	0.182
	(3ST)		0.008	0.008	0.470
	(TK)		0.008	0.008	0.470
			± 0.013	± 0.008	± 0.088
S. purpuratus vs. P. miliaris H3					
		ω	0.380	0.493	0.530
		θ	0.000	0.000	— 0.160
	(3ST)		0.086	0.020	0.479
	(TK)		0.087	0.020	0.484
			± 0.032	± 0.016	± 0.104
S. purpuratus vs. P. miliaris H2B					
		ω	0.303	0.436	0.535
		θ	1.00	0.000	0.536
	(3ST)		0.307	0.371	0.768
	(TK)		0.303	0.374	0.862
			± 0.062	± 0.060	± 0.259
Mouse $\psi\alpha$ *vs.* Rabbit α					
		ω	0.381	0.587	0.258
		θ	— 0.458	0.867	0.115
	(3ST)		0.133	0.121	0.658
	(TK)		0.129	0.138	0.883
Mouse α-1 *vs.* Rabbit α					
		ω	0.383	0.576	0.232
		θ	— 9.22	0.952	0.009
	(3ST)		0.209	0.300	0.337
	(TK)		0.207	0.294	0.358
			± 0.045	± 0.114	± 0.074
Mouse α-1 *vs.* Mouse $\psi\alpha$					
		ω	0.352	0.564	0.376
		θ	0.527	2.57	1.18

First, we prepared a random nucleotide sequence to be used as a common ancestor. It consisted of n sites, with frequencies of U, A, C and G being given by $\omega/2$, $\omega/2$, $(1-\omega)/2$ and $(1-\omega)/2$. From this sequence, two descendent sequences were derived by independent nucleotide substitutions according to the scheme shown in the previous section. In simulation experiments, we assigned values of substitutional rates so that k in equation (9) is of the order of 10^{-3}. The experiments were continued until one substitution per site had occurred on the average in each lineage. We compared the two sequences through time and counted the actual number of nucleotide substitutions involved in two lineages. The total

number of nucleotide substitutions, K, was monitored by summing the actual numbers of substitutions observed until a given time T. On the other hand, the expected number of nucleotide substitutions over T, as denoted by K_E, was calculated by $2kT$ for a given value of k. Note that, strictly speaking, K_E is different from K, although no significant differences between these two were observed in the simulations experiments. We also observed the relative frequencies of the various classes in Table 1 at specified times and calculated the estimtaed evolutionary distances using equation (18) or (18a). In a similar way, we obtained the estimate $\tilde{K}_{JC} = -\frac{3}{4}\ln(1-\frac{4}{3}\lambda)$ as a reference point; where λ is the fraction of different sites between two nucleotide sequences (see JUKES and CANTOR 1969; KIMURA and OHTA 1972). These processes were repeated 100 times, assuming $n = 100$. Each quantity of concern was obtained by taking the average.

The results are illustrated through Figures 2 to 4. As typical situations, we assigned the parameter values similar to the ones derived from comparisons of DNA sequences of exons 1 and 2 in the mouse and the rabbit α-globin genes, together with a mouse pseudogene. The parameters are determined separately at different codon positions. Figures 2, 3 and 4, respectively, represent situations at

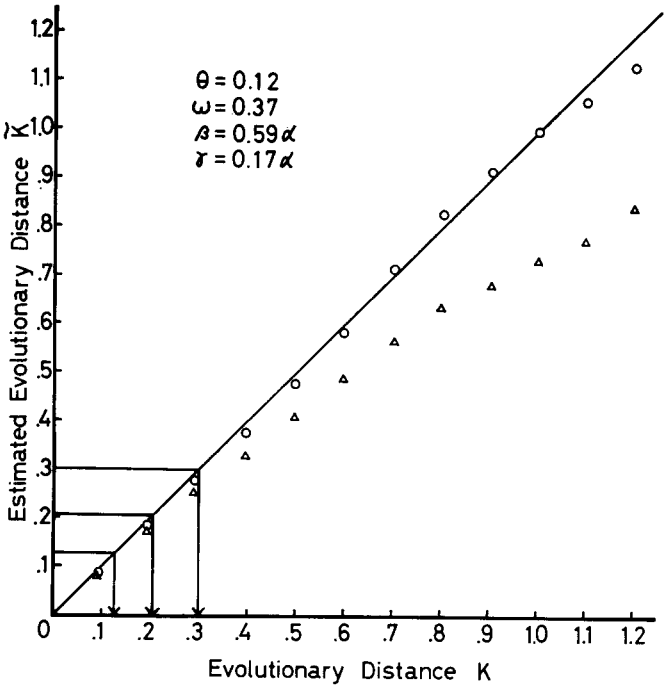

FIGURE 2 to 4.—Relationship between the actual evolutionary distance K and the estimated evolutionary distance $\tilde{K}$ based on the formula (18) (the broken lines). The solid lines represent $K = \tilde{K}$. The parameters used in these figures are determined by taking the average for the data on nucleotide sequences of the α globin gene and the α pseudogene ($\alpha 3$) of the mouse, and the α-globin gene of the rabbit. FIGURE 2: First codon position.

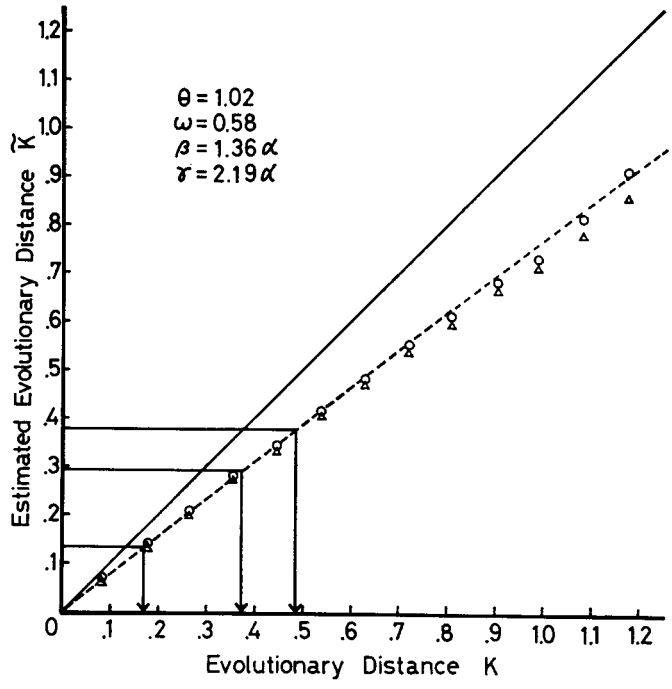

FIGURE 3.—Second codon position.

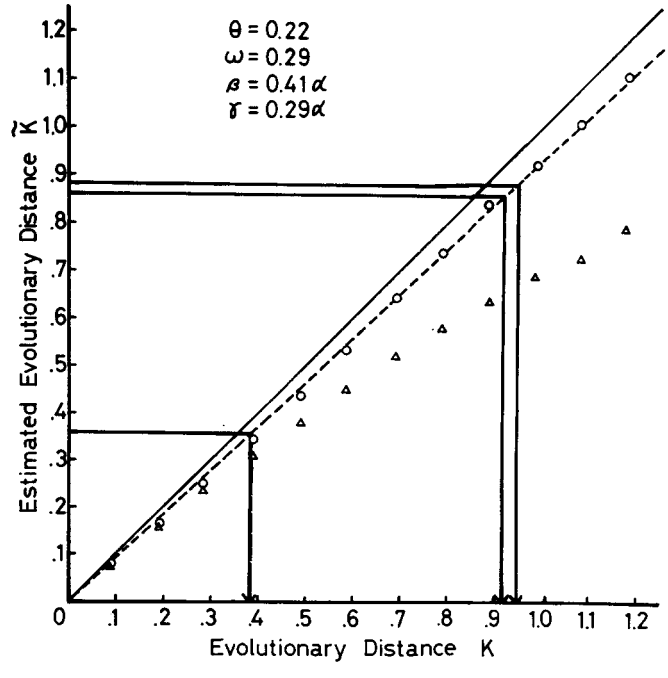

FIGURE 4.—Third codon position.

the first, second and third positions of codons within the above two exons. The abscissa stands for the actual evolutionary distance K (in terms of the number of base substitutions), while the ordinate represents the estimated evolutionary distance $\tilde{K}$. Because the difference between K and K_E turned out to be quite small in all experiments, we will not discriminate between them. Estimates obtained by using equation (18) are marked by open circles, while those estimated by using the fomula of JUKES and CANTOR (1969) are indicated open triangles. It is evident from Figure 3 that both sets of estimates are close to each other when θ is near 1 and ω is about 0.5 and that they need appropriate corrections when K is large. Such a discrepancy for the second codon position seems to have been caused by the relatively high transversion type substitution rate γ assumed. On the other hand, when θ and ω are, respectively, less than 1 and 0.5, marked difference between the two sets of estimators occur (see Figures 2 and 4). Fortunately, equation (18) provides good estimates of the actual evolutionary distance if K does not exceed unity, and the linearity is almost completely preserved in this range of K values. The simulation experiments show that the present formula (18) is useful, especially when the rates of transversion-type substitutions are low and the frequencies of two classes of bases differ greatly from each other. Although not perfectly proven, if the real pattern of base substitutions is something like this, the present method is the most accurate, followed by the 3ST (KIMURA 1981) and JUKES and CANTOR methods. On the other hand, if substitution rates are equal in all directions, all three methods of these give almost the same result.

Note that cases arise in which we cannot estimate the evolutionary distances from these formulae. This occurs when arguments of logarithms become zero or negative. Such situations should occur frequently when a great many substitutions are involved. Because we excluded such inapplicable cases from our calculations, the estimated evolutionary distance $\tilde{K}$ gives an underestimate for K. In some cases, the fraction of such inapplicable cases for formula (18) became more than 50% if $K \geqq 1$. When we analyze actual data, this difficulty often arises if we compare the sequences in which many substitutions have occurred. One example is afforded by the third position of codons in the C peptide of insulin when comparison is made between human and rat (see Table 2). Such a problem arises because the true nature of substitutional processes is stochastic; whereas, our present treatment is deterministic. As the number of nucleotide sites actually compared is finite, sometimes less than 100, the small sample size creates a large sampling error. Particularly, when $K > 2$, no estimator seems to provide good information on the actual evolutionary distance, as simulation experiments show. We also note that in such cases $\tilde{K}$ has a very large error variance.

RESULTS AND DISCUSSION

Using equation (18), we calculated the evolutionary distances between various DNA sequences, as shown in Table 2. This table also contains evolutionary distances estimated using the "three-substitution-type" (3ST) model of KIMURA (1981) for the purpose of comparison (this corresponds to the case where $\alpha = \beta$

and $\varepsilon = \delta$ in Figure 1). It can be seen from this table, that equation (18) often provides larger estimates than the corresponding estimates obtained from the 3ST model. Particularly, when the estimated value of θ from equation (19) is small and that of ω from (20) differs noticeably from 0.5, the discrepancy between the two sets of estimates becomes large. Such a dependence of equation (18) on ω seems to represent a favorable property as an estimator of evolutionary distance because, as pointed out by Kimura (1981), the U and A content, particularly at the third position of codons, is often much less than 0.5. For instance, the value of ω is about 0.23 at the third codon positions in rabbit α and β globins. This bias often results in unreliable estimates of the evolutionary distance, as we mentioned before.

On the other hand, the value of θ is quite sensitive to changes of observed values of various classes involved in equation (19). In most cases, θ is less than unity, but, in some cases, it happens to be negative or exceed unity. The estimated value of θ for each case may not be reliable, but the results suggest that transition-type substitutions can occur more frequently than those of transversion types (*i.e.*, from U to G and A to C, or *vice versa*). It is fortunate that equation (18) does not contain θ.

It may be clear from Table 2 that evolutionary base substitution is faster (roughly 2.5 ~ 5.9 times) at the third position of codons than at the second positions in the functional globin genes. This characteristic is particularly conspicuous in histone genes. The ratio per site of the rates of third to the second positions is about 59 for the comparison of *S. purpuratus* and *P. miliaris* H3 sequences, while it is about 24 for the H2B sequences in the same comparison (for data, see SURES, LOWRY and KEDES 1978; SCHAFFNER *et al.* 1978). These species probably diverged between 6×10^7 and 16×10^7 years ago (DURHAM 1966; KEDES 1979); therefore, the rate k_3 per site per year is $(1.5 \sim 3.9) \times 10^{-9}$ for the H3 sequences and $(1.5 \sim 4.0) \times 10^{-9}$ for the H2B sequences. These values are very similar; moreover, the rates k_3 estimated for other genes using several other comparisons show roughly the same values. For example, we have 4.3×10^{-9} for the human and rat pregrowth hormone comparison (with the divergence time $T = 8 \times 10^7$ years) and 1.2×10^{-9} for the chicken and rabbit β globin comparison ($T = 3 \times 10^8$ years). In some cases, formula (18) gives 1.4 to 1.7 times higher estimates for k_3 than does the 3ST model, but whether or not the model can decrease the estimated variance of k_3 is still uncertain until more data are available. The rough equality of the evolutionary rates at the third position of codons among genes is in sharp contrast to wide differences of the rate at the second position, where most substitutions alter amino acids. At any rate, we can confirm the conclusion of KIMURA (1980, 1981) and MIYATA, YASUNAGA and NISHIDA (1980) that the rates of nucleotide substitutions at the third position of codons are not only very high but also roughly equal to each other between genes even when amino acid altering substitutional rates are quite different.

Now, let us examine the history of the α globin pseudogene in the mouse, as studied by VANIN *et al.* (1980) and NISHIOKA, LEDER and LEDER (1980). We apply equation (18) to estimate evolutionary distances. The sequences of the

normal α-globin genes, including the noncoding regions, have also been determined in the mouse and the rabbit (see KONKEL, MAIZEL and LEDER 1979; HARDISON *et al.* 1979, and references therein). Thus, we can compare DNA sequences of these three genes. Our aim is to estimate the time of occurrence of duplication leading to the α pseudogene in the mouse line and the relative evolutionary rates at each codon position in the pseudogene relative to those in the normal α-globin genes. Although several models concerning the appearance of pseudogenes are conceivable (see for example PROUDFOOT and MANIATIS 1980; MIYATA and YASUNAGA 1981; LI, personal communication), we assume here a simple one.

Let us assume that the duplication occurred T_d years ago, and thereafter a duplicated gene became "dead" and started to evolve at the rate k'_i instead of k_i, where i (= 1, 2 or 3) denotes the codon position. At the incipient stage, the mouse population must have been polymorphic with respect to the number of α-globin genes per individual. However, it is likely that the duplicate gene could accumulate mutations at a higher rate than the normal gene, due to its multiplicity. Let T_0 be the divergence time of the mouse and the rabbit, and let $\tilde{K}_i(X - Y)$ be the evolutionary distance in the ith codon position of homologous genes between species X and Y. For example, $\tilde{K}_i(M\psi\alpha - R\alpha)$ denotes the evolutionary distance in the ith position between the mouse α pseudogene and the rabbit α gene. In the following study, we make no correction for $\tilde{K}$ and compare only the part including exon 1 and 2, excluding the exon 3 region from the calculation because of an unusual characteristic of this region, as pointed out by MIYATA and YASUNAGA (1981). Then, T_d and k_i', relative to T_0 and k_i, can be calculated for each codon position by

$$\frac{T_d}{T_0} = \frac{\tilde{K}_i(M\alpha - M\psi\alpha) - \tilde{K}_i(M\psi\alpha - R\alpha) + \tilde{K}_i(M\alpha - R\alpha)}{\tilde{K}_i(M\alpha - R\alpha)}$$

and

$$\frac{k_i'}{k_i} = \frac{\tilde{K}_i(M\alpha - M\psi\alpha) + \tilde{K}_i(M\psi\alpha - R\alpha) - \tilde{K}_i(M\alpha - R\alpha)}{\tilde{K}_i(M\alpha - M\psi\alpha) - \tilde{K}_i(M\psi\alpha - R\alpha) + \tilde{K}_i(M\alpha - R\alpha)} .$$

Substituting the values of Table 2 in the above equations, the ratios of T_d/T_0 are respectively, 0.26, 0.42 and 0.43 for the first, second and third codon positions, while $k_1'/k_1 = 11.5$, $k_2'/k_2 = 13.9$ and $k_3'/k_3 = 0.9$. Roughly speaking, this means that the duplication responsible for the mouse α pseudogene occurred about $(0.3 \sim 0.4)T_0$ years ago. If we take 8.0×10^7 as T_0, T_d becomes about $20 \sim 30$ million years. On the other hand, the rates in the first and second positions in the pseudogene turn out to be roughly 10 times faster than those of normal genes; whereas, the rate in the third positions remain unaltered. The estimated values of $2k_1'T_0$ (= 1.48) and $2k_2'T_0$ (= 1.92) are both about 2 times greater than the estimated value of $2k_3T_0$ (= 0.883). This might indicate that there are some selective constraints even against the changes in the third positions in the normal gene. Another possibility is that equation (18) still gives an underestimate for the evolu-

tionary distance. Considering the fact that estimated values of k_3 and k_3' are similar, the latter might be more probable in the light of the results of Monte Carlo experiments. We could make some correction for the values of $\tilde{K}$ based on Monte Carlo experiments, but we did not take such an approach here because it seemed unlikely that we can get more precise estimates of these values because of the inevitable large sampling errors.

In the above analysis, we have tacitly assumed that the α pseudogene is fixed in the mouse population. In fact, it is likely that several million years are sufficient for such a nonfunctional pseudogene to become fixed in a population (see MARUYAMA and TAKAHATA 1981; TAKAHATA 1981). Therefore, it is highly probable that the α pseudogene is fixed in the mouse population. This conclusion, however, is tentative in the sense that we ignored the effect of recurrent unequal crossing over. As pointed out by OHTA (1981), it is possible that unequal crossing over plays a prominent role in the evolution of duplicate genes, even when a small number of them are tightly linked (*i.e.*, multigene family of small size). If unequal crossing over occurs frequently in the course of evolution, the fixation of a pseudogene at a specific locus may be considered transient. However, a preliminary study incorporating such a mechanism still supports the view that all individuals carry a pseudogene in their genome for several million years, although its location on a chromosome may vary from individual to individual or in time, (the details will be published elsewhere).

We conclude that a duplicate gene leading to the α pseudogene in the mouse line was introduced $20 \sim 30$ million years ago by unequal crossing over and became fixed in the population several million years after the duplication occurred, and that many nucleotide substitutions have accumulated at a high rate, irrespective of codon position, due to the loss of selective constraints.

We thank K. AOKI for his helpful comments in composing the manuscript and T. OHTA for stimulating discussion.

LITERATURE CITED

DURHAM, J. W., 1966 *Echinoides*. pp. 270–295. In: *Treatise on Invetebrate Paleontology, Part U, Echinodermata 3*. Edited by R. C. MOORE. Univ. Kansas Press, Lawrence, Kansas.

FITCH, W. M. and E. MARGOLIASH, 1967 Construction of phylogenetic trees. Science **155:** 279–284.

HARDISON, R. C., E. T. BUTLER III, E. LACY, T. MANIATIS, N. ROSENTHAL and A. EFSTRATIDIS, 1979 The structure and transcription of four linked rabbit β-like globin genes. Cell **18:** 1285–1297.

HOLMQUIST, R., 1980 Evolutionary analysis of α and β hemoglobin genes by REH theory under the assumption of equiprobability of genetic events. J. Mol. Evol. **15:** 149–159.

HOLMQUIST, R. and D. PEARL, 1980 Theoretical foundations for quantitative paleogenetics. III. The molecular divergence of nucleic acids and proteins for the case of genetic events of unequal probability. J. Mol. Evol. **16:** 211–267.

JUKES, T. H. and C. H. CANTOR, 1969 Evolution of protein molecules. pp. 21–123. *Mammalian Protein Metabolism*. Edited by H. N. MUNRO. Academic Press, New York.

KEDES, L. H., 1979 Histone genes and histone messengers. Ann. Rev. Biochem. **48:** 837–870.

KIMURA, M., 1968 Evolutionary rate at the molecular level. Nature **217**: 624–626. ——, 1979 The neutral theory of molecular evolution. Scientific American **241**(5): 98–126. ——, 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**: 111–120. ——, 1981 On estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. U.S. **78**: 454–458.

KIMURA, M. and T. OHTA, 1972 On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol. **2**: 87–90.

KONKEL, D. A., J. V. MAIZEL, JR. and P. LEDER, 1979 The evolution and sequence comparison of two mouse chromosomal β-globin genes. Cell **18**: 865–873.

MARUYAMA, T. and N. TAKAHATA, 1981 Numerical studies of the frequency trajectories in the process of fixation of null genes at duplicated loci. Heredity **46**: 49–57.

MAXAM, A. and W. GILBERT, 1977 A new method for sequencing DNA. Proc. Natl. Acad. Sci. U.S. **74**: 560–564.

MIYATA, T. and T. YASUNAGA, 1980 Molecular evolution of mRNA: A method for estimating evolutionary rates of synonymous amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**: 23–36. ——, 1981 Rapid evolving mouse alpha globin-related pseudogene and its evolutionary history. Proc. Natl. Acad. Sci. U.S. **78**: 450–453.

MIYATA, T., T. YASUNAGA and T. NISHIDA, 1980 Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc. Natl. Acad. Sci. U.S. **77**: 7328–7332.

NEI, M., 1975 *Molecular Population Genetics and Evolution.* North-Holland Publishing Company: Amsterdam, Oxford.

NISHIOKA, Y., A. LEDER and P. LEDER, 1980 Unusual α-globin-like gene that has clearly lost both globin intervening sequences. Proc. Natl. Acad. Sci. U.S. **77**: 2806–2809.

OHTA, T., 1981 Genetic variation in small multigene families. Genet. Research **37**: 133–149.

OHTA, T. and M. KIMURA, 1971 On the constancy of the evolutionary rate of cistrons. J. Mol. Evol. **1**: 18–25.

PROUDFOOT, N. J. and T. MANIATIS, 1980 The structure of a human α-globin pseudogene and its relationship to α-globin gene duplication. Cell **21**: 537–544.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S. **74**: 4563–4567.

SCHAFFNER, W., G. KUNZ, H. DAETWYLER, J. TELFORD, H. O. SMITH and M. L. BIRNSTIEL, 1978 Genes and spacers of cloned sea urchin histone DNA analyzed by sequencing. Cell **14**: 655–671.

SURES, I., J. LOWRY and L. H. KEDES, 1978 The DNA sequence of sea urchin (*S. pupuratus*) H2A, H2B and H3 histone coding and spacer regions. Cell **15**: 1033–1044.

TAKAHATA, N., 1981 On the disappearance of duplicate gene expression. In: *Molecular Evolution, Protein Polymorphism and The Neutral Theory.* Edited by M. KIMURA. Japan Scientific Societies Press, Tokyo. (In press).

VANIN, E. F., G. I. GOLDBERG, P. W. TUCKER and O. SMITHIES, 1980 A mouse α-globin-related pseudogene lacking intervening sequences. Nature **286**: 222–226.

ZUCKERKANDL, E. and L. PAULING, 1965 Evolutionary divergence and convergence in proteins. pp. 97–166. In: *Evolving Genes and Proteins.* Edited by V. BRYSON and H. J. VOGEL. Academic Press: New York and London.

Corrsponding editor: M. NEI