Evolution of DNA Base Composition Under No-Strand-Bias Conditions When the Substitution Rates Are Not Constant

J. R. Lobry and C. Lobry†*

*Laboratoire BGBP-CNRS, Université Claude Bernard, Villeurbanne, France; and †Centre International de Mathématiques Pures et Appliquées, Nice, France

The evolution of DNA base composition evolution is simplified to a six-parameter model when there are no strand biases for mutation and selection. We analyzed the dynamics of this model with special attention to the influence of a change in substitution rates. The G+C content of the DNA sequence tends to an equilibrium value that is controlled by four parameters of the model. When the substitution rates are not constant, the G+C equilibrium position is not constant. The DNA sequence base frequencies always tend to a state in which A = T and G = C within a strand, regardless of substitution rates. This is true even when the substitution rates are not constant over time. This provides a simple way of rejecting the model from inspection of present-day DNA base composition.

Introduction

Substitutions result in the DNA base composition of genomes changing during evolution. Let **X** be the column matrix,

$$\mathbf{X} = \begin{pmatrix} \mathrm{A}(t) \\ \mathrm{T}(t) \\ \mathrm{G}(t) \\ \mathrm{C}(t) \end{pmatrix}, \tag{1}$$

whose elements are the nucleotide frequencies at time *t*. Let **M** be the matrix for the continuous process of evolution of base frequencies:

$$\frac{d\mathbf{X}}{dt} = \mathbf{M}\mathbf{X}. \tag{2}$$

The entries in matrix **M** are the substitution rates. Many parametric forms of matrix **M** have been published (for review, see Rodriguez et al. 1990; Zharkikh 1994; Li 1997, pp. 59–78). The peculiar form we are interested in is based on the assumption of no-strand-bias conditions (Sueoka 1995). Under no-strand-bias conditions, the transition matrix **M** is given by

$$\mathbf{M} = \begin{pmatrix} -a-e-c & a & b & d \\ a & -a-e-c & d & b \\ c & e & -b-d-f & f \\ e & c & f & -b-d-f \end{pmatrix}, \tag{3}$$

where the six parameters ($a, \ldots, f$) represent the six substitution rates depicted in figure 1.

A general property of this model is that the solutions tend to equilibrium when the substitution rates are constant.

$$\mathbf{X}^* = \frac{1}{2}\begin{pmatrix} 1-\theta^* \\ 1-\theta^* \\ \theta^* \\ \theta^* \end{pmatrix}, \tag{4}$$

where

$$\theta^* = \frac{e+c}{b+c+d+e} \tag{5}$$

is the G+C content at equilibrium. This equilibrium is such that A = T and G = C within a strand, regardless of parameter values (Sueoka 1995; Lobry 1995).

However, there is conclusive evidence that the substitution matrix **M** is not constant during evolution. The most clear evidence is that the genome G+C content differs between species (Sueoka 1961, 1962). A more sensible model,

$$\frac{d\mathbf{X}}{dt} = \mathbf{M}(t)\mathbf{X}, \tag{6}$$

is obtained when substitution rates are allowed to change with time, while the matrix $\mathbf{M}(t)$ has the same structure as the constant one:

$$\mathbf{M}(t) = \begin{pmatrix} -a(t)-e(t)-c(t) & a(t) & b(t) & d(t) \\ a(t) & -a(t)-e(t)-c(t) & d(t) & b(t) \\ c(t) & e(t) & -b(t)-d(t)-f(t) & f(t) \\ e(t) & c(t) & f(t) & -b(t)-d(t)-f(t) \end{pmatrix}. \tag{7}$$

Despite the fact that matrix $\mathbf{M}(t)$ and the constant matrix **M** have the same structure, it is not obvious that the nonautonomous system (eq. 6) has the same properties as the autonomous one (eq. 2), i.e., that A(t) tends to T(t) and G(t) tends to C(t) as t tends to infinity.

For instance, the simple differential equation

$$\frac{dx}{dt} = -a(t)x, \quad \text{with} \quad a(t) = \frac{1}{1+t^2}, \tag{8}$$

has the property that for each fixed value of t, say $\underline{t}$, the parameter a($\underline{t}$) is strictly positive and thus the autonomous system obtained with the frozen parameter $a(\underline{t})$,

$$\frac{dx}{dt} = -a(\underline{t})x, \tag{9}$$

has all its solutions exponentially decreasing to zero. However, explicit solutions of equation (10),

$$x(t) = Ce^{-\mathrm{Arctg}(t)}, \tag{10}$$

tend to $Ce^{(-\pi/2)}$, a nonzero value. It is easy to find sys-

Key words: molecular evolution, DNA base composition, GC content, parity rules, base composition skew, mathematical model.

Address for correspondence and reprints: J. R. Lobry, Laboratoire BGBP-CNRS UMR 5558, Université Claude Bernard, 43 Bd. du 11 Novembre 1918, F-69622 Villeurbanne cedex, France. E-mail: lobry@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 16(6):719–723. 1999

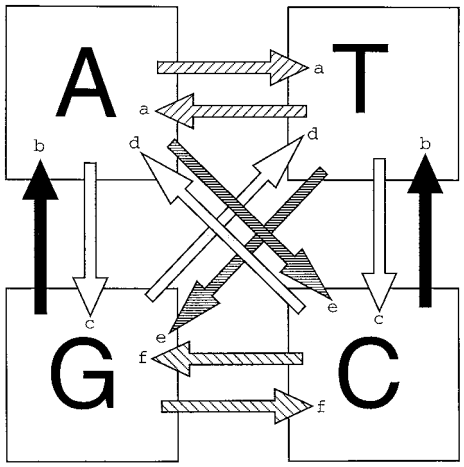

FIG. 1.—Diagram showing the substitution rates from one nucleotidic basis to another under no-strand-bias conditions. The 12 substitution rates are represented by arrows. Some substitution rates are equal under no-strand-bias conditions; this is denoted by the pattern filling the arrows. The notation for the six substitution rates ($a, \ldots, f$) is the same as that in Lobry (1995) and Sueoka (1995).

tems for more complex linear systems in two dimensions,

$$\frac{d\mathbf{X}}{dt} = \mathbf{A}(t)\mathbf{X}, \tag{11}$$

such that each system

$$\frac{d\mathbf{X}}{dt} = \overline{\mathbf{A}}(\underline{t})\mathbf{X} \tag{12}$$

with frozen $\mathbf{A}(\underline{t})$ is stable, but all the solutions of the nonautonomous system (eq. 11) are unstable. There are examples of such counterintuitive behaviors in the dynamics of populations in variable environments (Lobry, Sciandra, and Nival 1994).

This paper shows that the intrastrand equalities hold asymptotically provided the entries in the matrix $\mathbf{M}(t)$ satisfy the following criterion:

Hypothesis. There is a strictly positive constant μ such that for every t:

$$a(t) > \mu; \quad b(t) > \mu; \quad c(t) > \mu;$$
$$d(t) > \mu; \quad e(t) > \mu; \quad f(t) > \mu. \tag{13}$$

From a biological point of view, because all kinds of substitutions occur when sufficiently distant homologous sequences are compared, this assumption is a weak prerequisite. Since there is no analytical expression for the solutions of nonautonomous linear systems of differential equations, we verified our result using some differential inequality, which is normal in these situations.

Splitting the System into Two Subsystems

Let $\mathbf{P}$ be the nonsingular matrix,

$$\mathbf{P} = \frac{1}{2}\begin{pmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 1 \end{pmatrix}, \tag{14}$$

whose columns are the components in the initial basis of four vectors defining a new basis. This new basis is orthogonal, so the inverse of $\mathbf{P}$ is easily obtained:

$$\mathbf{P}^{-1} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}. \tag{15}$$

Let $\mathbf{Y}$ be the column matrix,

$$\mathbf{Y} = \begin{pmatrix} y_1(t) \\ y_2(t) \\ y_3(t) \\ y_4(t) \end{pmatrix} = \mathbf{P}^{-1}\mathbf{X}, \tag{16}$$

whose elements are the components in the new basis of the vector of nucleotide base frequencies at time t. The first component represents the $A(t)+T(t)$ content, the second represents the $G(t)+C(t)$ content, the third represents the $A(t)$-$T(t)$ skew, and the last represents the $C(t)$-$G(t)$ skew. The expression of system (eq. 2) in the new basis is given by

$$\frac{d\mathbf{Y}}{dt} = \mathbf{P}^{-1}\mathbf{M}\mathbf{P}\mathbf{Y} = \mathbf{N}\mathbf{Y}, \tag{17}$$

where

$$\mathbf{N} = \begin{pmatrix} -(e+c) & b+d & 0 & 0 \\ e+c & -(b+d) & 0 & 0 \\ 0 & 0 & -(2a+e+c) & d-b \\ 0 & 0 & e-c & -(2f+b+d) \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{N}_{11} & 0 \\ 0 & \mathbf{N}_{22} \end{pmatrix} \tag{18}$$

is a block-diagonal matrix, so studying the system is simplified by splitting it into two smaller subsystems.

First Subsystem: Changes in A+T and G+C Contents

When the substitution rates are constant, the first subsystem,

$$\begin{cases} \dfrac{dy_1}{dt} = -(e+c)y_1 + (b+d)y_2 \\ \dfrac{dy_2}{dt} = +(e+c)y_1 - (b+d)y_2, \end{cases} \tag{19}$$

is easily solved, because $y_1(t) + y_2(t) = 1$. The solutions are given by

$$\begin{cases} y_1(t) = 1 - \theta^* + (\theta^* - \theta_0)e^{-(b+c+d+e)t} \\ y_2(t) = \theta^* + (\theta_0 - \theta^*)e^{-(b+c+d+e)t}, \end{cases} \tag{20}$$

where θ_0 is the initial G+C content and θ^* is the G+C content at equilibrium (eq. 5).

Such an analysis is impossible when substitution rates b, c, d, and e are not constant, except when (i) the parameter functions are piecewise constant, or (ii) the parameter functions vary slowly. In the first case, the previous analysis holds after each discontinuity of a pa-

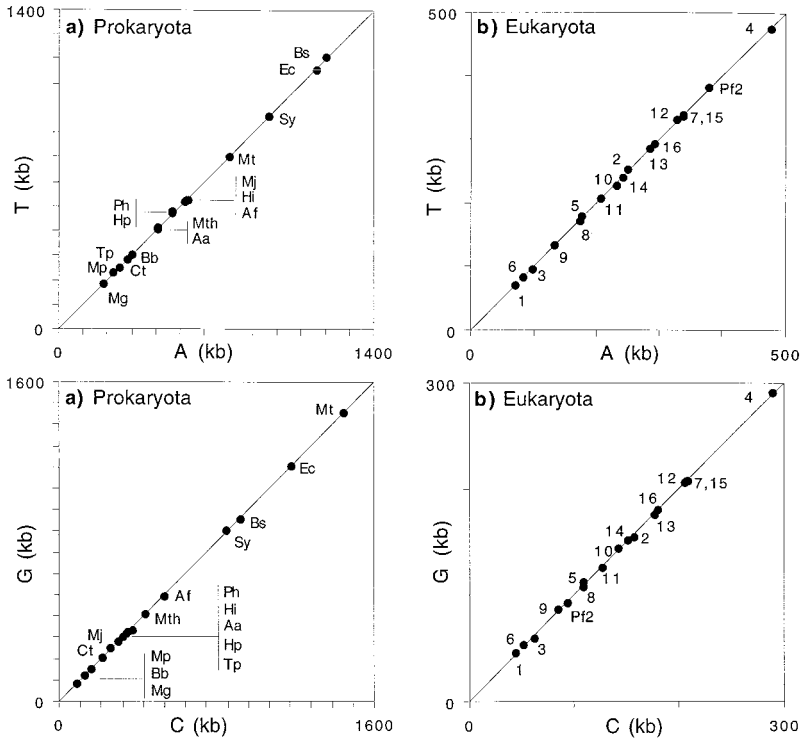

FIG. 2.—Base counts for all of the complete genomes available to date. The lines are the main diagonals ($y = x$). Points should be on these lines if A = T and C = G. Base counts are from the single-strand DNA data published on the NCBI ftp site (ncbi.nlm.nih.gov/genbank/genomes). *a*, Sixteen complete prokaryotic genomes, including 4 archaebacteria (Af, *Archaeoglobus fulgidus* [AE000782]; Mj, *Methanococcus jannaschii* [L77117]; Mth, *Methanobacterium thermoautotrophicum* [AE000666]; Ph, *Pyrococcus horikoshii* [AP000001–AP000007]) and 12 eubacteria (Aa, *Aquifex aeolicus* [AE000657]; Bs, *Bacillus subtilis* [AL009126]; Bb, *Borrelia burgdorferi* [AE000783]; Ct, *Chlamydia trachomatis* [AE001273]; Ec, *Escherichia coli* [U00096]; Hi, *Haemophilus influenzae* [L42023]; Hp, *Helicobacter pylori* [AE000511]; Mt, *Mycobacterium tuberculosis* [AL123456]; Mp, *Mycoplasma pneumoniae* [U00089]; Mg, *Mycoplasma genitalium* [L43967]; Sy, *Synechocystis* sp. [AB001339]; Tp, *Treponema pallidum* [AE000520]) *b*, Sixteen complete yeast (*Saccharomyces cerevisiae*) chromosomes (1, 2, . . . , 16) and for the malaria parasite (*Plasmodium falciparum*) chromosome 2 (Pf2).

rameter as long as the parameters are constant. Thus, the solutions $y_1(t)$ and $y_2(t)$ tend to the equilibrium defined by the parameters. Equilibrium is (nearly) attained if the parameters are constant for long enough. In the second case, the parameters are said to vary slowly if the maximum of their derivative is small compared with the sum $b(t) + c(t) + d(t) + e(t)$. As the variation of the equilibrium of the frozen system is slow compared with the decay rate of the solution, this equilibrium of the frozen system is a ''quasi-equilibrium'' of the actual system.

Points (i) and (ii) can be represented by mathematically rigorous statements, but this requires mathematical sophistication which is of little interest here, since we are mainly interested in the status of the intrastrand equalities and not the variations in G+C content.

Second Subsystem: Changes in A-T and C-G Skews

When the substitution rates are constant, the expression describing the second subsystem is

$$\begin{cases} \dfrac{dy_3}{dt} = -(2a + e + c)y_3 + (d - b)y_4 \\ \dfrac{dy_4}{dt} = (e - c)y_3 - (2f + b + d)y_4. \end{cases} \tag{21}$$

The trivial solution (0, 0) for system (21) is the only constant one, regardless of parameter values, because the determinant of the subsystem,

$$\det(\mathbf{N}_2) = 4af + 2ab + 2ad + 2ef + 2eb + 2cf + 2cd \tag{22}$$

is always strictly positive from hypothesis (13). This solution is also stable. Note the difference with the G+C content at equilibrium: here, the equilibrium position is independent of parameter values, such that even for a time-dependent process, the equilibrium position is always the same.

To deal with the nonconstant case, we introduce the city block distance from the equilibrium position at time t:

$$r(t) = |y_3(t)| + |y_4(t)|. \tag{23}$$

It decreases exponentially with time. To show this, we have to consider four cases corresponding to the signs of $y_3(t)$ and $y_4(t)$. For example, when both are positive,

$$y_3(t) \geq 0$$

and

$$y_4(t) \geq 0, \tag{24}$$

equation (23) becomes

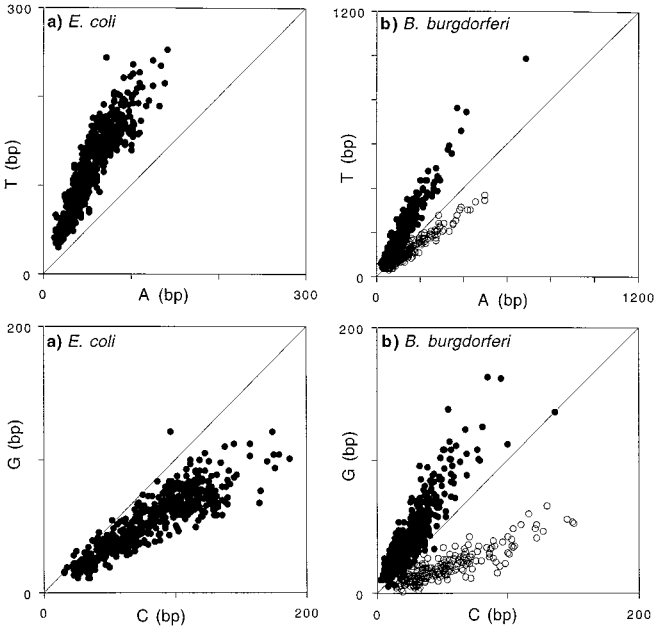

FIG. 3.—Base count comparisons (*a*) in the second codon positions of 512 coding sequences from *Escherichia coli* corresponding to integral membrane proteins (the GRAVY score [Kyte and Doolittle 1982] of the encoded proteins is greater than +0.5), and (*b*) in the third codon positions of 772 coding sequences from *Borrelia burgdorferi*, with black dots corresponding to the leading orientation with respect to the origin of replication (they are transcribed in the same direction as the replication fork motion) and white dots corresponding to the lagging orientation. Base counts are always from the sense strand.

$$r(t) = y_3(t) + y_4(t). \tag{25}$$

The rate of change of $r(t)$ with time is then given by

$$\frac{dr}{dt} = \frac{dy_3}{dt} + \frac{dy_4}{dt} \tag{26}$$

such that we obtain the following expression from subsystem (21):

$$\frac{dr}{dt} = -2(a + c)y_3 - 2(f + b)y_4. \tag{27}$$

Then, from hypothesis (13), the rate of variation of $r(t)$ is less than a threshold,

$$\frac{dr}{dt} \leq -2\mu y_3 - 2\mu y_4, \tag{28}$$

or, equivalently, from (eq. 25),

$$\frac{dr}{dt} \leq -2\mu r, \tag{29}$$

corresponding to an exponentially decreasing upper boundary for $r(t)$ values:

$$r(t) \leq r_0 e^{-2\mu t}. \tag{30}$$

The same inequality is also found for the three remaining cases for the signs of $y_3(t)$ and $y_4(t)$.

Hence, there is convergence to the equilibrium point ($y_3 = 0$, $y_4 = 0$). Returning to the meaning of these components, the equalities A = T and G = C are then an asymptotic property that is insensitive to a change in the substitution rates during the course of evolution.

Discussion

The evolution of DNA base composition under no-strand-bias conditions can be dissociated into two phenomena: (1) the evolution to the G+C equilibrium value and (2) the evolution to the equifrequencies A = T and G = C. The G+C equilibrium value and the convergence rate are controlled by the four substitution rates *b, c, d,* and *e.* A change in one of these parameters will then change the G+C equilibrium composition. The evolution to the equifrequencies A = T and G = C is very different: the position of the equilibrium position at which A = T and G = C is unaffected by a change in substitution rates; only the rate of convergence is affected. Then, if there is a statistically significant departure from these equalities, the model can be rejected with a high degree of confidence, because the equilibrium assumption is not required; it means that the no-strand-bias conditions were violated during the course of evolution of the DNA sequence under consideration. From a biological point of view, it means that either mutation, selection, or both were not symmetric with respect to the two DNA strands.

Figure 2 shows that A = T and C = G are sensible approximations for all of the complete genomes available to date (including representatives of archaebacteria, eubacteria, and eukaryota). However, systematic local deviations from A = T and G = C have been reported for many genomes (Smithies et al. 1981; Wu and Maeda 1987; Filipski 1990; Tanaka and Ozawa 1994; Jermiin, Graur, and Crozier 1995; Lobry 1996*a*, 1996*b*; Francino and Ochman 1997; Blattner et al. 1997; Kunst et al. 1997; Freeman et al. 1998; Grigoriev 1998; Mrázek and Karlin 1998; Reyes et al. 1998). Therefore, the model should be rejected on a local scale, meaning that the substitution rates are not symmetric with respect to the two strands. The substitution rates are dependent in a complex fashion on mutation and selection. Rejection of the model does not identify the cause of the asymmetry between the two strands, and extra biological information is needed before this interesting point can be discussed. We use two extreme cases to show how the violation of the model could be interpreted as the result of asymmetric selective pressure or asymmetric mutational pressure.

Coding sequences for integral membrane proteins in *Escherichia coli* do not follow A = T and C = G in their second codon positions (T > A and C > G; fig. 3*a*). There is a selective pressure on the amino acid content of these proteins to maintain their subcellular location by avoiding polar or charged amino acids (Asp, Glu, Lys, Arg, His, Asn, Gln) and favoring hydrophobic amino acids (Phe, Leu, Ile, Met, Val, Tyr, Trp). Because of the structure of the genetic code, this implies an excess of T over A and of C over G in the second codon positions. The violation of A = T and C = G is therefore the consequence of an asymmetric selective pres-

sure between the two strands, because only one strand corresponds to the sense strand.

Coding sequences from *Borrelia burgdorferi* do not follow A = T and C = G in their third codon positions, and there are clearly two groups with opposite deviations (fig. 3*b*). The third codon positions are nearly neutral, and the two groups correspond to the orientations of the coding sequences with respect to replication. The violation of A = T and C = G is therefore the consequence of an asymmetric mutation pressure between the leading and lagging strands of replication.

LITERATURE CITED

BLATTNER, F. R., G. PLUNKETT III, C. A. BLOCH et al. (14 co-authors). 1997. The complete genome sequence of *Escherichia coli* K12. Science **277**:1453–1462.

FILIPSKI, J. 1990. Evolution of DNA sequence, contributions of mutational bias and selection to the origin of chromosomal compartments. Pp. 1–54 *in* G. OLE, ed. Advances in mutagenesis research 2. Springer-Verlag, Berlin.

FRANCINO, M. P., and H. OCHMAN. 1997. Strand asymmetries in DNA evolution. Trends Genet. **13**:240–245.

FREEMAN, J. M., T. N. PLASTERER, T. F. SMITH, and S. C. MOHR. 1998. Patterns of genome organization in bacteria. Science **279**:1827a.

GRIGORIEV, A. 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. **26**:2286–2290.

JERMIIN, L. S., D. GRAUR, and R. H. CROZIER. 1995. Evidence from analyses of intergenic region for strand-specific directional mutation pressure in metazoan mitochondrial DNA. Mol. Biol. Evol. **12**:558–563.

KUNST, F., N. OGASAWARA, I. MOSZER et al. (149 co-authors). 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.* Nature **390**:249–256.

KYTE, J., and R. F. DOOLITTLE. 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. **157**:105–132.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland Mass.

LOBRY, C., A. SCIANDRA, and P. NIVAL. 1994. Paradoxal effects on growth and competition induced by fluctuations in environment. C. R. Acad. Sci. Life Sci. **317**:102–107.

LOBRY, J. R. 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. **40**: 326–330 (erratum in J. Mol. Evol. **41**:680).

———. 1996*a*. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13**:660–665.

———. 1996*b*. Origin of replication of *Mycoplasma genitalium.* Science **272**:745–746.

MRÁZEK, J., and S. KARLIN. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA **95**:3720–3725.

REYES, A., C. GISSI, G. PESOLE, and C. SACCONE. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. **15**:957–966.

RODRÍGUEZ, F., J. L. OLIVER, A. MARÍN, and J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. **142**:485–501.

SMITHIES, O., W. R. ENGELS, J. R. DEVEREUX, J. L. SLIGHTOM, and S.-H. SHEN. 1981. Base substitution, length differences and DNA strand asymmetries in the human $^{G}\gamma$ and $^{A}\gamma$ fetal globin region. Cell **26**:345–353.

SUEOKA, N. 1961. Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old and new data. J. Mol. Biol. **3**:31–40.

———. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48**:582–592.

———. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. **40**:318–325 (erratum in J. Mol. Evol. **42**:323).

TANAKA, M., and T. OZAWA. 1994. Strand asymmetry in human mitochondrial DNA mutations. Genomics **22**:327–335.

WU, C.-I., and N. MAEDA. 1987. Inequality in mutation rates of the two strands of DNA. Nature **327**:169–170.

ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. **39**:315–329.

DAN GRAUR, reviewing editor

Accepted January 18, 1999