J Mol Evol (1995) 40:326–330

JOURNAL OF MOLECULAR EVOLUTION

Properties of a General Model of DNA Evolution Under No-Strand-Bias Conditions

J.R. Lobry

Laboratoire de Biométrie, CNRS URA 243, Université Claude Bernard, 43 Bd. du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France

Received: 28 September 1994 / Accepted: 12 December 1994

Abstract. Under the hypothesis of no-strand-bias conditions, the Watson and Crick base-pairing rule decreases the complexity of models of DNA evolution by reducing to six the maximum number of substitution rates. It was shown that intrastrand equimolarity between A and T ($A^* \quad T^*$) and between G and C ($G^* \quad C^*$) is a general asymptotic property of this class of models. This statistical prediction was observed on 60 long genomic fragments (>50 kbp) from various kingdoms, even when the effect of the two opposite orientations for coding sequences is removed. The practical consequence of the model for estimating the expected number of substitutions per site between two homologous DNA sequences is discussed.

Key words: DNA evolution — No-strand bias — Coding sequence

Introduction

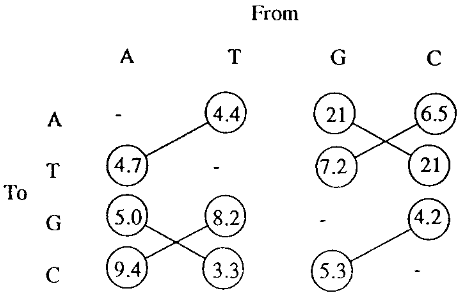
The genome of living organisms is a heteropolymer, the DNA double helix, described by the sequence of its constitutive monomers (symbolized A, T, G, and C) of one strand of the double helix, the other strand being unambiguously deduced by the famous one-to-one application known as the Watson and Crick (1953) interstrand base-pairing rule (BPR). For evolutionary studies on present-day homologous DNA sequences (i.e., that share a common ancestral DNA sequence), comparisons are made on the basis of only one strand, since considering the complementary strand would yield perfectly isomorphic results. The BPR is then always put in the background, and this may explain why the importance of BPR for building models of DNA evolution was not recognized sooner. Under the working hypothesis of *no-strand-bias conditions,* the BPR reduces the complexity of models of DNA evolution. This simplification, explained in the accompanying paper by Sueoka (1995), is called the intrastrand type-1 parity rule (PR1). In the following I will use the power of PR1 to derive formally a simple asymptotic property of the state of DNA sequences evolving under a PR1-compatible model—namely, the type-2 parity rule (PR2), which is the intrastrand equimolar frequencies between A and T and between G and C at equilibrium. Then I will challenge this statistical prediction with actual long DNA sequences. The agreement is good enough to say that the *no-strand bias conditions* are satisfied on average.

Definition of the Model

Models of DNA evolution are used to correct for multiple base-substitution events when estimating the distance between homologous DNA sequences, the correction being more and more important as we want to go back deeper in the past. In continuous time, these models are usually written in the form of an autonomous and homogeneous system of differential equations,

$$\frac{dX}{dt} = MX \qquad (1)$$

Abbreviations: BPR, Watson and Crick base pairing rule (A:T, G:C); PR1, Intrastrand type-1 parity rule ($i \neq j$, $m(i,j) \quad m(\bar{i},\bar{j})$); PR2, Intrastrand type-2 parity rule ($A^* \quad T^*$, $G^* \quad C^*$)



		From			
		A	T	G	C
To	A	-	4.4	21	6.5
	T	4.7	-	7.2	21
	G	5.0	8.2	-	4.2
	C	9.4	3.3	5.3	-

Fig. 1. Pattern of nucleotide substitution, in percent, from 13 pseudogene sequences (Li *et al.* 1984). Connected values should be equal under PR1 hypothesis.

where X $'(A(t),T(t),G(t),C(t))$ is the vector which represents the state of the system—that is, the frequencies of the bases A, T, G, and C at time t in the DNA sequence.

The substitution matrix, M, contains the substitution rates from one base to an other. By construction, the sum by column of M is always zero, so

$$A(t) + T(t) + G(t) + C(t) \quad K \tag{2}$$

is a first integral of system (1) whose biologically relevant solutions are for positive frequency values and $K = 1$.

Under the assumption *no-strand-bias conditions* the substitution rate from base i to base j equals the substitution rate from base $\bar{i}$ to base $\bar{j}$, where the notation $\bar{x}$ denotes the image of base x by BPR (Wu and Maeda 1987; Sueoka 1995). This hypothesis on the nature of the process, PR1, is consistent with data from pseudogenes (Fig. 1) and divides by two the maximum number of distinct parameters in the substitution matrix:

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

The six substitution rates $(a, \ldots, f)$ are assumed to be strictly positive. The notations are the same as in the accompanying paper (Sueoka 1995). The reduced number of parameters in (3) makes formal treatments of system (1) easier, yielding general properties that do not depend on a particular set of parameter values.

Expected Asymptotic Behavior

By solving numerically $dX/dt = 0$ under the constraint given by (2) with $K = 1$, Sueoka found that, for a wide range of different substitution rates, if the system (1) is at equilibrium then the base frequencies at equilibrium (A^*, T^*, G^* and C^*) are such that A^* T^* and G^* C^*. This proposition is called PR2. Note that the reciprocal proposition of PR2 is not true: observing that A T and G C for a DNA sequence does not mean that the equilibrium is reached. However, it is possible to show that A T and G C is an asymptotic property of all solutions of system (1) under PR1 hypothesis.

I found that the eigenvalues of M are given by

$$\begin{cases} \lambda_1 = 0 \\ \lambda_2 = -(b+c+d+e) \\ \lambda_3 = \dfrac{1}{2}(\alpha - \beta) \\ \lambda_4 = \dfrac{1}{2}(\alpha + \beta) \end{cases}$$

with

$$\begin{aligned} &\alpha = -(2a + \lambda_2 + 2f) \\ &\beta = \sqrt{\Delta} \qquad \text{if } \Delta \geqslant 0 \\ &\beta = i\sqrt{-\Delta} \qquad \text{if } \Delta < 0 \\ &\Delta = 4(b-d)(c-e) + (-2a + d + b - e - c + 2f)^2 \end{aligned}$$

As the first eigenvalue is zero, its associated eigenvector,

$$V_1 = \begin{pmatrix} d+b \\ d+b \\ e+c \\ e+c \end{pmatrix}$$

is a basis of the subspace of equilibrium points of system (1). By equation (2) with $K = 1$ we obtain the biologically relevant equilibrium point

$$\begin{cases} A^* = T^* = \dfrac{d+b}{2(b+c+d+e)} \\ G^* = C^* = \dfrac{e+c}{2(b+c+d+e)} \end{cases} \tag{4}$$

which respects PR2 proposition for any acceptable parameter values. The base frequencies at equilibrium are expressed by (4) as a simple function of substitution rates; in particular, the initial condition of system (1) does not influence the equilibrium position. However, equation (4) by itself is not sufficient to say that A T and C G is an asymptotic property because it contains no information on the *stability* of the equilibrium point. When λ_3 and λ_4 are complex numbers there is an oscillatory component in the solution of system (1), so the stability of the equilibrium point is far from being intuitive.

To show that the equilibrium point is stable, one may check that the real parts of eigenvalues λ_2, λ_3, and λ_4 are strictly negative. (This is easy when $\Delta \leqslant 0$ but a little more tedious when $\Delta > 0$.) More simply, we note that M belongs to the class of compartmental matrices, which are known to have no eigenvalue with a positive real part and no purely imaginary eigenvalue (Jacquez and Simon 1993). Moreover, as M corresponds to a closed system with no internal traps (all parameters are strictly positives) the multiplicity of λ_1 is one by Foster-Jacquez theorem (Foster and Jacquez 1975). Then, starting from any initial condition in the positive orthant $\Re^n_+$ for which

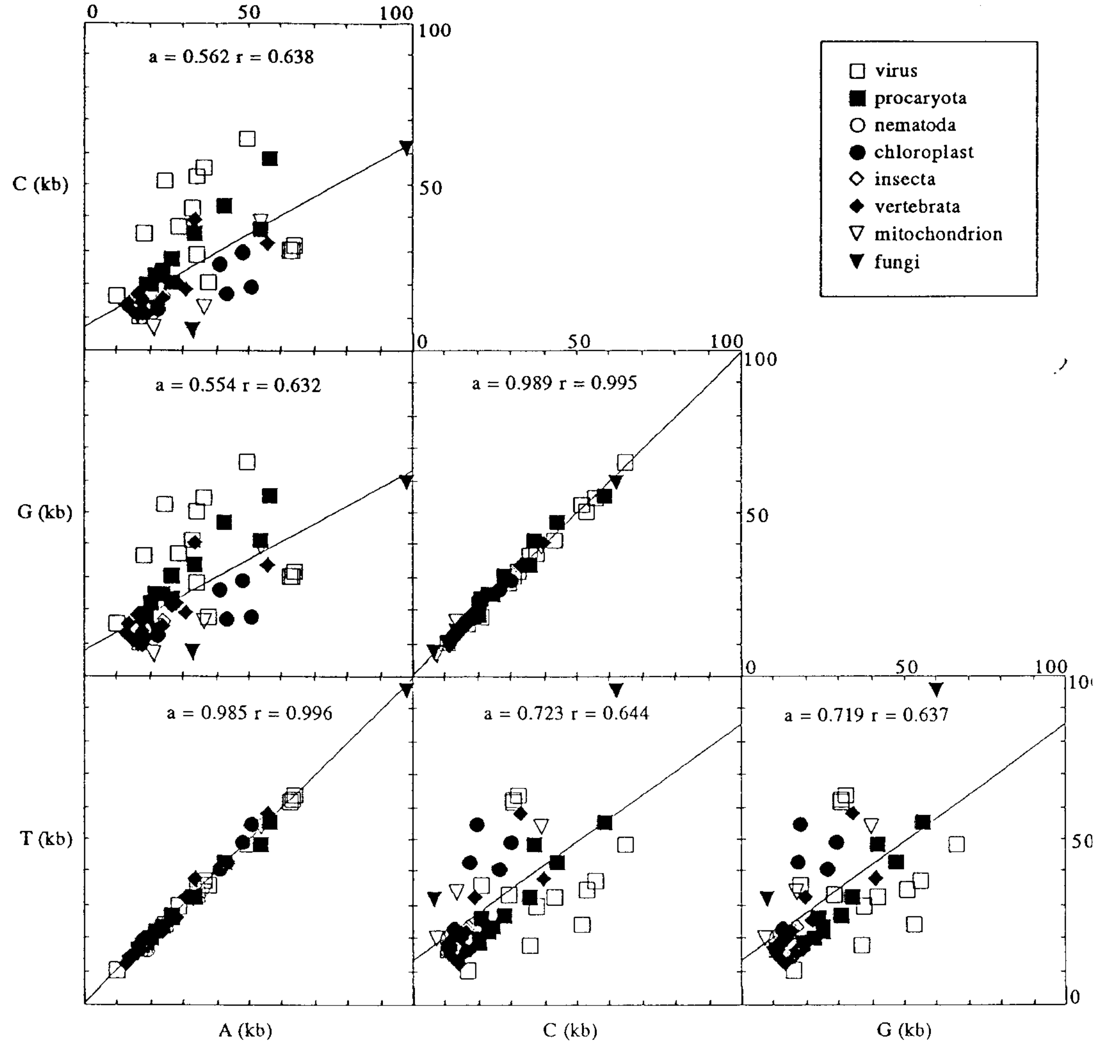


Fig. 2. Base composition of 60 long genomic fragments (>50 kbp) extracted from GenBank release 83 (15 June 1994). For each plot, the slope, *a*, of the regression line, *y*   *ax* + *b*, and the linear correlation coefficient, *r*, are given.

equation (2) is true with $K = 1$, trajectories will tend exponentially to frequencies at equilibrium given by (4).

To summarize, when the process of DNA evolution is governed by PR1, the equilibrium point defined by (4) is stable and will be reached exponentially regardless of the initial state of the DNA sequence. This equilibrium point is such that intrastrand equimolarity between A and T (A^* T^*) and between G and C (G^* C^*) holds.

Observed Asymptotic Behavior

Using long DNA sequences it is possible to test PR2 with extraordinary precision. By extracting from GenBank release 83 the 60 DNA sequences longer than 50 kbp the correlation between the total number of A and T and between C and G is striking (Fig. 2) and holds for DNA sequences from various kingdoms having very different genomic organizations. This observation of PR2 was incidentally reported by Chargaff and co-workers (Rudner et al. 1968; Chargaff 1979) and Prabhu (1993).

At first glance, this is a rather surprising result because of the unbearable lightness of the hypothesis of *no-strand-bias conditions*. This hypothesis *cannot* be locally true, especially in prokaryotes, where the density of coding sequences is high along the DNA sequence. For coding sequences there is a strong selective pressure to maintain the amino-acid composition. As only one strand is the coding strand, high strand bias is locally (i.e., at the scale of a coding sequence) expected, and local violations of PR2 are indeed observed (Fig. 3).

One may suspect that local strand biases are canceled out with two opposite orientations for coding sequences. This does not seem to be a crucial requirement since PR2 still holds when all coding sequences are concatenated in a chimeric sequence that removes the opposite orientation effect (Fig. 4). A similar observation was reported

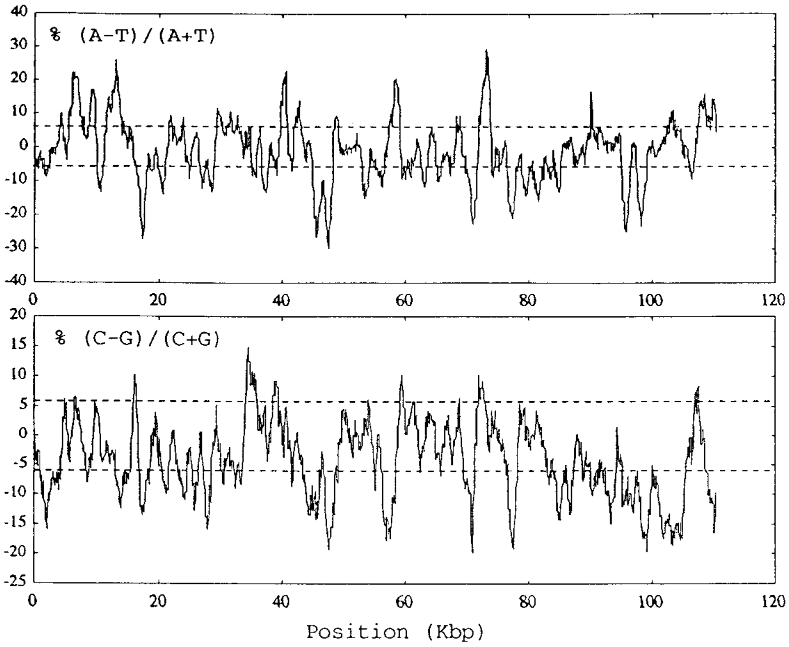


Fig. 3. Local violatiońs of PR2. The DNA sequence ECO110K from *Escherichia coli* (0–2.4-min region) was analyzed with a moving window of 1,000 bp (roughly the size of a coding sequence) with an incremental step of 100 bp. If PR2 was locally true the two plotted indices (*A* − *T*)/(*A* + *T*) and (*C* − *G*)/(*C* + *G*) should be zero, which is clearly not the case. The *dashed lines* represent the error lines expected from the random deviation from PR2 at a critical level of 5%.

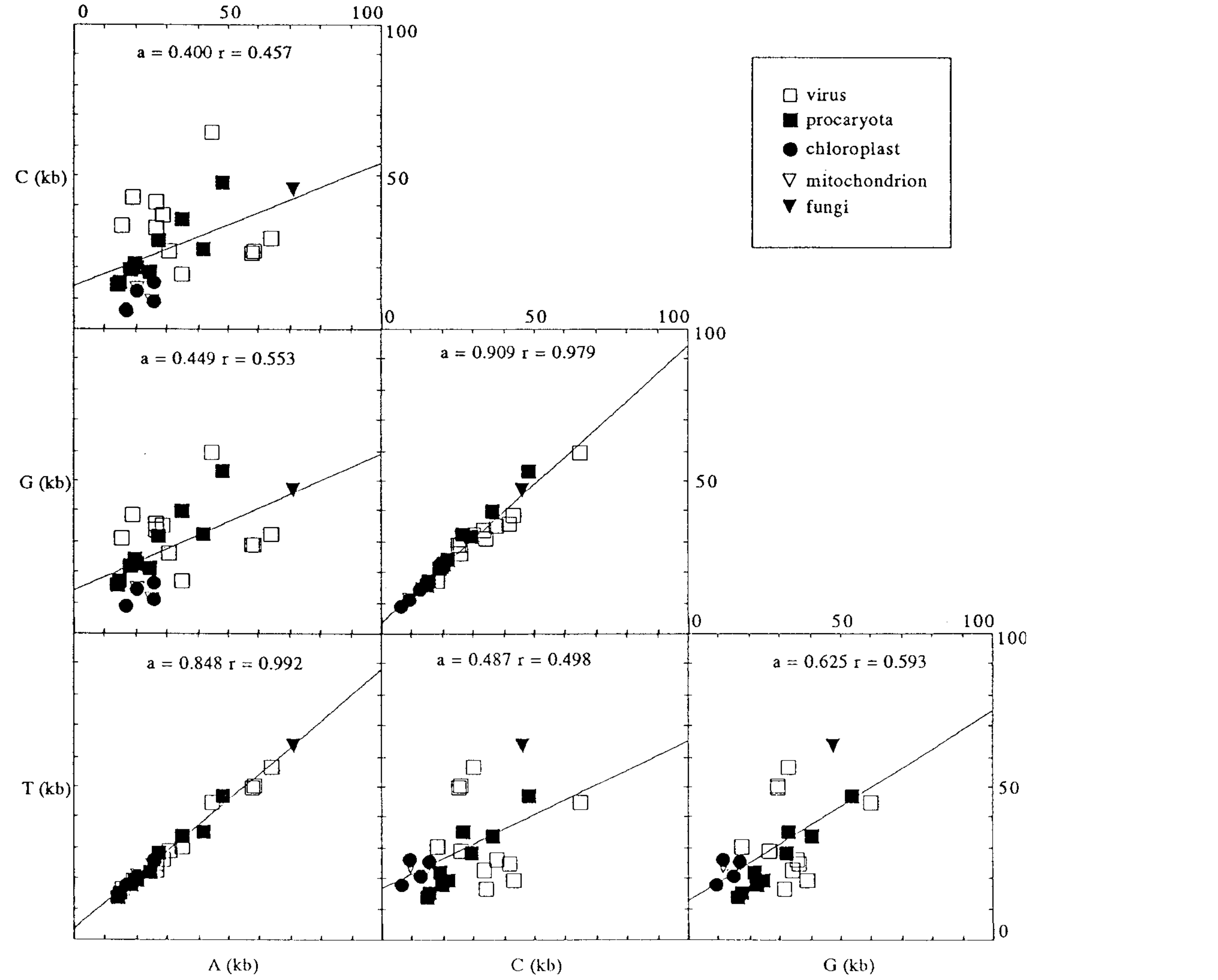


Fig. 4. Base composition of the concatenated coding sequences extracted from the 60 long genomic fragments in Fig. 2. Chimeric sequences of less than 50 kbp were removed. For each plot, the slope, *a*, of the regression line, *y* *ax* + *b*, and the linear correlation coefficient, *r*, are given.

for the composition of mRNA from whole cell extracts (Elson and Chargaff 1955).

Intrastrand DNA base-composition rule PR2 seems to hold by the sole virtue of the law of large numbers: the local strand biases influence base frequencies like a random variable with mean zero and tend to neutralize when a large number of base is taken into account, so on average the *no-strand-bias conditions* are satisfied.

Discussion

The model defined by equation (3) is the most general model that could be written under the assumption of no-strand-bias conditions. This assumption seems sensible because the asymptotic property (A^* T^* and G^* C^*) is not rejected by the data. This is of course an indirect argument because we do not know whether the equilibrium is reached or not in present-day DNA sequences. However, simulations seem to indicate that transients are short as compared with evolutionary time scales (Sueoka 1993).

Thanks to the simplification of the process of DNA evolution it implies, this model could be useful for estimating the expected number of substitutions per site between two homologous DNA sequences. The model does not correspond to a reversible process, so we cannot use the mathematically convenient properties of this class of models (Yang 1994) without introducing an extra *ad hoc* constraint on substitution parameters.

References

Chargaff E (1979) How genetics got a chemical education. Ann NY Acad Sci 325:345–360

Elson D, Chargaff E (1955) Evidence of common regularities in the composition of pentose nucleic acids. Biochim Biophys Acta 17: 367–376

Foster DM, Jacquez JA (1975) Multiple zeros for eigenvalues and the multiplicity of traps of a linear compartmental system. Math Biosci 26:89–97

Jacquez JA, Simon CP (1993) Qualitative theory of compartmental systems. SIAM Rev 35:43–79

Li WH, Wu CI, Luo CC (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Prabhu VV (1993) Symmetry observation in long nucleotide sequences. Nucleic Acids Res 21:2797–2800

Rudner R, Karkas JD, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands: III. Direct analysis (1968) Proc Natl Acad Sci USA 60:921–922

Sueoka N (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. J Mol Evol 37:137–153

Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J Mol Evol 40:318–325

Watson JD, Crick FHC (1953) A structure for deoxyribose nucleic acid. Nature 171:737–738

Wu CI, Maeda N (1987) Inequality in mutation rates of the two strands of DNA. Nature 327:169–170

Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

J Mol Evol (1995) 41:680

JOURNAL OF MOLECULAR EVOLUTION

Erratum

Properties of a General Model of DNA Evolution Under No-Strand Bias Conditions

J.R. Lobry

Laboratoire de Biométrie, CNRS URA 243, Université Claude Bernard, 43 Bd. du 11 Novembre 1918, F-69622 Villeurbanne Cedex, France

J Mol Evol (1995) 40:326–330

Due to a printer's error, 19 equal signs (=) were omitted in the published version of this article.

Page 326, Abstract, Line 6 should read:

($A^* = T^*$) and between G and C ($G^* = C^*$) is a general

Page 326, the footnote should read:

Abbreviations: BPR, Watson and Crick base pairing rule (A:T, G:C); PR1, Intrastrand type-1 parity rule ($i \neq j$, $m(i,j) = m(\bar{i},\bar{j})$); PR2, intrastrand type-2 parity rule ($A^* = T^*$, $G^* = C^*$)

Page 327, Column 1, Line 1 should read:

where $X = {}^t(A(t),T(t),G(t),C(t))$ is the vector which represents the state

Page 327, Column 1, Eq. (2) should read:

$$A(t) + T(t) + G(t) + C(t) = K \qquad (2)$$

Page 327, Column 1, Last paragraph should read:

By solving numerically $dX/dt = 0$ under the constraint given by (2) with $K = 1$, Sueoka found that, for a wide range of different substitution rates, if the system (1) is at equilibrium then the base frequencies at equilibrium (A^*, T^*, G^* and C^*) are such that $A^* = \mathrm{T}^*$ and $G^* = \mathrm{C}^*$. This proposition is called PR2. Note that the reciprocal proposition of PR2 is not true: observing that $A = T$ and $G = C$ for a DNA sequence does not mean that the equilibrium is reached. However, it is possible to show that $A = T$ and $G = C$ is an asymptotic property of all solutions of system (1) under PR1 hypothesis.

Page 327, Column 2, Sentence beginning, "However, equation (4) . . . ," should read:

However, equation (4) by itself is not sufficient to say that $A = T$ and $C = G$ is an asymptotic property because it contains no information on the *stability* of the equilibrium point.

Page 328, Fig. 2 caption should read:

Fig. 2. Base composition of 60 long genomic fragments (>50 kbp) extracted from GenBank release 83 (15 June 1994). For each plot, the slope, a, of the regression line, $y = ax + b$, and the linear correlation coefficient, r, are given.

Page 328, Column 1, Paragraph 1 should read:

To summarize, when the process of DNA evolution is governed by PR1, the equilibrium point defined by (4) is stable and will be reached exponentially regardless of the initial state of the DNA sequence. This equilibrium point is such that intrastrand equimolarity between A and T ($A^* = T^*$) and between G and C ($G^* = C^*$) holds.

Page 329, Fig. 4 caption should read:

Fig. 4. Base composition of the concatenated coding sequences extracted from the 60 long genomic fragments in Fig. 2. Chimeric sequences of less than 50 kbp were removed. For each plot, the slope, a, of the regression line, $y = ax + b$, and the linear correlation coefficient, r, are given.

Page 330, Paragraph 2 should read:

The model defined by equation (3) is the most general model that could be written under the assumption of no-strand-bias conditions. This assumption seems sensible because the asymptotic property ($A^* = T^*$ and $G^* = C^*$) is not rejected by the data. This is of course an indirect argument because we do not know whether the equilibrium is reached or not in present-day DNA sequences. However, simulations seem to indicate that transients are short as compared with evolutionary time scales (Sueoka 1993).