Proc. Natl. Acad. Sci. USA
Vol. 78, No. 1, pp. 454–458, January 1981
Genetics

Estimation of evolutionary distances between homologous nucleotide sequences

(molecular evolution/comparison of base sequences/base substitution rate/neutral mutation-random drift hypothesis)

MOTOO KIMURA

National Institute of Genetics, Mishima, 411 Japan

Contributed by Motoo Kimura, October 3, 1980

ABSTRACT By using two models of evolutionary base substitutions—"three-substitution-type" and "two-frequency-class" models—some formulae are derived which permit a simple estimation of the evolutionary distances (and also the evolutionary rates when the divergence times are known) through comparative studies of DNA (and RNA) sequences. These formulae are applied to estimate the base substitution rates at the first, second, and third positions of codons in genes for presomatotropins, preproinsulins, and α- and β-globins (using comparisons involving mammals). Also, formulae for estimating the synonymous component (at the third codon position) and the standard errors are obtained. It is pointed out that the rates of synonymous base substitutions not only are very high but also are roughly equal to each other between genes even when amino acid-altering substitution rates are quite different and that this is consistent with the neutral mutation-random drift hypothesis of molecular evolution.

Data on nucleotide sequences of various parts of the genome in diverse organisms are appearing at an accelerated, almost explosive, rate. Many of these sequences are of interest for studies of molecular evolution. Before long, comparative studies of amino acid sequences, which have played a major role during the last 15 years or so, will be superseded by studies of nucleotide sequences. Already it has become increasingly evident that a preponderance of synonymous and other silent base substitutions is a general but remarkable feature of molecular evolution and that this is consistent with the neutral theory of molecular evolution (1–8).

In estimating the evolutionary distances between homologous sequences in terms of the number of base substitutions, corrections for multiple and revertant changes at homologous sites are essential. This is because only four kinds of bases exist in nucleotide sequences and even two random sequences show a 25% average match at individual sites. In this paper, I derive some formulae which are useful for estimating evolutionary distances between nucleotide sequences by using two models of evolutionary base substitutions.

THREE-SUBSTITUTION-TYPE (3ST) MODEL

Consider a pair of homologous sites in two sequences being compared. We investigate how these sites have diverged from each other during their descent from a common ancestor T years back. At each individual site, bases are successively substituted one after another in the course of time. To formulate this, we assume a model of evolutionary base substitutions as shown in Fig. 1*a*. Throughout this paper we use RNA codes so that the four bases are expressed by letters U, C, A, and G. Let α, β, and γ be the rates ot base substitutions as indicated in Fig. 1*a*, α being the rate of "transition" type substitutions and β and γ being rates of "transversion" type substitutions. The total rate of base substitutions per unit time (year) is $k = \alpha + \beta + \gamma$.

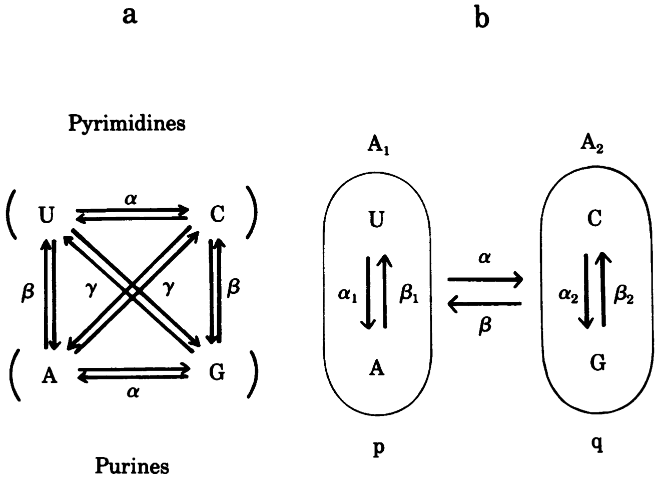

FIG. 1. Two models of evolutionary base substitutions: (*a*) "three-substitution-type" (3ST) model; (*b*) "two-frequency-class" (2FC) model.

It is important to note here that α, β, and γ refer to evolutionary rates by which bases are substituted in the species rather than ordinary mutation rates at the level of individuals. The total number of base substitutions per site which separate the two sequences and therefore involve two branches each with length T is given by $2Tk$ which we denote by K.

$$K = 2Tk = 2(\alpha + \beta + \gamma)T. \qquad [1]$$

When we compare homologous sites of sequences 1 and 2, we note that there are 12 combinations of different bases (Table 1). Let P be the probability (relative frequency) that, at time T, homologous sites are occupied by base pair UC, CU, AG, or GA. In other words, P is the probability of homologous sites showing the transition type base differences. Similarly, let Q be the probability of homologous sites being occupied by pair UA, AU, CG, or GC, and R be the probability of UG, GU, CA, or AC. Thus, $Q + R$ represents the probability of homologous sites showing transversion type differences.

Then it can be shown that P, Q, and R satisfy the following set of differential equations (details of the derivation will be published elsewhere).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: 3ST model, three-substitution-type model; 2FC model, two-frequency-class model.

$$\begin{aligned} dP/dT &= 2\alpha - 2(2\alpha + \beta + \gamma)P \\ &\quad - 2(\alpha - \gamma)\,Q - 2(\alpha - \beta)R \\ dQ/dT &= 2\beta - 2(\beta - \gamma)P \\ &\quad - 2(\alpha + 2\beta + \gamma)Q - 2(\beta - \alpha)R \\ dR/dT &= 2\gamma - 2(\gamma - \beta)P \\ &\quad - 2(\gamma - \alpha)Q - 2(\alpha + \beta + 2\gamma)R. \end{aligned} \qquad [2]$$

The solution of this set of equations that satisfies the condition

$$P = Q = R = 0 \text{ at } T = 0, \qquad [3]$$

i.e., the two sequences are identical at the start, is

$$\left.\begin{aligned} P &= (1 - e^{\lambda_1 T} - e^{\lambda_2 T} + e^{\lambda_3 T})/4 \\ Q &= (1 - e^{\lambda_1 T} + e^{\lambda_2 T} - e^{\lambda_3 T})/4 \\ R &= (1 + e^{\lambda_1 T} - e^{\lambda_2 T} - e^{\lambda_3 T})/4 \end{aligned}\right\} \qquad [4]$$

in which $\lambda_1 = -4(\alpha + \beta)$, $\lambda_2 = -4(\alpha + \gamma)$, and $\lambda_3 = -4(\beta + \gamma)$. From these equations, we get

$$\begin{aligned} 4(\alpha + \beta)T &= -\ln[1 - 2(P + Q)] \\ 4(\alpha + \gamma)T &= -\ln[1 - 2(P + R)] \\ 4(\beta + \gamma)T &= -\ln[1 - 2(Q + R)]. \end{aligned} \qquad [5]$$

Because the evolutionary distance in terms of the number of base substitutions between the two sequences is given by Eq. 1, we obtain

$$K = -(1/4)\ln[(1 - 2P - 2Q)(1 - 2P - 2R)(1 - 2Q - 2R)]. \quad [6]$$

This formula has the desirable property that, as P, Q, and R approach zero, it converges to $K = P + Q + R$.

If the divergence time T is known, the base substitution rate per year is then given by

$$k_{\text{nuc}} = K/(2T), \qquad [7]$$

in which the subscript nuc means that the estimate refers to the rate per nucleotide site.

In the special case in which two types of transversion substitutions occur equally frequently so that $\gamma = \beta$, it can be shown that Eq. 6 reduces to

$$K = -(1/2)\ln[(1 - 2P - Q')\sqrt{1 - 2Q'}], \qquad [8]$$

in which $Q' = Q + R$ is the total proportion of transversion differences (9). This formula is useful when only two types of differences (i.e., transition and transversion) are distinguished in comparative studies of sequences (as in ref. 10). In a still simpler situation in which $\alpha = \beta = \gamma$, Eq. 6 reduces to

$$K = -(3/4)\ln[1 - (4/3)\lambda], \qquad [9]$$

in which $\lambda = P + Q + R$ is the proportion of sites that differ in the two sequences. This formula was obtained by Jukes and Cantor (11), and a formula for the large sample standard error of this estimator was given by Kimura and Ohta (12).

It is known that a large fraction of base substitutions at the third position in the codons are synonymous (i.e., do not lead to amino acid changes). So, it may be of interest to derive a formula for estimating the synonymous component of the number of base substitutions at position 3. From the code table we note that, roughly speaking, for a given combination of bases in the first and second codon positions, base substitutions at the third position are either completely synonymous or synonymous within purines or pyrimidines. Since these two situations occur approximately in equal frequencies, we can estimate the synonymous component of the number of substitutions per year at the third position of codons by

$$k_S' = \frac{1}{2}(\alpha + \beta + \gamma) + \frac{1}{2}\alpha = \frac{1}{2}(\alpha + \beta) + \frac{1}{2}(\alpha + \gamma). \qquad [10]$$

The corresponding distance is then given by $K_S' = 2Tk_S' = (1/4)[4(\alpha + \beta)T + 4(\alpha + \gamma)T]$, and noting Eqs. 5, we obtain

$$K_S' = -(1/4)\ln[(1 - 2P - 2Q)(1 - 2P - 2R)]. \qquad [11]$$

In the special case $\gamma = \beta$, this reduces to $K_S' = -(1/2)\ln(1 - 2P - Q')$, where $Q' = Q + R$ (ref. 9).

It is desirable to have a formula for the error variance (due to sampling) of the estimated value of K. If n is the number of nucleotide sites for which the two sequences are compared, then it can be shown that the large sample variance of K is

$$\sigma_K^2 = (1/n)[a^2P + b^2Q + c^2R - (aP + bQ + cR)^2], \qquad [12]$$

in which $a = (C_{12} + C_{13})/2$, $b = (C_{12} + C_{23})/2$, and $c = (C_{13} + C_{23})/2$ in which $C_{12} = 1/(1 - 2P - 2Q)$, $C_{13} = 1/(1 - 2P - 2R)$ and $C_{23} = 1/(1 - 2Q - 2R)$.

Similarly, for the estimate of the synonymous component K_S', the error variance is

$$\sigma_{K'S}^2 = (1/n)[a_S^2P + b_S^2Q + c_S^2R - (a_SP + b_SQ + c_SR)^2], \qquad [13]$$

in which $a_S = (C_{12} + C_{13})/2$, $b_S = C_{12}/2$, and $c_S = C_{13}/2$.

As an example, let us compare the nucleotide sequence of human presomatotropin (13) with that of rat presomatotropin (14). Excluding insertions or deletions ("gaps") that amount to three codons, there are 214 homologous codon positions that can be compared. For the first codon positions, we find $P = 28/214$, $Q = 9/214$, and $R = 10/214$, and, from Eqs. 6, we obtain $K = 0.264$. It is likely that the human and the rat diverged from each other late in the Mesozoic, some 80 million years ago, so we may take $T = 8 \times 10^7$. The evolutionary rate per site at codon position 1 for presomatotropin is then $k_{\text{nuc}} = K/(2T) = 1.65 \times 10^{-9}$ per year. From Eq. 12, the error variance of K becomes (taking $n = 214$) $\sigma_K^2 = 1.34 \times 10^{-3}$ so that the standard error is $\sigma_K = 3.66 \times 10^{-2}$. We can calculate the corresponding estimates of K for positions 2 and 3, and also for the synonymous component, as shown in the first line in Table 2. The table also lists (in the lines marked 3ST) estimates of evolutionary distances similarly computed by using data on the human preproinsulin gene (15, 16), rat preproinsulin gene I (17, 18), rabbit β-globin (19), mouse β-globin (20), rabbit α-globin (21), and mouse α-1-globin genes (22). Note that in the first four comparisons the diver-

Table 1. Combinations of bases

Type of difference	Transition type	Transversion type	
Sequence 1	U C A G	U A C G	U G C A
Sequence 2	C U G A	A U G C	G U A C
Frequency	P	Q	R

Various types of different base pairs at homologous nucleotide sites in two sequences compared.

Table 2. Estimates of K

Comparison	Model	Evolutionary distance per nucleotide site			
		K_1	K_2	K_3	K_S'
Human vs. rat presomatotropins	3ST	0.26 ± 0.04	0.18 ± 0.03	0.53 ± 0.07	0.44 ± 0.07
	2FC	0.28 ± 0.05	0.18 ± 0.04	0.75 ± 0.20	—
Human vs. rat I preproinsulins:	3ST	0.04 ± 0.03	0.00[*]	0.46 ± 0.12	0.38 ± 0.12
A + B chains	2FC	0.04 ± 0.04	0.00	0.60 ± 0.39	—
C peptide	3ST	0.18 ± 0.06	0.27 ± 0.10	0.95 ± 0.46	0.77 ± 0.51
	2FC	0.15 ± 0.08	0.30 ± 0.14	—[†]	—
Rabbit vs. mouse β-globins	3ST	0.16 ± 0.03	0.13 ± 0.03	0.43 ± 0.07	0.36 ± 0.07
	2FC	0.17 ± 0.05	0.14 ± 0.04	0.49 ± 0.11	—
Rabbit vs. mouse[‡] α-globins	3ST	0.12 ± 0.03	0.12 ± 0.03	0.54 ± 0.09	0.47 ± 0.09
	2FC	0.13 ± 0.04	0.13 ± 0.05	0.65 ± 0.17	—
Rabbit α- vs. rabbit β-globins	3ST	0.60 ± 0.08	0.44 ± 0.04	0.90 ± 0.14	0.68 ± 0.13
	2FC	0.64 ± 0.12	0.53 ± 0.14	1.19 ± 0.38	—

Evolutionary distances per site (together with standard errors) as estimated by using two models (3ST and 2FC). K_i, (i = 1, 2, 3), denotes the number of base substitutions at codon position i that separates the two sequences compared, and K_S' is the synonymous component at position 3.

[*] No observed changes among 51 codons.

[†] Inapplicable case.

[‡] Mouse α-1-globin gene of ref. 22

gence time may be taken as $T = 8 \times 10^7$ years and that the evolutionary rates per year can be obtained by dividing these values by $2T = 1.6 \times 10^8$.

TWO-FREQUENCY-CLASS (2FC) MODEL

This model is motivated by the observation that, in mammalian mRNAs, bases C and G are much higher in frequency than U and A at the third codon positions. For example, the average base composition at the third positions in several mammalian globin sequences (computed from data in table 2 of ref. 5) are: C, 40%; G, 32%; A, 6%; and U, 22%.

Let us group the four bases into two classes, U + A in one class (called A_1) and C + G in the other (called A_2). Let α and β be, respectively, the substitution rate of A_2 for A_1 and vice versa as shown in Fig. 1*b*. We denote by X and Y the respective frequencies of A_1A_1 and A_2A_2 pairs, and by Z the frequencies of the sum of A_1A_2 and A_2A_1 pairs when two homologous sequences are compared ($X + Y + Z = 1$). Then it can be shown that X, Y, and Z satisfy the differential equations

$$\begin{aligned} dX/dT &= -2\alpha X + \beta Z \\ dY/dT &= -2\beta Y + \alpha Z \\ dZ/dT &= 2\alpha X + 2\beta Y - (\alpha + \beta)Z. \end{aligned} \qquad [14]$$

To simplify the analysis, we assume that frequencies of A_1 and A_2 are in equilibrium so that they do not change with time. This means that the frequencies of A_1 and A_2 are given by p and $q = 1 - p$,

$$p = \beta/(\alpha + \beta). \qquad [15]$$

Under this assumption, the evolutionary distance between two sequences with respect to substitutions between A_1 and A_2 is $K = 2T(p\alpha + q\beta) = 4pq(\alpha + \beta)T$, and, incorporating the relevant solution of Eqs. 14, this leads to

$$K = -\theta \ln(1 - Z/\theta), \qquad [16]$$

in which $\theta = 2pq$, p is the frequency of base group A_1, and $q = 1 - p$ is that of A_2. Also, Z is the fraction of sites by which the two sequences differ from each other (i.e., A_1A_2 and A_2A_1). Note that this formula has the desirable property of converging to $K = Z$ as Z approaches zero, irrespective of the value of θ. If θ is in the range 0.4–0.5, then K does not depend much on θ if Z is less than 0.2. Note also that Eq. 9 is equivalent to this formula when $\theta = 3/4$. In applying this formula it may be desirable to estimate p not simply from the two sequences being compared but from a number of related sequences (if they are available). For example, for the comparison of globin sequences, we take $p = 0.28$ which is the average frequency of U + A at the third codon positions for six globins (rabbit α-, mouse α-, human β-, rabbit β-, mouse β-, and chicken β-). Let us suppose then that, in general, p is estimated by a sample of size N, and Z is estimated from a sample of size n. Then it can be shown that the standard error of K is given by

$$\sigma_K = \sqrt{a^2\sigma_\theta^2 + b^2\sigma_Z^2}, \qquad [17]$$

in which $\sigma_\theta^2 = 4(1 - 2p)^2p(1 - p)/N$, $\sigma_Z^2 = Z(1 - Z)/n$, $a = K/\theta - Z/(\theta - Z)$, and $b = \theta/(\theta - Z)$. Note that, for $p = 0.28$, which we assume for codon position 3 of globins, we have $\theta = 0.4$. If we take conservatively $N = 500$ (because, the six globins used to estimate p are not wholly independent), then $\sigma_\theta^2 = 3.12 \times 10^{-4}$. Because θ is not very sensitive to the change of p at the neighborhood of 0.5, we may take $\theta = 0.5$ unless p and q differ greatly from each other. Note that, for $\theta = 0.5$, we have $\sigma_\theta^2 = 0$ so that N is irrelevant for computing σ_K^2.

In order to estimate the total distance K by using this model, we first estimate the component K_b ("between-class component") by applying Eq. 16, classifying the four bases into two groups A_1 and A_2. Next, we apply Eq. 16 to the first class A_1, proceeding as if the two bases U and A make up 100%. This yields an estimate for the component K_{w1} ["within-class A_1 component," corresponding to $2T(\alpha_1 + \beta_1)$ of Fig. 1*b*]. Similarly, we obtain K_{w2} ("within-class A_2 component"). Then the total distance is obtained by

$$K = K_b + pK_{w1} + qK_{w2}. \qquad [18]$$

For codon position 3 of globins, we take $\theta = 0.4$, but for codon positions 2 and 3, we take $\theta = 0.5$. The difference between the estimate obtained by using the 2FC model and that obtained by the 3ST model becomes significant only when the base composition deviates greatly from equality and at the same time the evolutionary distance involved is large. This is evident when we compare values estimated by these two methods for K_1 and K_2 as listed in Table 2.

At the third codon positions, and particularly when the distance is large, however, the difference may become large. As an example, let us compare the α- and β-globins of the rabbit. Excluding insertions and deletions (gaps) that amount to 9 codons, there are 139 codons that can be compared ($n = 139$). For the third positions of these codons, we find $nX = 8$ (UU 4, AA 2, UA 1, AU 1), $nY = 82$ (CC 29, GG 32, CG 16, GC 5), and $nZ = 49$. Applying Eqs. **16** and **17** with $\theta = 0.4$ ($p = 0.28$), $Z = 49/139$, $n = 139$, and $N = 500$, we obtain $K_b = 0.836 \pm 0.334$. Also, applying these equations within classes A_1 (consisting of U and A) and A_2 (C + G), we get $K_{w1} = 0.346 \pm 0.306$ and $K_{w2} = 0.359 \pm 0.099$. Altogether we get $K_3 = 1.19 \pm 0.38$. On the other hand, the corresponding estimate obtained by using the 3ST model turns out to be $K_3 = 0.90 \pm 0.14$, which is likely to be an underestimate (see bottom line in Table 2).

CORRECTION FOR EXCLUDING INAPPLICABLE CASES

Equations for K, such as Eqs. **6**, **11**, and **16**, are derived by deterministic methods which are based on the assumption that the lengths of sequences involved are infinite. In other words, the sampling effect due to finite number of codons is disregarded. On the other hand, the actual sequences are all finite in length, and the observed numbers of differences are subject to statistical fluctuation. The most serious consequence of such fluctuation is that cases arise, particularly when the true value of K is large and n is small, for which the equations cannot be applied. I shall explain this using Eq. **16**. Let n be the total number of homologous sites and let j be the observed number of sites for which the two sequences differ from each other—that is, the number of A_1A_2 plus A_2A_1 pairs ($j = 0, 1, \ldots, n$). Then, j follows the binomial distribution

$$f(j) = \binom{n}{j} Z^j(1 - Z)^{n-j}, \qquad [19]$$

in which $Z = \theta[1 - \exp(-K/\theta)]$. If j happens to become equal to or larger than $n\theta$, then Eq. **16** cannot be used to estimate K by letting $Z = j/n$ in this equation, because $(1 - Z/\theta)$ becomes negative. If we exclude such "inapplicable cases," the estimate of K becomes biased and a correction will be required. Let $\tilde{K}$ be the average value of K obtained under the condition that inapplicable cases are excluded—i.e., $\tilde{K} = E\{K|j < n\theta\}$.

$$\tilde{K} = -\sum_{j=0}^{L} \theta \ln\{1 - j/(n\theta)\}f(j) \bigg/ \sum_{j=0}^{L} f(j), \qquad [20]$$

in which L is the maximum integer such that $L < n\theta$. Fig. 2 depicts the relationship between the true distance K and the conditional distance $\tilde{K}$, assuming $\theta = 0.4$ and $n = 140$. The graph suggests the possibility of a serious underestimate for K when its estimated value (applying Eq. **16** to position 3 of globins) turns out to be larger than about 1.0.

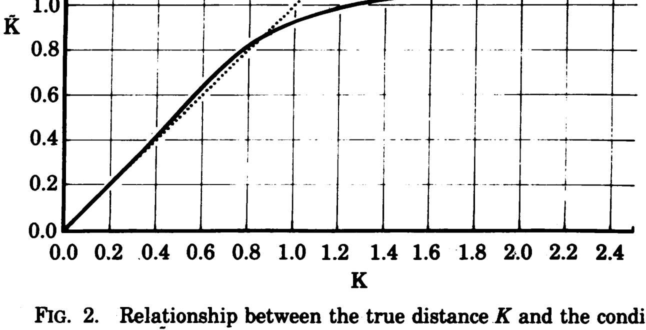

FIG. 2. **Relationship between the true distance K and the conditional distance $\tilde{K}$. For details, see text.**

DISCUSSION

Table 2 shows that the evolutionary rates of synonymous base substitutions at the third positions of codons are not only high but also are roughly equal (the two seemingly higher values are for C peptide, which has a large standard error, and the bottom comparison which involves a much longer time period, probably $T = 5 \times 10^8$). This is particularly evident if we contrast the evolutionary distances in presomatotropin with those of insulin (preproinsulin A + B chains), both involving human vs. rat comparisons. In presomatotropin hormone, the distance due to amino acid-altering substitutions per site, as estimated by $(K_1 + K_2)/2$, is 0.22, but the corresponding distance is only 0.02 in insulin. This means that amino acid-altering substitutions proceed some 10 times faster in presomatotropin than in insulin. On the other hand, the synonymous component at position 3, as estimated by K_S', is roughly equal in these two proteins. Furthermore, in α- and β-globins, the rates of synonymous substitutions are about equal to those of presomatotropin and insulin, although their amino acid-altering substitutions are intermediate between those of somatotropin and insulin. Note that the divergence time of rabbit and mouse is approximately the same as that of man and rat. Such uniform rate of synonymous substitutions has also been brought out by Miyata *et al.* (7).

These observations can be explained readily by the neutral mutation-random drift hypothesis of molecular evolution (the neutral theory, in short; see ref. 2). Unlike the Darwinian paradigm, this theory states that the majority of evolutionary mutant substitutions in the species are caused by random fixation of selectively neutral (i.e., selectively equivalent, but not necessarily functionally equivalent) mutants rather than by positive Darwinian selection. Although favorable mutations no doubt occur, the theory assumes that they are so rare as to be neglected in calculating rates of molecular evolution. The neutral theory predicts that the probability of a mutation being selectively neutral (that is, not harmful) is larger the less the mutation disrupts the existing structure and function of the molecule. At the limit in which all the mutations are selectively neutral, the rate of evolution per site (k) becomes equal to the total mutation rate (v) per site. In my opinion (see ref. 1), synonymous mutations are not very far from this limit and therefore the evolutionary rates of synonymous substitutions per site are nearly equal between different molecules.

Recently, an opposing view was proposed by Perier *et al.* (23). They claimed that the driving force for fixation is positive natural selection operating on some fraction of amino acid-altering ("replacement") changes and, that such a selected fixation carries along with it neutral alterations (including changes at silent sites) that have accumulated in that region of the DNA. In other words, they invoke the "hitchhiking" effect to explain fixation of synonymous changes.

I would like to point out that, unless we ignore the principles of population genetics, such an explanation cannot account for actual observations. In fact, such hitchhiking cannot bring about substitutions of neutral mutants at a very high rate when the selected changes occur at a very low rate. For example, take the histone H4 gene. The rate of replacement changes is almost zero, yet synonymous base substitutions occur at a rate comparable to that of replacement changes in fibrinopeptides, one of the most rapidly evolving molecules (1).

We can treat the problem in more detail. Because the hitchhiking effect extends only over short distances around a selectively driven gene, particularly in bringing associated mutations to fixation in the population, we consider a small segment of DNA, such as a gene locus, within which the incidence of crossing over is so low as to be neglected. Let us suppose that a new, advantageous, mutant allele at this gene locus appeared, first singly represented, in the population. In order that this selected mutant can bring other unselected (neutral) mutants to fixation by hitchhiking, the gene copy in which this advantageous mutant appeared must also contain at the same time a number of neutral mutants. Furthermore, in order to make the rate of substitution of neutral mutants per site *m* times higher than that of selectively driven mutations (in this case, amino acid-altering changes), each gene copy in the population must contain on the average *m* neutral mutants, irrespective of whether an advantageous mutation happened to occur in it or not. This factor *m* must be very large, probably 1000 or more in histone H4. On the other hand, if each gene copy contains a large number of neutral mutants, the corresponding (homologous) genes in different individuals differ from each other in so many bases that there is no such thing as a species-specific nucleotide sequence of a particular gene, say histones, hemoglobins, etc. In other words, every individual in the species would have quite different homologous sequences. This is contrary to observations.

Furthermore, the hitchhiking theory cannot explain the observation that, when genes of different proteins are studied, the evolutionary rates of synonymous substitutions are not only high, but also they are roughly equal to each other, even when their amino acid-altering substitution rates differ greatly.

I thank Drs. T. H. Jukes and K. M. Weiss for reading the manuscript and correcting the English. Thanks are also due to Dr. J. F. Crow who read the first draft carefully and made many useful suggestions to improve the presentation and to Dr. N. Takahata for stimulating discussions. Contribution no. 1339 from the National Institute of Genetics, Mishima, Shizuoka-ken, 411 Japan.

1. Kimura, M. (1977) *Nature (London)* **267**, 275–276.
2. Kimura, M. (1979) *Sci. Am.* **241** (5), 94–104.
3. Jukes, T. H. (1978) *J. Mol. Evol.* **11**, 267–269.
4. Jukes, T. H. & King, J. L. (1979) *Nature (London)* **281**, 605–606.
5. Jukes, T. H. (1980) *Naturwissenschaften*, **67**, 534–539.
6. Jukes, T. H. (1980) *Science*, **210**, 973–978.
7. Miyata, T., Yasunaga, T. & Nishida, T. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 7328–7332.
8. Nichols, B. P. & Yanofsky, C. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5244–5248.
9. Kimura, M. (1980) *J. Mol. Evol.*, in press.
10. van Ooyen, A., van den Berg, J., Mantel, N. & Weissmann, C. (1979) *Science* **206**, 337–344.
11. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism, III*, ed. Munro, H. N. (Academic, New York), pp. 21–132.
12. Kimura, M. & Ohta, T. (1972) *J. Mol. Evol.* **2**, 87–90.
13. Martial, J. A., Hallewell, R. A., Baxter, J. D. & Goodman, H. M. (1979) *Science* **205**, 602–607.
14. Seeburg, P. H., Shine, J., Martial, J. A., Baxter, J. D. & Goodman, H. M. (1977) *Nature (London)* **270**, 486–494.
15. Bell, G. I., Pictet, R. L., Rutter, W. J., Cordell, B., Tischer, E. & Goodman, H. M. (1980) *Nature (London)* **284**, 26–32.
16. Sures, I., Goeddel, D. V., Gray, A. & Ullrich, A. (1980) *Science* **208**, 57–59.
17. Cordell, B., Bell, G., Tischer, E., DeNoto, F. M., Ullrich, A., Pictet, R., Rutter, W. J. & Goodman, H. M. (1979) *Cell* **18**, 533–543.
18. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18**, 545–558.
19. Efstratiadis, A., Kafatos, F. C. & Maniatis, T. (1977) *Cell* **10**, 571–585.
20. Konkel, D. A., Tilghman, S. M. & Leder, P. (1978) *Cell* **15**, 1125–1132.
21. Heindell, H. C., Liu, A., Paddock, G. V., Studnicka, G. M. & Salser, W. A. (1978) *Cell* **15**, 43–54.
22. Nishioka, Y., Leder, A. & Leder, P. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2806–2809.
23. Perier, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20**, 555–566.